



UNIVERSITÀ DEGLI STUDI DI PADOVA

DEPARTMENT OF MATHEMATICS "TULLIO LEVI-CIVITA"

PH.D. SCHOOL IN BRAIN, MIND, AND COMPUTER SCIENCE

CURRICULUM IN COMPUTER SCIENCE AND INNOVATION FOR SOCIETAL CHALLENGES

XXXVI CYCLE

ADVANCING SOCIAL NETWORK ANALYTICS: RESILIENCE AND SECURITY

THESIS WRITTEN WITH THE FINANCIAL CONTRIBUTION OF CHISITO S.R.L.

PH.D. SCHOOL COORDINATOR

PROF. ANNA SPAGNOLLI

SUPERVISOR

PROF. MAURO CONTI

CO-SUPERVISOR

PROF. ANNA SPAGNOLLI

PH.D. STUDENT

PIER PAOLO TRICOMI

“IMPOSSIBLE? WE DID A LOT OF IMPOSSIBLE THINGS ON THIS JOURNEY.
I’M TIRED OF HEARING THAT THINGS ARE IMPOSSIBLE OR USELESS.
THOSE WORDS MEAN NOTHING TO ME.”

– JOTARO KUJO

Abstract

In the digital age, Online Social Networks (OSNs) have emerged as epicenters of human interaction, facilitating the creation, sharing, and dissemination of information at an unprecedented scale. The vast reservoir of user-generated data within OSNs has become a valuable resource for researchers, analysts, and practitioners. Social Network Analytics (SNA) has arisen as a powerful tool to extract insights from these data, enabling a better understanding of social structures and dynamics. However, the ever-changing landscape of OSNs, marked by the emergence of new platforms and shifts in user behavior, necessitates constant adaptation of SNA methodologies and tools. This dissertation advances SNA in three dimensions: (i) explaining influence and engagement mechanisms in trending OSNs; (ii) developing resilient SNA tools designed to function effectively in adversarial environments, and (iii) exploring security and privacy concerns in modern social platforms.

The first part of this thesis begins by examining how virtual influencers are transforming OSNs and influencer marketing. While major companies and brands increasingly embrace them, individuals remain divided between enthusiasm and apprehension regarding this phenomenon. The thesis then unveils Instagram engagement mechanisms to optimize content creation and delves into TikTok's unique influence dynamics, emphasizing how influencers can expand their reach. These studies demonstrate that influence and engagement patterns are strictly related to the tiers and categories of influencers, an aspect not considered in the existing literature. The part concludes with a case study exemplifying information manipulation on Twitter orchestrated by social bots. In the context of the 2022 Italian General Elections, these bots engineered the public discourse surrounding the Russo-Ukrainian conflict, frequently initiating and steering the discussions.

Motivated by the existence of such adversarial activities, the second part of the thesis focuses on developing resilient SNA tools tailored for adversarial contexts. The part begins by identifying Instagram crowdturfing, an emerging phenomenon wherein individuals generate fake engagement using their authentic profiles, behind a monetary reward. The analysis reveals that over 20% of mega influencers' engagement is artificial. Then, the thesis delves into categorizing simple but powerful obfuscation techniques OSN users adopt to evade moderators, and proposes a detection mechanism based on supervised and unsupervised Deep Learning (DL) strategies. The part concludes by introducing the innovative concept of social honeypots for examining OSN communities and trends. These honeypots are fully automated Instagram pages, powered by generative AI, that attract users for subsequent analysis.

Eventually, the notion of social networks has expanded beyond traditional OSNs to encompass contemporary digital landscapes like video games and the Metaverse. In fact, within these virtual worlds, people engage, communicate, and forge connections, giving rise to online communities and social interactions. However, the widespread use of these modern social platforms also results in the generation of massive amounts of (public) data, which can be exploited for malicious purposes. Unfortunately, numerous threats within this evolving landscape remain unknown to the research community, while techniques to identify malicious users can be the key to mitigating these risks. This thesis's final part focuses on enhancing security and privacy in modern social platforms, such as video games and the Metaverse. First, it introduces PvP, a DL-based framework that can effectively identify gamers based on their play style. Then, the thesis assesses an attribute inference attack in Dota 2, with far-reaching privacy consequences for millions of gamers. In fact, it demonstrates that players' private information, including their age, gender, or personality traits, can be inferred with up to 96% precision. The thesis concludes by presenting a comprehensive user profiling framework for augmented and virtual reality, addressing privacy and security challenges within the Metaverse's enabling technologies.

Contents

ABSTRACT	v
LIST OF FIGURES	xiii
LIST OF TABLES	xvii
I INTRODUCTION	1
1.1 Research Motivation and Contribution	2
1.1.1 Publications	8
I Influence and Engagement in Trending Online Social Networks	11
2 VIRTUAL INFLUENCERS IN ONLINE SOCIAL MEDIA	13
2.1 Virtual Influencers Timeline	14
2.2 Popular Virtual Influencers	15
2.3 The Influencer Marketing	17
2.4 Virtual vs. Real Influencers	18
2.4.1 Opportunities of Virtual Influencers	18
2.4.2 Threats of Virtual Influencers	18
2.4.3 Top Real vs. Virtual Influencers	19
2.4.4 Ontology and Ethics	19
2.5 Opinions about Virtual Influencers	20
2.5.1 People’s Opinion on Virtual Influencers	20
2.5.2 Creator’s Opinion	21
2.5.3 Virtual Influencer’s Opinion	22
2.6 The Future of Virtual Influencers	22
3 FOLLOW US AND BECOME FAMOUS! INSIGHTS AND GUIDELINES FROM INSTA- GRAM ENGAGEMENT MECHANISMS	25
3.1 Related Works	27
3.1.1 Limitations of Existing Literature	27
3.2 Dataset & Preliminary Assessments	28
3.2.1 Dataset Description	28
3.2.2 Engagement Metrics: Likes & Comments	29
3.2.3 The Importance of Tiers and Categories	29
3.2.4 Features Extraction	30
3.3 Predict & Interpret the Engagement	32
3.3.1 Correlation Analysis	32
3.3.2 Engagement Prediction & Guidelines Methodology	34
3.4 Spotting Instagram Hot Topics	38

3.4.1	Methodology	38
3.4.2	General vs User-specific Hot topics	40
3.5	Guidelines Insights	41
3.5.1	Guidelines for Likes, Comments, and Topics	42
3.5.2	Guidelines for Engaging Captions	43
3.5.3	Limitations	43
3.6	Conclusion	44
4	CLIMBING THE INFLUENCE TIERS ON TIKTOK: A MULTIMODAL STUDY	45
4.1	Related Work	46
4.2	Our New TIDES Dataset	47
4.3	Feature Extraction	48
4.3.1	Traditional Features	48
4.3.2	Audio Features	48
4.3.3	Video Features	49
4.3.4	Text Features	49
4.4	Descriptive Statistical Exploration of Specific Features	50
4.4.1	Non-Actionable Features	50
4.4.2	Actionable Features	51
4.5	Influencer Tier Classification	55
4.5.1	Data Aggregation	55
4.5.2	All Features Analysis	56
4.6	Actionable Features analysis	59
4.6.1	Which features to improve?	59
4.7	Conclusion and Future Works	63
5	TWITTER BOTS INFLUENCE ON THE RUSSO-UKRAINIAN WAR DURING THE 2022 ITALIAN GENERAL ELECTIONS	65
5.1	Related Works	67
5.2	Dataset Creation	67
5.3	The Russo-Ukrainian War in Italian Politics	68
5.3.1	The Importance of Conflict for Italian Political Parties	69
5.3.2	Temporal Analysis of Russo-Ukrainian Discussions	70
5.4	Bots Influence Analysis	72
5.4.1	Bots Presence Analysis	72
5.4.2	Bots Topics Distortion Analysis	74
5.4.3	Bots Temporal Influence Analysis	76
5.5	Discussion	79
5.5.1	Limitations	79
5.6	Conclusion and Future Works	80
II	Developing Resilient Social Network Analytics Tools	81
6	ARE WE ALL IN A TRUMAN SHOW? SPOTTING INSTAGRAM CROWDTURFING THROUGH SELF-TRAINING	83
6.1	Related Works	84
6.1.1	Crowdturfing in Online Social Media	85

6.1.2	Instagram Fake Accounts Detection	85
6.1.3	Semi-supervised Fake Accounts Detection	86
6.2	Crowdturfing Providers Analysis	86
6.3	Crowdturfing Profiles Detection	87
6.3.1	Dataset and feature selection	88
6.3.2	Our Semi-Supervised Model	88
6.3.3	Baseline Comparison	90
6.4	Crowdturfing Analysis: Profiles Information	91
6.4.1	Followers and following ratio analysis	92
6.4.2	Fake profiles biography analysis	94
6.4.3	Fake profiles external URLs analysis	94
6.5	Real vs Crowdturfing Comments Analysis	96
6.5.1	Stylometric Analysis	96
6.5.2	Common Words Analysis	97
6.5.3	Number of Comments per User	98
6.5.4	Language Analysis	98
6.5.5	Topics Analysis	99
6.6	Conclusion	100
7	TURNING CAPTCHAS AGAINST HUMANITY: CAPTCHA-BASED ATTACKS IN ON-LINE SOCIAL MEDIA	103
7.1	Introduction	103
7.2	Background & Related Works	105
7.2.1	Security of Machine Learning Applications	106
7.2.2	Moderators in OSN	106
7.2.3	CAPTCHA	107
7.3	Captcha Attack: A Taxonomy	108
7.3.1	Challenges from OSN's users: a Taxonomy	108
7.3.2	OCR-failure	110
7.3.3	Classifier-failures	110
7.3.4	Statistics from the wild	110
7.4	Attack Execution	111
7.4.1	Motivation	111
7.4.2	CC-CAPA Generation Procedure	112
7.4.3	CC-CAPA Dataset	113
7.5	Attack Results	116
7.5.1	Overview	116
7.5.2	Image Moderators	117
7.5.3	Cross-domain Moderators	117
7.6	CC-CAPA Detection Strategies	119
7.6.1	Overview	119
7.6.2	Supervised Approach: Classification	120
7.6.3	Unsupervised Approach: Outlier Detection	123
7.6.4	Toward Preventing <i>CAPA</i>	126
7.6.5	Comparison with State of the Art	127
7.7	Conclusions and Future Works	127
8	SOCIAL HONEYPOT FOR HUMANS: LURING PEOPLE THROUGH SELF-MANAGED INSTAGRAM PAGES	129

8.1	Related Works	130
8.2	Methodology	132
8.2.1	Overview & Motivation	132
8.2.2	Topic Selection	132
8.2.3	Post Generation Strategies	133
8.2.4	Engagement Plans	135
8.3	Implementation	136
8.3.1	Topic Selection	136
8.3.2	Testbed	136
8.4	Honeypots Evaluation	137
8.4.1	Overall Performance	137
8.4.2	Honeypot Trends Analysis	138
8.4.3	The Impact of Honeypots Configuration	140
8.4.4	Baseline Comparison	141
8.5	Social Analyses	143
8.5.1	Comments analysis	143
8.5.2	Followers analysis	144
8.5.3	Reached Audience	144
8.6	Toward a Real Implementation	145
8.6.1	Use Cases	145
8.6.2	Challenges and Limitations	146
8.7	Conclusions	147
8.7.1	Implementation details	149
8.7.2	Models	149
8.7.3	Sponsored Content Analyses	151

III Security and Privacy Concerns in Modern Social Platforms I 53

9	PvP: PROFILING VERSUS PLAYER! A FRAMEWORK FOR USER IDENTIFICATION IN ONLINE VIDEO GAMES	I 55
9.1	Related Work	157
9.1.1	Video games related works	157
9.1.2	Dota 2 Related Works	158
9.1.3	CS: GO related works	159
9.2	Our Framework: PvP (Profiling vs Player)	159
9.2.1	Overview and Goal of the Framework	160
9.2.2	Data acquisition phase	160
9.2.3	Match Engineering Phase	161
9.2.4	Identification Phase	162
9.3	Testing The Framework: Games Selection	163
9.3.1	Online Gaming Panorama	163
9.3.2	Dota 2	164
9.3.3	Counter-Strike: Global Offensive	164
9.3.4	Considerations on the Selected Games	166
9.4	Data Collection	166
9.4.1	Dota 2	166
9.4.2	CS: GO data collection	168

9.5	Framework Testing on Dota 2 and CS: GO	170
9.5.1	Dota 2	170
9.5.2	CS: GO	173
9.6	Further experiments	175
9.6.1	General features evaluation	175
9.6.2	Sequence length evaluation	176
9.6.3	Sequence picking interval evaluation	176
9.6.4	Game-specific additional case studies	177
9.6.5	Overall comparison of the two games using our framework	180
9.7	Discussions	180
9.7.1	Real world implications	180
9.7.2	Applicability to other games	182
9.8	Conclusions and Future Work	183
10	ATTRIBUTE INFERENCE ATTACKS IN ONLINE MULTIPLAYER VIDEO GAMES: A CASE STUDY ON DOTA 2	185
10.1	Background and Related Work	187
10.1.1	The Competitive Video-Game Ecosystem	187
10.1.2	Attribute Inference Attacks	189
10.2	DOTA2 Attribute Inference Attacks	190
10.2.1	Proposed Threat Model	190
10.2.2	Feasibility of AIA in DOTA2	191
10.3	Preliminary Assessment	193
10.3.1	Collection of personal attributes (survey)	193
10.3.2	Collection of in-game statistics (TW)	195
10.3.3	Correlation between DOTA2 in-game statistics and personal attributes	196
10.4	Proactive Evaluation of AIA in DOTA2	197
10.4.1	Simple AIA (aggregated player data)	199
10.4.2	One-match AIA (ablation study)	200
10.4.3	Sophisticated AIA	201
10.4.4	Reflection: AIA in research and in practice	202
10.5	Practical AIA (The true threat)	203
10.5.1	Indiscriminate ‘many-to-many’ AIA	203
10.5.2	Targeted ‘many-to-one’ AIA	204
10.6	Discussion	206
10.6.1	Countermeasures to AIA in DOTA2	206
10.6.2	Extension to other E-Sports	207
10.7	Conclusion	208
11	YOU CAN’T HIDE BEHIND YOUR HEADSET: USER PROFILING IN AUGMENTED AND VIRTUAL REALITY	209
11.1	Background & Related Work	211
11.1.1	XR use-cases and benefits of profiling	211
11.1.2	Privacy in XR Technologies	212
11.1.3	Users Profiling in AR and VR Applications	213
11.2	Methodology	214
11.2.1	Scope of the work	214
11.2.2	Inference Framework Overview	215
11.2.3	Framework Detailed Description	215

11.3	Dataset overview	217
11.3.1	AR experiment	217
11.3.2	VR experiment	219
11.3.3	Ethics	220
11.4	Experimental setting	220
11.4.1	Profiling Targets	220
11.4.2	Implementation	221
11.5	Results	222
11.5.1	Task-Level	223
11.5.2	Action-Level	226
11.5.3	Sensors Relevance - Ablation Study	228
11.6	Discussion	230
11.6.1	Limitations	232
11.7	Conclusion	232
12	CONCLUSION AND FUTURE WORK	233
	REFERENCES	235
	ACKNOWLEDGMENTS	273

Listing of figures

2.1	Timeline of significant events related to virtual characters.	14
2.2	Examples of virtual characters.	15
2.3	Visual aspect of the Top 7 Virtual Influencers.	16
2.4	Virtual Influencers' Instagram posts.	21
3.1	Engaging and not engaging posts for each category.	26
3.2	Box plots of Likes and Comments for the different categories and tiers. Note that the y-axes have two different scales, giving a lower number of comments in general.	30
3.3	Top-3 features per absolute correlation value (ρ -value < 0.001) in comments engagement for each category.	33
3.4	Example of guidelines generated by the decision tree for category Beauty, tier nano, likes engagement. The representation is limited at a maximum depth of 3.	36
3.5	Percentage of pure neighborhoods in mid-tier for engagement metrics.	39
3.6	Types of topics for Engagement and User Diversity.	40
3.7	Examples of hot topics found in our categories.	41
4.1	Box-plots of the total number of likes per tier.	50
4.2	Box-plots of the average number of likes per tier.	51
4.3	Box-plots of the average number of comments per tier.	51
4.4	Box-plots of the average number of views per tier.	52
4.5	Box-plots of the total number of videos per tier.	52
4.6	Box-plots of the average inter-posting time per tier.	53
4.7	Box-plots of the number of users an influencer follows per tier.	53
4.8	Box-plots of the number of videos liked per tier.	54
4.9	Presence of bio link info.	54
4.10	Box-plots of the average percentage of videos an influencer post per tier per week of the day.	55
4.11	Classification results to predict the influencer tier using several classifiers.	55
4.12	Feature importance per Tier - All features.	58
4.13	Feature importance - All actionable features.	60
4.14	Feature importance - Traditional Actionable features.	61
4.15	Feature importance - Audio Actionable features.	62
4.16	Feature importance - Video Actionable features.	63
4.17	Feature importance - Text Actionable features.	63
5.1	Bot response to an Italian politician expressing a strong-sided opinion regarding the conflict.	66
5.2	Word Clouds for the tweets of parties captured.	69
5.3	Temporal trends for the war-related tweets, 15 days aggregation.	71
5.4	Number of shared bots between profiles belonging to the same coalition. Colors are representative of the parties, according to the Italian press.	74
5.5	Comparison between "Spider Graphs" of the Mixed and No-Bots Scenario in the Center-Left coalition.	76

5.6	Comparison between “Spider Graphs” of the Complete and No Bots Scenario in the Center-Right coalition.	77
5.7	Mean number of posts and mean posting time for war-related tweets in the last month of Italian elections. Data are reported for both real accounts and bots.	78
6.1	Example of fake vs crowdturfing profiles.	84
6.2	Self-Training process. Dashed arrows represent the training cycle.	89
6.3	Logistic Regression weights to discriminate Crowdturfing (positive label) vs Real (negative label) profiles.	92
6.4	Followers and following avg and std of CT users (Fake) and different categories of real users. Y-axis is in log scale.	93
6.5	Distribution of fake accounts’ following.	93
6.6	Categories of External URLs of the fake profiles.	95
6.7	Results provided by the fraud prevention and detection service on the URLs in the “Other” category.	96
6.8	Most used words by fake and real users.	98
6.9	Languages detected in comments.	99
7.1	Instagram alert of sensitive content.	104
7.2	Example of memes with different obfuscations (e.g., typos, letters-shaped objects, hard background).	104
7.3	Overview of a content moderator in the text and image domains.	107
7.4	Representative samples with obfuscation techniques we identified in online social networks. Blue boxes represent the ACM component that might fails. Green boxes represent different obfuscation techniques.	109
7.5	Overview of CC-CAPA execution pipeline.	113
7.6	Captchas’ styles used in the experiments.	114
7.7	User-study performance distribution. We report accuracy (the higher, the better), and CER (the lower, the better).	116
7.8	Cross-domain evaluation. On the left, the Attack Success Rate (ASR). On the right, the average Normalized Levenshtein Distance (NLD). For both measures, the higher, the more successful the attack.	119
7.9	T-SNE 2D visualization of 2000 samples benign (Pinterest) and 2000 captchas (C111 and CAPA).	124
7.10	F1-score of different Outlier Detection at the varying of the OSN.	125
7.11	Accuracy of different Outlier Detection on known captcha styles at the varying of the OSN.	125
7.12	Accuracy of different Outlier Detection on unknown captcha styles at the varying of the OSN.	125
7.13	Example of false positive found among the outliers that should be moderated.	126
8.1	Pipeline overview to create a social honeypot. After the owner decides on the topic, generation strategy, and engagement plan, the honeypot automatically generates posts to interact with the social network. After the post is automatically generated, the owner can approve it or request a new one to meet the desired quality.	132
8.2	Overview of Post Generation strategies.	133
8.3	Likes trend of our honeypots grouped by engagement plan.	139
8.4	Distribution of likes at the varying of topic, model generation strategy, and engagement plan.	140
8.5	Baseline comparison (average likes) with PLAN1 social honeypots.	142
8.6	Baseline comparison (followers) with PLAN1 social honeypots.	142
9.1	Schema of the proposed framework	159

9.2	An example of a game map. The two red areas, labeled A and B, are the two possible areas in which the bomb can be planted, and the two green areas are the spawn areas for the team. In this case, the upper one (nearer to the planting areas for the bombs) is the spawn for counter-terrorists, while the other one is for the terrorists' team.	165
9.3	Training and validation accuracy and loss in Dota 2 identification model.	172
9.4	Training and validation accuracy and loss in CS: GO identification model.	175
9.5	Identification accuracy for different sequence lengths.	176
9.6	Accuracy with sequences picked at different 2-minute time windows.	177
9.8	Confusion Matrix with Unknown Players involved (class 50).	179
10.1	A TW, showing the statistics of the professional DOTA2 player "Dendi" [1]. All such information is constantly updated and publicly accessible: https://dotabuff.com/players/70388657	186
10.2	The E-Sport ecosystem. <i>Players</i> engage in <i>matches</i> of a <i>video-game</i> , which publicly releases data on such matches. These data are collected by <i>tracking websites</i> , whose elaborations are made public.	188
10.3	Overview of our proposed AIA against DOTA2 players. Public information is used to infer personal (private) attributes.	192
10.4	Top-3 Spearman's ρ between \mathcal{P} and \mathcal{A} (at $p < 0.01$). Higher absolute values denote stronger correlation, while the sign indicates the direction of the correlation.	198
10.5	Impact of Sophisticated AIA. We post-process the predictions of the ML model over multiple matches of the same targeted player.	202
10.6	Targeted 'many-to-one' AIA. We train our ML models by maximizing the <i>precision</i> on a single targeted class. Such AIA are very effective after analyzing ~ 10 matches for each player in the test-set.	206
11.1	Overview of the proposed framework for user profiling in Augmented and Virtual Reality.	214
11.2	Virtual environments adopted in the experiments.	218
11.3	User Identification on task-level.	224
11.4	Age profiling on task-level.	225
11.5	Gender profiling on task-level.	225

Listing of tables

2.1	The Top 7 Virtual Influencers.	15
2.2	Top 3 Real vs. Virtual Influencers.	19
3.1	N. posts (influencers) for categories and tiers.	29
3.2	Performance of Decision Trees (DT) against a dummy classifier (Dum.). In bold, the best scores for likes and comments for each category. Values reported are F1-Score, macro-weighted <i>mean</i> \pm <i>std.</i>	34
3.3	Comparison of Mean F1-Score between our model (DT) and baselines in predicting Likes. <u>Underlined results</u> are statistically significantly higher (two-tailed p -value < 0.05) than the second-best.	37
3.4	Features importance of category Fashion, tier micro.	37
4.1	Category wise list of features used. Actionable features are highlighted with a “*” symbol.	49
4.2	F1-Score of XGB Classifier using top N Features.	57
4.3	Impact, in percentage, of feature groups when considering all features - Increase of F-1 Score.	58
4.4	Impact, in percentage, of feature groups when considering actionable features - Increase of F-1 Score.	59
5.1	Complete overview of the dataset.	68
5.2	Top-7 topics and the number of tweets for each party.	70
5.3	Percentages of bots and non-bots for each profile.	73
5.4	Categories of bots distribution replying to the tweets of the leaders.	73
6.1	Characteristics of Crowdturfing providers. The table reports information claimed by the provider and retrieved by analyzing 100 profiles bought from each. The last row reports info on real profiles for comparison.	86
6.2	Final set of features of our dataset.	88
6.3	Average \pm std of classification results of the best models during cross validation. Sup = Supervised.	90
6.4	Baseline comparison in detecting CT profiles.	91
6.5	Emoji-based Stylometric analysis. CE = Comments with Emoji, EPC = Avg Emoji per comment.	97
6.6	Top-10 topics extracted from fake and real comments.	99
7.1	Percentage of obfuscation techniques observed in different Instagram sources.	111
7.2	List of transformations for textual captchas variants.	114
7.3	Captcha schemes used in our experiment coming from [2].	121
7.4	Datasets’ statistics.	122
7.5	Avg retrieval results of 11 captcha schemes in different OSNs.	122
7.6	Percentage of \mathbb{D}_{tox}^m captchas detected by models trained on data coming from different OSNs.	123
8.1	Honeypots deployed.	138

8.2	Honeypots overall performance. On the left side, we report the average (and std) engagement generated by the honeypots. On the right, we report the number of honeypots with a non-stationary trend. The results are reported based on the topic, generation strategy, and engagement plan.	139
8.3	Percentage of real people, pages, and bots for the best social honeypot in each topic.	144
8.4	Overview of the sponsored content attracted users	151
9.1	Table representing the most played online video games. The monthly players refer to May 2023.	164
9.2	Initial set of Dota 2 features considered for the Identification task.	169
9.3	List of parsed features for CS: GO	170
9.4	Features we kept after the feature selection process	174
9.5	Test Accuracy using all features and general features for CS: GO and Dota 2 identification.	176
10.1	Personal attributes considered in our study. Our population is of 484 DOTA2 players. The distribution resembles the one in [3].	194
10.2	Data returned by the player and matches OpenDota APIs.	196
10.3	Impact of the <i>simple</i> AIA (based on \mathcal{P}) as measured by the F1-score. Rows report the attributes and columns our ML models (boldface denotes the best model for a given attribute).	200
10.4	Impact of the <i>one-match</i> AIA (F1-score). Columns refer to the ‘naive’ attacker (using \mathcal{M}), ‘expert’ attacker (using $\overline{\mathcal{M}}$), and the Dummy (random guess). The expert attacker is always superior.	201
10.5	Results of prior work on AIA. Cells denote the value of a given ‘Metric’ for each of the attributes considered in this work.	203
10.6	Indiscriminate ‘many-to-many’ AIA (mid column). Compared to the baseline (cf. Fig. 10.5), the accuracy substantially increases.	204
10.7	Overview of E-Sports VG. Numbers are taken from various sources [4, 5, 6, 7, 8].	207
11.1	State of the art overview. Legend: \bullet = AR, \circ = VR, \bullet = AR & VR.	212
11.2	Augmented Reality actions organized per type of action and workload level.	218
11.3	Virtual Reality actions organized per type of action and workload level.	220
11.4	User identification on action-level organized per type of operation and workload level. Random guess at 0.03 for both AR and VR tasks. All the measures in F1-Score.	226
11.5	Age profiling on action-level organized per type of operation and workload level. Random guess at 0.5 for both AR and VR tasks. All the measures in F1-Score.	227
11.6	Gender profiling on action-level organized per type of operation and workload level. Random guess at 0.5 for both AR and VR tasks. All the measures in F1-Score.	227
11.7	Ablation study of sensor importance at task-level in AR. All the measures in F1-Score.	229
11.8	Ablation study of sensor importance at task-level in VR. All the measures in F1-Score.	229
11.9	Ablation study of sensor importance at action-level in AR. All the measures in F1-Score.	230
11.10	Ablation study of sensor importance at action-level in VR. All the measures in F1-Score.	231

1

Introduction

In the digital age, society is characterized by unprecedented connectivity facilitated by Online Social Networks (OSNs). These widespread networks have become epicenters of human interaction, enabling individuals to create, share, and disseminate information at an unprecedented scale. Consequently, the vast amount of user-generated data within OSNs has become a valuable resource for researchers, analysts, and practitioners. Social Network Analytics (SNA) has emerged as a powerful tool to extract insights and patterns from social network data, facilitating a better understanding of social structures and dynamics. However, the landscape of OSNs is constantly changing, marked by the emergence of new platforms, shifts in user behavior, and the advent of novel communication modalities. As these networks evolve, SNA methodologies and tools must also adapt, developing innovative approaches for data collection, analysis, and interpretation.

Researchers in diverse domains have leveraged SNA to glean valuable insights. From community detection [9] to sentiment analysis [10], from information diffusion [11] to prevent cybercrime [12], the applications of SNA are as wide-ranging as the networks themselves. Among its myriad topics, SNA focuses on comprehending engagement and influence patterns in these social platforms [13, 14]. Engagement analysis delves into users' interactions with content and other users, for instance, by analyzing metrics such as likes, comments, or views. Conversely, influence analysis concentrates on individuals or entities that have a significant impact on others' opinions, behaviors, or decisions. These influential users, often called "influencers," may possess a large following or exhibit specific characteristics that enable them to shape discussions and trends.

Understanding the dynamics of engagement and influence in recent social networks is crucial. Platforms such as Instagram and TikTok have revolutionized the way individuals communicate, share content, and interact with each other. These modern social networks have become hubs of cultural exchange, shaping opinions, trends, and consumer behavior on a global scale. Therefore, grasping how influence operates within these platforms is essential, not only for individuals seeking to build their online presence but also for businesses and marketers. Companies invest substantial resources in influencer marketing campaigns, with billions of dollars at stake [15]. Identifying the right influencers and accurately predicting the impact of their campaigns are critical for achiev-

ing a return on investment. Similarly, revealing the attributes of trending and engaging content can significantly enhance users' influence and the effectiveness of their campaigns.

On the other hand, understanding the mechanisms of negative influence, including the spread of fake news, misinformation, and the manipulation of public opinion, is essential for combating the growing threat of disinformation [16]. Unfortunately, SNA faces formidable challenges in these adversarial settings. The proliferation of fake profiles, social bots, and the deliberate spread of misinformation introduces noise that can seriously impede the accuracy and reliability of SNA outcomes. In this intricate landscape, distinguishing genuine user behaviors from manipulative activities becomes a challenging task. Accordingly, there is an urgent need to develop innovative methodologies and resilient algorithms to detect and mitigate the impact of these malicious actors.

Finally, the vast amount of data shared on social platforms exposes their users to security and privacy risks. Researchers have dedicated substantial effort to analyzing such threats in traditional OSNs [17, 18]. Nevertheless, the notion of social networks has transcended traditional OSNs, extending its reach into contemporary digital domains like video games and the Metaverse. Within these virtual landscapes, people engage, communicate, and establish connections, creating online communities and fostering social interactions. These platforms often facilitate multiplayer experiences through in-game chats and collaborative events. Furthermore, with the advent of esports and streaming platforms, players and fans unite to watch, discuss, and interact with their favorite games and players. However, as their adoption continued to soar, it did not take long for malicious users to appear, posing novel security and privacy threats. This evolving landscape demands a fresh perspective on privacy protection, exploring novel threats and pioneering countermeasures to ensure that users can engage in these immersive experiences with confidence in their data security.

1.1 RESEARCH MOTIVATION AND CONTRIBUTION

This dissertation advances social network analytics by focusing on three key aspects:

1. *Influence and engagement*: studying of engagement and influence patterns in trending online social networks. Chapter 2 explores the phenomenon of Virtual Influencers. Chapter 3 investigates patterns of Instagram engagement, while Chapter 4 focuses on the characteristics of TikTok influencers. Chapter 5 presents a case study illustrating how bots engage and manipulate information on Twitter.
2. *Resilient Social Network Analytics Tools*: developing algorithms to enhance Social Network Analytics, particularly in challenging or adversarial contexts. Chapter 6 presents a novel tool to uncover Instagram crowdturfing. Chapter 7 analyzes obfuscations implemented by OSN users to spread adversarial content, proposing a detection mechanism. Chapter 8 introduces Social Honey pots, an SNA tool for effectively studying communities and topics on Instagram.
3. *Security and Privacy Concerns in Modern Social Platforms*: investigating novel cybersecurity threats within social platforms like video games and the Metaverse. Chapter 9 introduces PvP, a user identification framework for online video games to track cybercriminals and mitigate harmful behaviors. Chapter 10 presents an Attribute Inference Attack in Dota 2, underscoring a subtle privacy threat affecting millions of video gamers. Chapter 11 proposes a user profiling framework for Augmented and Virtual Reality, the enabling

technologies of the Metaverse, improving the user experience while enhancing the platforms' security and privacy.

Now, a concise overview of each chapter, including its contributions, will be presented. In this dissertation, some passages have been quoted verbatim, and some figures have been reused from the works [19, 20, 21, 22, 23, 24, 25, 26], all coauthored by the author of the thesis.

VIRTUAL INFLUENCERS IN ONLINE SOCIAL MEDIA

Influencers are people on social media who distinguish themselves by the high number of followers and the ability to influence other users. While Influencers are a long-standing phenomenon in social media, Virtual Influencers have made their appearance on such platforms only recently: they are Computer-Generated Imagery (CGI) characters that act and resemble humans, even if they do not physically exist in the real world. This recent phenomenon has sparked interest in society, and several questions arise regarding their evolution, opinions, ethics, purpose in marketing, and future perspective.

CONTRIBUTION In Chapter 2, we conduct an exhaustive review of the virtual influencer phenomenon. Through an extensive study of the literature, press articles, social platforms data, blogs, and interviews, we give a comprehensive reflection on Virtual Influencers. Starting from their evolution, we analyze their opportunities and threats. We provide detailed information about the most popular ones and their marketing collaborations, with a comparative analysis of virtual and real (human) influencers. Moreover, we conducted an online survey to grasp people's perspectives. From the 360 participants' answers, we draw conclusions about Virtual Influencers' ethics, importance, overall feelings, and future. Results show controversial opinions on this recent phenomenon.

FOLLOW US AND BECOME FAMOUS! INSIGHTS AND GUIDELINES FROM INSTAGRAM ENGAGEMENT MECHANISMS

With 1.3 billion users, Instagram (IG) has become an essential business tool. IG influencer marketing, expected to generate \$33.25 billion in 2022, encourages companies and influencers to create trending content. Various methods have been proposed for predicting a post's popularity, i.e., how much engagement (e.g., Likes) it will generate. However, these methods are limited: first, they focus on forecasting the likes, ignoring the number of comments, which became crucial in 2021. Secondly, studies often use biased or limited data. Third, researchers focused on Deep Learning models to increase predictive performance, which are difficult to interpret. As a result, end-users can only estimate engagement after a post is created, which is inefficient and expensive. A better approach is to generate a post based on what people and IG like, e.g., by following guidelines.

CONTRIBUTION In Chapter 3, we uncover part of the underlying mechanisms driving IG engagement. We rely on statistical analysis and interpretable models rather than Deep Learning (black-box) approaches to achieve this goal. Leveraging innovative domain-relevant features, we first build classifiers to predict posts' engagement. Then, we interpret the best models to determine which type of content will generate the most engagement, maximizing influencers' and companies' profits. We conduct extensive experiments using a worldwide dataset of 10 million

posts created by 3.4K global influencers in nine different categories. Our simple yet powerful algorithms can effectively predict engagement, making us comparable and even superior to Deep Learning-based methods, reaching up to 94% F1-Score. Furthermore, we propose a novel unsupervised algorithm for finding highly engaging topics on IG. Thanks to our interpretable approaches, we conclude by outlining guidelines for creating successful posts.

CLIMBING THE INFLUENCE TIERS ON TIKTOK: A MULTIMODAL STUDY

Unlike most social media platforms, TikTok remunerates content creators based on their video views, motivating them to create highly engaging content and strive to expand their audience reach. Corporate social media analysts categorize influencers into five escalating tiers of significance, based on their number of followers: Nano, Micro, Mid, Macro, and Mega influencers. In addition to earnings from TikTok’s remuneration system, influencers frequently receive direct marketing opportunities from companies, with their compensation scaling up in accordance with their tier. Therefore, influencers are strongly incentivized to ascend from their current tier to the next higher one.

CONTRIBUTION In Chapter 4, we perform a comprehensive study of TikTok influencers in these five tiers with two goals: (i) what factors distinguish influencers in each of these five tiers from the adjacent tier(s)? (ii) out of the features influencers could directly control (actionable feature), which ones are more impactful to reach the next tier? We build and release a novel TikTok dataset consisting of over 230K videos (published by 5000 influencers—1000 from each tier) and corresponding video information, ranging from the number of likes to facial action units, people’s emotions, or music information (taken from Spotify). To find the most important features for distinguishing influencers in a given tier from those in the tier directly above, we perform a thorough analysis of traditional features as well as text, audio, and video features using both statistically valid hypotheses and ablation testing. Through our classifiers achieving an F1-score of over 80%, we identified the most impactful actionable features within traditional characteristics, e.g., increasing the posting frequency or refining profile information, as well as within video-related attributes, including enhancing video pleasure, quality, and emphasizing facial expressions.

TWITTER BOTS INFLUENCE ON THE RUSSO-UKRAINIAN WAR DURING THE 2022 ITALIAN GENERAL ELECTIONS

In February 2022, Russia launched a full-scale invasion of Ukraine. This event had global repercussions, especially on the political decisions of European countries. As expected, the role of Italy in the conflict became a major campaign issue for the Italian General Election held on 25 September 2022. Politicians frequently use Twitter to communicate during political campaigns, but bots often interfere and attempt to manipulate elections. Hence, understanding whether bots influenced public opinion regarding the conflict and, therefore, the elections is essential.

CONTRIBUTION In Chapter 5, we investigate how Italian politics responded to the Russo-Ukrainian conflict on Twitter and whether bots manipulated public opinion before the 2022 general election. We first analyze 39,611

tweets of six major political Italian parties to understand how they discussed the war during the period February-December 2022. Then, we focus on the 360,823 comments under the last month’s posts before the elections, discovering around 12% of the commenters are bots. By examining their activities, it becomes clear they both distorted how war topics were treated and influenced real users during the last month before the elections.

ARE WE ALL IN A TRUMAN SHOW? SPOTTING INSTAGRAM CROWDTURFING THROUGH SELF-TRAINING

Influencer Marketing generated \$16 billion in 2022. Usually, the more popular influencers are paid more for their collaborations. Thus, many services were created to boost profiles’ popularity metrics through bots or fake accounts. However, real people recently started participating in such boosting activities using their real accounts for monetary rewards, generating unguanine content that is extremely difficult to detect. To date, no works have attempted to detect this new phenomenon, known as crowdturfing (CT), on Instagram.

CONTRIBUTION In Chapter 6, we propose the first Instagram CT engagement detector. Our algorithm leverages profiles’ characteristics through semi-supervised learning to spot accounts involved in CT activities. Compared to the supervised approaches used so far to identify fake accounts, semi-supervised models can exploit huge quantities of unlabeled data to increase performance. We purchased and studied 1293 CT profiles from 11 providers to build our self-training classifier, which reached 95% F1-score. We tested our model in the wild by detecting and analyzing CT engagement from 20 mega-influencers (i.e., with more than one million followers), and discovered that more than 20% was artificial. We analyzed the CT profiles and comments, showing that it is difficult to detect these activities based solely on their generated content.

TURNING CAPTCHAS AGAINST HUMANITY: CAPTCHA-BASED ATTACKS IN ONLINE SOCIAL MEDIA

Nowadays, people generate and share massive amounts of content on online platforms (e.g., social networks, blogs). In 2021, the 1.9 billion daily active Facebook users posted around 150 thousand photos every minute. Content moderators constantly monitor these online platforms to prevent the spreading of inappropriate content (e.g., hate speech, nudity images). Based on deep learning (DL) advances, Automatic Content Moderators (ACM) help human moderators handle high data volume. Despite their advantages, attackers can exploit weaknesses of DL components (e.g., preprocessing, model) to affect their performance. Therefore, an attacker can leverage such techniques to spread inappropriate content by evading ACM.

CONTRIBUTION In Chapter 7, we analyzed 4600 potentially toxic Instagram posts, and we discovered that 44% of them adopt obfuscations that might undermine ACM. As these posts are reminiscent of captchas (i.e., not understandable by automated mechanisms), we coin this threat as Captcha Attack (**CAPA**). Our contributions start by proposing a **CAPA** taxonomy to better understand how ACM is vulnerable to obfuscation attacks. We then focus on the broad sub-category of **CAPA** using textual Captcha Challenges, namely CC-CAPA, and we empirically demonstrate that it evades real-world ACM (i.e., Amazon, Google, Microsoft) with 100% accuracy. Our investigation revealed that ACM failures are caused by the OCR text extraction phase. The training of OCRs to with-

stand such obfuscation is therefore crucial, but huge amounts of data are required. Thus, we investigate methods to identify CC-CAPA samples from large sets of data (originated by three OSN – Pinterest, Twitter, Yahoo-Flickr), and we empirically demonstrate that supervised techniques identify target styles of samples almost perfectly. Un-supervised solutions, on the other hand, represent a solid methodology for inspecting uncommon data to detect new obfuscation techniques.

SOCIAL HONEYPOT FOR HUMANS: LURING PEOPLE THROUGH SELF-MANAGED INSTAGRAM PAGES

Social Honey Pots are tools deployed in Online Social Networks (OSN) to attract malevolent activities performed by spammers and bots. To this end, their content is designed to be of maximum interest to malicious users. However, by choosing an appropriate content topic, this attractive mechanism could be extended to *any* OSN users, rather than only luring malicious actors. As a result, honeypots can be used to attract individuals interested in a wide range of topics, from sports and hobbies to more sensitive subjects like political views and conspiracies. With all these individuals gathered in one place, honeypot owners can conduct many analyses, from social to marketing studies.

CONTRIBUTION In Chapter 8, we introduce a novel concept of social honeypot for attracting OSN users interested in a generic target topic. We propose a framework based on fully-automated content generation strategies and engagement plans to mimic legit Instagram pages. To validate our framework, we created 21 self-managed social honeypots (i.e., pages) on Instagram, covering three topics, four content generation strategies, and three engaging plans. In nine weeks, our honeypots gathered a total of 753 followers, 5387 comments, and 15739 likes. These results demonstrate the validity of our approach, and through statistical analysis, we examine the characteristics of effective social honeypots.

PvP: PROFILING VERSUS PLAYER! A USER IDENTIFICATION FRAMEWORK FOR ONLINE VIDEO GAMES

The rapid proliferation of online video games has opened up many avenues for fraudulent activities. In-game purchases and one-click payments have led to a significant rise in scams and account takeovers, impacting millions of gamers. Prominent security breaches and user information leaks, such as those affecting major corporations like Steam, Nintendo, or Bandai Namco, and games like League of Legends and Fortnite, highlight the gravity of the issue. Moreover, the gaming community continues to struggle with problems such as cyberbullying, grooming, and harassment despite efforts to identify and ban malicious actors. Indeed, these individuals routinely create new accounts to perpetuate their malevolent activities. All these issues could be mitigated through the capability to uniquely identify a player, regardless of the account they employ. Much like a fingerprint in the real world, a virtual fingerprint could enable the identification and subsequent banning of malicious actors across all their existing or newly created profiles.

CONTRIBUTION In Chapter 9, we present a novel player identification framework called PvP (Profiling vs Player). PvP introduces and leverages the concept of a “video game fingerprint” derived from a player’s unique

play style and interaction with the digital world. The framework extracts game-related features, aggregates in-game data, and employs deep learning techniques to identify and distinguish players. We thoroughly tested PvP on data from 50 Dota 2 and 50 CS: GO players, encompassing 10,000 matches. Using only two minutes of gaming data, PvP achieved over 90% accuracy in both games. Notably, Dota 2 and CS: GO represent diverse gaming genres, underscoring PvP's versatility. While PvP holds promise for enhancing player identification and thwarting malicious activities, it also raises awareness of potential vulnerabilities. Victims seeking to evade tormentors (e.g., cyberbullies) may face themselves pursued through the analysis of their play styles. Therefore, PvP not only establishes the feasibility of player identification but also raises awareness about a subtle threat that may already be impacting millions of users in the gaming community.

ATTRIBUTE INFERENCE ATTACKS IN ONLINE MULTIPLAYER VIDEO GAMES: A CASE STUDY ON DOTA2

The rapid expansion of esports, along with their substantial prize pools, serves as a strong incentive for millions of gamers to compete and enhance their skills in pursuit of becoming professional players. Tracking websites have emerged to assist gamers in analyzing their performance and drawing insights from fellow players. In this ecosystem, the norm is for data to be publicly accessible, exemplified by over 70 million DOTA2 players freely sharing their in-game data. However, this situation raises the question: What if such data were exploited for malicious purposes? We are pioneering the investigation of this critical issue.

CONTRIBUTION In Chapter 10, motivated by the widespread popularity of video games, we propose the first threat model for Attribute Inference Attacks (AIA) in the DOTA2 context. We explain *how* (and *why*) attackers can exploit the abundant public data in the DOTA2 ecosystem to infer private information about its players. Due to lack of concrete evidence on the efficacy of our AIA, we empirically prove and assess their impact in reality. By conducting an extensive survey on ~ 500 DOTA2 players spanning over 26k matches, we verify whether a correlation exists between a player's DOTA2 activity and their real-life. Then, after finding such a link ($p < 0.01$ and $\rho > 0.3$), we ethically perform diverse AIA. We leverage the capabilities of machine learning to infer real-life attributes of the respondents of our survey by using their publicly available in-game data. Our results show that, by applying domain expertise, some AIA can reach up to 98% precision and over 90% accuracy. This chapter hence raises the alarm on a subtle, but concrete threat that can potentially affect the entire competitive gaming landscape. We alerted the developers of DOTA2.

YOU CAN'T HIDE BEHIND YOUR HEADSET: USER PROFILING IN AUGMENTED AND VIRTUAL REALITY

Augmented and Virtual Reality (AR, VR), collectively known as Extended Reality (XR), are increasingly gaining traction thanks to their technical advancement and the need for remote connections, recently accentuated by the pandemic. Remote surgery, telerobotics, and virtual offices are only some examples of their successes. As users interact with XR, they generate extensive behavioral data usually leveraged for measuring human activity, which could be used for profiling users' identities or personal information (e.g., gender). However, several factors affect the efficiency of profiling, such as the technology employed, the action taken, the mental workload, the presence

of bias, and the sensors available. To date, no study has considered all of these factors together and in their entirety, limiting the current understanding of XR profiling.

CONTRIBUTION In Chapter 11, we provide a comprehensive study on user profiling in virtual technologies (i.e., AR, VR). Specifically, we employ machine learning on behavioral data (i.e., head, controllers, and eye data) to identify users and infer their individual attributes (i.e., age, gender). Toward this end, we propose a general framework that can potentially infer any personal information from any virtual scenarios. We test our framework on eleven generic actions (e.g., walking, searching, pointing) involving low and high mental loads, derived from two distinct use cases: an AR everyday application (34 participants) and VR robot teleoperation (35 participants). Our framework limits the burden of creating technology- and action-dependent algorithms, also reducing the experimental bias evidenced in previous work, providing a simple (yet effective) baseline for future works. We identified users up to 97% F1-score in VR and 80% in AR. Gender and Age inference was also facilitated in VR, reaching up to 82% and 90% F1-score, respectively. Through an in-depth analysis of sensors' impact, we found VR profiling resulting more effective than AR mainly because of the eye sensors' presence.

1.1.1 PUBLICATIONS

This section summarizes the manuscripts produced during my Ph.D. and published or submitted in peer-reviewed journals and conferences. All the manuscripts are listed in chronological order of acceptance and submission.

JOURNAL PUBLICATIONS

- Conti, M., Gathani, J., & Tricomi, P. P. (2022). Virtual influencers in online social media. *IEEE Communications Magazine*, 60(8), 86-91. (Q1, JCR IF 2022: 11.2) [19].
- Tricomi, P. P., Nenna, F., Pajola, L., Conti, M., & Gamberini, L. (2023). You can't hide behind your headset: User profiling in augmented and virtual reality. *IEEE Access*, 11, 9859-9875. (Q1, JCR IF 2022: 3.9) [26].
- Cardaioli, M., Conti, M., Orazi, G., Tricomi, P. P., & Tsudik, G. (2023). BLUFADER: Blurred face detection & recognition for privacy-friendly continuous authentication. *Pervasive and Mobile Computing*, 92, 101801. (Q1, JCR IF 2022: 3.848) [27].
- Conti, M., Pajola, L., & Tricomi, P. P. (2023). Turning captchas against humanity: Captcha-based attacks in online social media. *Online Social Networks and Media*, 36, 100252. (Q1, Scopus IF 2022: 4.419) [23].
- Mondini, S., Pucci, V., Pastore, M., Gaggi, O., Tricomi, P. P., & Nucci, M. (2023). s-CRIq: the online short version of the Cognitive Reserve Index Questionnaire. *Aging Clinical and Experimental Research*. (Q2, Scopus IF 2022: 4.204) [28].
- Conti, M., Kostadinov, S., & Tricomi, P.P. (2023). PvP: Profiling Versus Player! A Framework for User Identification in Online Video Games. *ACM Transactions on Privacy and Security*. (Q1, JCR IF 2022: 2.717.) Submitted.

CONFERENCE PUBLICATIONS

- Conti, M., Vinod, P., & Tricomi, P. P. (2021, October). Secure static content delivery for CDN using blockchain technology. In *International Workshop on Data Privacy Management* (pp. 301-309). **ESORICS 2021 Workshop** [29].
- Cardaioli, M., Conti, M., Tricomi, P. P., & Tsudik, G. (2022, March). Privacy-friendly de-authentication with BLUFADE: Blurred face detection. In *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom 2022)* (pp. 197-206). (**CORE: A++**, **LiveSHINE: A+**, **MA: A**, **Italian GGS: 1/A+**) [30].
- Tricomi, P. P., Facciolo, L., Apruzzese, G., & Conti, M. (2023, April). Attribute inference attacks in online multiplayer video games: A case study on Dota2. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy (CODASPY 2023)* (pp. 27-38). (**LiveSHINE: B**, **MA: C**) [25].
- Tricomi, P. P., Chilese, M., Conti, M., & Sadeghi, A. R. (2023, April). Follow us and become famous! Insights and guidelines from Instagram engagement mechanisms. In *Proceedings of the 15th ACM Web Science Conference 2023 (WebSci 2023)* (pp. 346-356). (**MA: A++**) [20].
- Bardi, S., Conti, M., Pajola, L., & Tricomi, P. P. (2023, May). Social Honey-pot for Humans: Luring People Through Self-managed Instagram Pages. In *International Conference on Applied Cryptography and Network Security (ACNS 2023)* (pp. 309-336). (**CORE: B**, **LiveSHINE: A**, **MA: A-**, **Italian GGS: 2/A-**) [24].
- Tricomi, P. P., Tarahomi, S., Cattai, C., Martini, F., & Conti, M. (2023, July). Are we all in a Truman show? Spotting Instagram crowdturfing through self-training. In *2023 32nd International Conference on Computer Communications and Networks (ICCCN 2023)* (pp. 1-10). (**CORE: B**, **LiveSHINE: A-**, **MA: B**, **Italian GGS: 3/B**) [22].
- De Faveri, F. L., Cosuti, L., Tricomi, P. P., & Conti M., Twitter Bots Influence on the Russo-Ukrainian War During the 2022 Italian General Elections (2023). *The 9th International Symposium on Security and Privacy in Social Networks and Big Data (SocialSec 2023)*. [21].
- Tricomi P. P., Kumar, S., Conti, M., Subrahmanian, V.S., (2023). Climbing the Influence Tiers on TikTok: A Multimodal Study. *The 18th International AAAI Conference on Web and Social Media (ICWSM 2024)*. (**LiveSHINE: A+**, **MA: A+**, **Italian GGS: 2/A**) Submitted

Part I

Influence and Engagement in Trending Online Social Networks

INTRODUCTION TO PART I

The first part of this doctoral thesis delves into analyzing engagement and influence patterns within trending online social networks. The part begins by delving into the emerging phenomenon of virtual influencers, primarily focusing on assessing how they are transforming social platforms, audience engagement, and influencer marketing. The focus then shifts towards Instagram, the leading image-sharing social network, where user engagement mechanisms are uncovered, defining algorithms and guidelines for creating highly engaging content, thereby optimizing profits for influencers and digital marketing efforts. The investigation proceeds to TikTok, the ever-expanding video-sharing platform, exploring the distinctive influence dynamics, focusing on how influencers can enhance their reach through behavior and content creation strategies. Lastly, the thesis explores the tumultuous landscape of Twitter, where pervasive social bot engagement often leads to information manipulation and repercussions for societal discourse.

2

Virtual Influencers in Online Social Media

As Social Media spread among people, companies began to embrace them as advertising tools. Marketing agencies rely on people with a high number of followers and ability to influence the mass (known as Influencers) for advertisements. The usage of visual content on these platforms has increased in the last decade, especially on Instagram, which became an effective way for brands to literally show their products and values. Constant innovation in the influencer marketing industry has led to a new phenomenon called Virtual Influencers (VI). We can describe a virtual influencer as a person or thing created by software that can influence others, primarily through marketing collaborations or participation in social campaigns, and is solely created and consumed via digital mediums [31]. They resemble human characteristics, behavior, and actions but do not correspond to any human in the real world. Companies are not releasing information about the software or technology they use to create VI. However, we expect they are created by 3D artists using CGI (Computer-Generated Imagery) and motion capture technologies to depict them as real people in real situations. Sometimes, VI are digitally-altered versions of real people or a digital combination of a CGI-made head and real person's body. We presume that even content related to VI (e.g., posts), which nowadays is mainly created by humans, will always be more generated by Artificial Intelligence (AI). In the following sections, we will refer to the above definition of virtual influencer.

One recent study stated that people like or comment on virtual influencers' posts three times more than Real (human) Influencers (RI) ones. This trend has followed from 2019 and was consistent in 2020 as well [32]. In the past three to four years, brands from every industry have exploded Instagram with digital avatars, demonstrating their commitment towards innovation and creativity. Examples are Renault, IKEA, Prada, or Samsung. During the difficult time of the COVID-19 pandemic, VI contributed to raise awareness about social distancing and other ways to help prevent COVID-19 from spreading [33]. Although the COVID-19 pandemic decelerated the growth of human influencers worldwide, virtual influencers were not affected. In reality, the COVID-19 crisis probably fueled the expansion of virtual influencer marketing strategy. People are clearly attracted to VI, given the growing trend of companies partnering with them. Viewers probably like the human emotions VI express in daily-life situations, although they are not real. Furthermore, VI can digitally be anywhere at any time, delivering

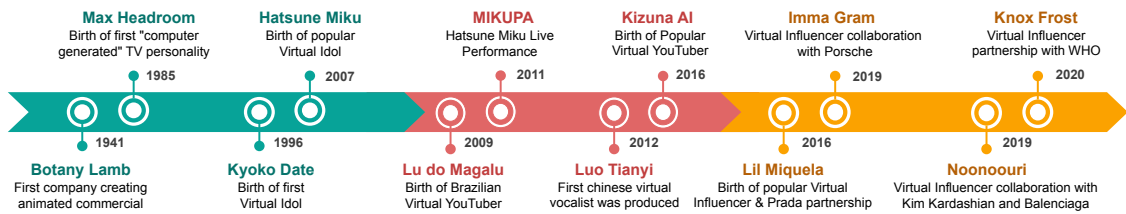


Figure 2.1: Timeline of significant events related to virtual characters.

their followers highly-catchy content.

CONTRIBUTION. We provide a wide overview of the emerging phenomenon of Virtual Influencers under several aspects: evolution, popularity, marketing, ethics, opportunities, and threats. For this purpose, we collected and analyzed several literature articles, online resources, and reports provided by websites specialized in influencers' analysis. Moreover, we carried out a comparative analysis between Real and Virtual Influencers. Finally, we conducted a survey with 360 participants to understand people's views on VI.

STRUCTURE. We first introduce the history and evolution of Virtual Influencers. Then, we present some of the most popular Virtual Influencers, followed by a discussion on their marketing. Next, we analyze the differences between Real and Virtual Influencers. Last, we consider opinions about Virtual Influencers, concluding the article by presenting some future directions.

2.1 VIRTUAL INFLUENCERS TIMELINE

To show the birth and evolution of virtual influencers, we reconstructed a timeline (Fig. 2.1) of significant events related to virtual characters. Figure 2.2 shows examples of them, which we now describe. Although the connection between virtual characters and virtual influencers has not been demonstrated in the literature, their similarity allow us to consider virtual characters as predecessors of Virtual Influencers.

The phenomena of virtual characters existed way back in the early 90s, with cartoon characters being the pioneers. Animation has been used as advertising tool since 1940s, given the high viewer engagement it creates. However, the first virtual celebrities were launched in Japan alongside virtual idols, which are media performances that occur independently of any living performer's referent [34]. The Japanese talent agency HoriPro teamed up with Visual Science Laboratory, a computer graphics company, to create Kyoko Date, the world's first 3D computer-generated female model [35]. Kyoko had released her first CD single, "Love Communication", which was well-received on Japanese radios. Another popular virtual idol is Hatsune Miku (Fig. 2.2). She was considered very well-known in the world of character entertainment [36]. Hatsune Miku, or Miku Hatsune, which translates to "first sound from the future", is a virtual singer brought to life in 2007, developed by Crypton Future Media using Vocaloid, a Yamaha voice synthesizer program. Her popularity skyrocketed, prompting her to record her own music in live concerts like any other pop star.

Table 2.1: The Top 7 Virtual Influencers.

Name	IG profile (URL link)	Followers	Engagement rate	Origin country	Birth date	Creator	Estimated Earnings per Post (EEP)	Brand collaboration
Lu do Magalu	@magazineluiza	5M	0.08%	Brazil	2009	Magazine Luiza	\$10,128–\$16,880	Magazine Luiza
Lil Miquela	@lilmiquela	3M	1.85%	USA	2016	Brud	\$6,056–\$10,093	Calvin Klein, Prada
Knox Frost	@knox frost	800K	1.02%	USA	2019	–	\$2,386–\$3,977	WHO
Thalasya Pov	@thalasya_	495K	0.95%	Indonesia	2018	Magnavem Studio	\$1,474–\$2,457	Chocolatos ID
Imma	@imma.gram	331K	1.61%	Japan	2018	Aww Inc.	\$987–\$1,646	Porsche, IKEA
Bermuda	@bermudaisbae	293K	7.29%	USA	2016	Brud	\$881–\$1,468	Chanel
Shudu	@shudu.gram	215K	3.12%	England	2017	The Digitals Agency	\$645–\$1,075	Balmain

Virtual idols grew rapidly, resulting in the birth of virtual YouTubers in early 2016. A virtual YouTuber, or “Vtuber”, is a fictional character in YouTube videos and live streams. These are 3D models that most commonly exist in the digital form and are typically associated with some voice to provide vocal performances [37]. With 3 million subscribers on YouTube, Kizuna AI (Fig. 2.2) is one of the most famous Vtubers. She has served as a spokeswoman for SoftBank and the Japan National Tourism Organization, hosted offline fan events, performed at music festivals, and was engaged in a talk with Japanese Nobel Prize winners.

In 2016, a relatively new phenomenon known as Virtual influencers emerged, which can be thought of as an evolution of virtual idols and virtual YouTubers. Since most of the virtual characters were used as influencers, this phenomena quickly gained traction on Instagram, which is one of the most effective platforms for influencer marketing [38]. These characters were more appealing than virtual idols because of their realistic human-like appearance. Moreover, they were actively involved in marketing and social campaigns, and thus identified as “influencers”. Lil Miquela (Fig. 2.2), launched in mid-2016, amassed more than 3 million followers on Instagram [39]. Following her success, many other VI were created. In this article, we will mainly focus on Virtual Influencers.

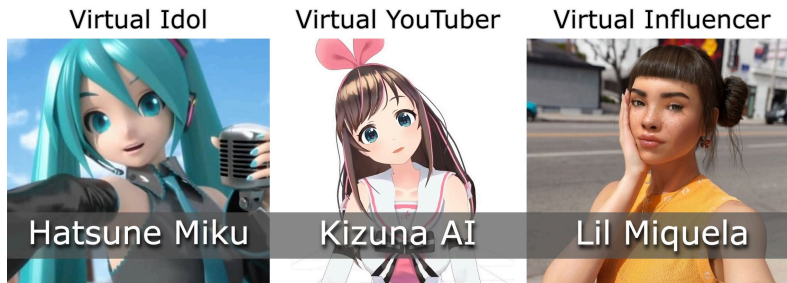


Figure 2.2: Examples of virtual characters.

2.2 POPULAR VIRTUAL INFLUENCERS

Nearly 70 percent of brands uses influencers on Instagram for their marketing campaigns, compared to around 45 percent on TikTok and Facebook [38]. This could be one reason for Virtual Influencers to be highly active on Instagram compared to other social media. Hence, we focus more on Instagram in this study. Table 2.1 provides information of top 7 most popular VI present on Instagram, based on the number of followers and collaborations with famous brands. Figure 2.3 shows all of them. The table is reconstructed using multiple sources on

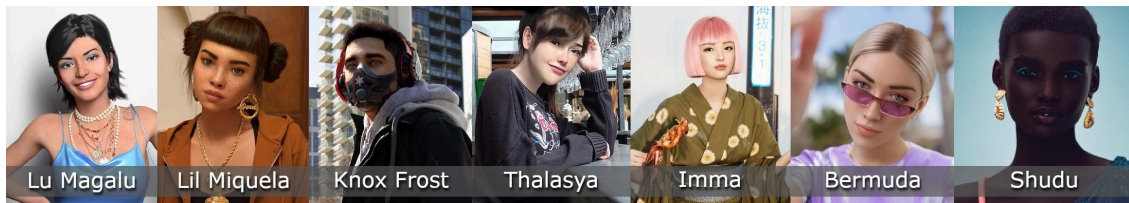


Figure 2.3: Visual aspect of the Top 7 Virtual Influencers.

the Internet. The number of followers and the engagement rate are referred from HyperAuditor, a website that offers a comprehensive Instagram account analysis report. Here, engagement rate refers to the percentage of the audience who likes or comments on the posts. Birth date, origin, creator, and brand collaborations are sourced from VirtualHumans, a website which provides detailed information about VI. Estimated per post earnings of each VI is calculated based on the engagement rate and the number of followers, using an online tool (Instagram Influencer Earnings Calculator). We now analyze the VI reported in Table 2.1.

LU DO MAGALU Lu made her YouTube debut in 2009, promoting iBlogTV on behalf of Magazine Luiza (“Magalu”), one of the largest Brazilian retail firms. Lu has been featuring unboxing videos, product reviews, and technological tips on behalf of the company. Although Lu has a low engagement rate of 0.08 percent, possibly because she is only famous in Brazil, she has a massive audience of 5 million followers on Instagram. She is also famous on Twitter and TikTok. Lu has worked with several fashion companies and supports social causes such as cancer, diversity, and violence against women.

LIL MIQUELA Miquela Sousa is a 19-year-old Brazilian-American influencer who debuted on Instagram in 2016, with more than 3 million followers [39]. She is a Computer-Generated Imagery (CGI) character developed by Brud, a Los Angeles-based company. She describes herself as a “musician, change-seeker, and drip robot”. Lil Miquela has collaborated with American music producer Baaer on the “Hate Me” album, and her Spotify page have gained a huge amount of monthly listeners. She supports social issues such as Black Lives Matter and transgender rights. *The Time* magazine named her one of the Internet’s 25 most influential people in 2018, alongside Donald Trump and Kanye West.

KNOX FROST Knox Frost, a 20-year-old “guy” from Atlanta with over 800k Instagram followers, is a top male VI. His content over Instagram sparks vibrant discussions in the comments sections of his posts. He has often provided his advice in supporting some social matters like self-empowerment and mental health. He also collaborated with the World Health Organization (WHO) on the COVID-19 public awareness and fundraising campaign.

THALASYA POV Thalasya Pov was brought to life in Indonesia in 2018. Since then, she has nearly hit around 500k followers. Magnavem Studio developed her, and she now owns a clothing store (Yipiiii). She is dressed as a typical Indonesian influencer, often sharing photos of herself in cafes and tourist attractions. She has also partnered with real influencers such as Gilang Dirga and Raditya Dika, and has been a Chocولاتos brand ambassador, sporadically sharing pictures of herself enjoying the snack.

IMMA Imma was developed by Aww Inc., a Japanese startup that produces virtual influencers. She was featured on the cover of *CGWorld* magazine and has gained more than 300k followers. Imma has partnered with several famous companies to promote their products and services. She is well-known for her edgy street-style images with very catchy expressions and poses. She looked so real in a photograph alongside two other actual human models in a makeup spread for Kate cosmetics, that it was impossible to tell she was a virtual character. She also supports Black Lives Matter.

BERMUDA Bermuda was brought to life in 2016 in Los Angeles. She is another creation of Brud. She has the highest engagement rate of 7.29 percent among all VI and aims to inspire more women to pursue careers in robotics. She is also a rapper with tracks available on Spotify, and is also known as “The Most Controversial” VI. Bermuda once hacked Miquela’s account, gaining both of them more followers and driving Miquela’s account past the million-follower mark. This attack was assumed to be a marketing strategy by Brud to gain attention.

SHUDU Shudu is the world’s first digital supermodel created by British photographer Cameron-James Wilson, the founder of The Diigitals Agency. She has over 200,000 followers on Instagram with a significantly higher engagement rate of 3.12 percent as compared to some of the most popular virtual influencers. Shudu has also landed some major brand collaborations.

2.3 THE INFLUENCER MARKETING

Influencer marketing is a phenomenon where companies approach famous or high-influence people on social media for a brand or product endorsements, extending the “word of mouth” marketing strategy. Influencers are content creators who have built their own personal brand image both online and offline, and are able to drive people’s purchasing decisions. The influencer marketing industry is expected to rise in value to \$13.8 billion by 2021, which is almost ten times to what it was in 2016 [38].

Big companies and major brands are moving towards digitalization by creating or partnering with virtual influencers. Lil Miquela, one of the most famous virtual influencers, has collaborated with companies such as Prada and Calvin Klein, alone or alongside human influencers. Her estimated earnings per post (EEP) range from \$6,056 to \$10,093, indicating her enormous success and popularity. Lu, who is the spokesperson of Magazine Luiza, has also collaborated with the fashion store Zattini for their winter collection clothes. Her EEP of \$10,000 is probably higher than those of many human influencers. Shudu, the world’s first digital model, has worked for famous brands like Ellesse and the high-end luxury fashion house Balmain, with an average EEP of about \$700 – \$1000. Even during the COVID-19 pandemic, where the whole world (and most human influencers) was at a stand-still, some big brands collaborated with virtual influencers. The WHO had partnered with Knox Frost to disseminate best practices against COVID-19. He supported the COVID-19 fundraising campaign in his Instagram feed, by also including a link to the WHO’s donation page [33]. Another big “virtual” collaboration during the COVID-19 pandemic involved IKEA. They inaugurated a new store in Tokyo, with the help of Imma. Imma has also collaborated with various well-known firms, including Magnum, Porsche, Amazon Fashion.

The above collaborations are just examples of how companies are moving toward “virtual” partnerships. Due to VI flexibility and increasing popularity, we expect to see more collaborations in the future.

2.4 VIRTUAL VS. REAL INFLUENCERS

In this section, we present some opportunities and threats of using VI, a comparative analysis between Virtual Influencers and Real (human) Influencers (RI), and VI ontology and ethics.

2.4.1 OPPORTUNITIES OF VIRTUAL INFLUENCERS

MORE FLEXIBILITY Virtual influencers are completely flexible and adaptable. Creators can use virtual influencers in whatever promotional capacity they wish, placing them at any place and at any given time. On the contrary, human influencers are constrained by factors such as photographic expertise and modeling abilities. During the COVID-19 pandemic, VI flexibility helped them to remain active in posting innovative content, while RI were confined to their homes.

EXCLUSIVITY Virtual influencers can be produced specifically for one particular brand and remain connected to it forever. On the other hand, human influencers often work with several brands simultaneously and are not solely known or affiliated with them.

BRAND SAFETY Since VI are digitally created, brands can customize VI personas to suit their image and comply with their brand values. This reduces the company's risk of exposure due to inappropriate behavior or tainted past of RI. This also avoids VI from publishing any material that is against the brand or its messages.

BRAND INNOVATION Among younger audiences, companies that partner with virtual influencers are perceived as being more innovative and tech-savvy than those that work with real influencers.

2.4.2 THREATS OF VIRTUAL INFLUENCERS

UNREALISTIC EXPECTATIONS VI are prone to inflating people's perceptions. By redrawing expectations for appearance, style, and culture, adolescents could feel forced to imitate and follow those standards. This could negatively affect the audience's mental and physical health without considering that these digital creations do not physically exist in the real world.

UNREALATABLE The relationship that consumers may develop with VI could be limited. Fans would never meet their favorite VI, and could also perceive a lack of human touch since it is not real, which can harm brand loyalty [31].

AUTHENTICITY Authenticity, trust, and transparency are important values for any influencer. In case of VI, is it possible for a virtual influencer to suggest a product that they have not physically tested? VI will never try on an outfit, a makeup set, or a weight-loss product since they are just digital creations. This raises suspicions about virtual influencers' trustworthiness and authenticity.

Table 2.2: Top 3 Real vs. Virtual Influencers.

	<i>Influencer</i>	<i>Followers</i>	<i>Yearly growth (%)</i>	<i>Most following country</i>	<i>Real followers (%)</i>	<i>Estimated reach</i>	<i>Engagement rate (%)</i>	<i>Estimated per Post Earnings (EEP)</i>
<i>Real</i>	Cristiano Ronaldo	280M	29.1	India (15%)	80.5	15M–80M	1.98	\$889K
	Ariana Grande	232M	26.1	USA (19%)	67.3	15M–60M	1.8	\$996K
	Kylie Jenner	228M	32.5	USA (19%)	61.0	10M–55M	2.5	\$1.2M
<i>Virtual</i>	Lu do Magalu	5M	65	Brazil (79%)	62.2	300k–1500k	0.05	\$13K
	Lil Miquela	3M	49.7	Brazil (14%)	64.2	150k–800k	2.1	\$8K
	Knox Frost	800k	-39.7	USA (33%)	60.6	50k–250k	1.10	\$3K

COSTS Considering the costs besides the partnership itself, content generation is very expensive for VI. Experts of computer graphics are always required behind their actions, which obviously need to be paid. On the contrary, RI can produce a lot of content with minimal effort, and therefore, be more active.

2.4.3 TOP REAL VS. VIRTUAL INFLUENCERS

We now compare real and virtual influencers by analyzing the best representative of both categories. In particular, we focus on top-three virtual/real influencers (i.e., with the highest number of Instagram followers and collaborations with famous brands). We considered Lu do Magalu, Lil Miquela, and Knox Frost for the VI, and Kylie Jenner, Cristiano Ronaldo, and Ariana Grande as the top RI. The comparison is based on reports released by HypeAuditor, summarized in Table 2.2.

The difference in the number of followers of real and virtual influencers is evident. Lu do Magalu, the most popular VI, has 5M followers, while Kylie Jenner, the less popular RI we considered, has ~228M followers. This huge discrepancy probably reflects that many popular RI are celebrities outside Instagram, while VI only exist in social platforms’ scope. Furthermore, VI joined these platforms late compared to RI, and many people still ignore their existence. Even the followers’ growth is substantially different. RI growth was stable around 30 percent in the last year, while for VI, it fluctuates from very high increases (e.g., 65 percent Lu do Magalu) to a substantial decrement (e.g., -39.7 percent Knox Frost). People might have unfollowed the influencer because of its content or the account lost bots. In fact, it is estimated that only around 60 percent of the followers of VI are real people, while the value increases for RI.

The estimated reach expresses the number of people who usually see an influencer post. RI present higher values because of their higher number of followers, but the percentage over the total number of followers is similar to VI one (from 5 to 30 percent). The same applies to the engagement rate. What differs is the diversity of the population they reach. RI are followed in many different countries, while the majority of VI audience usually comes from their origin country (e.g., Lu do Magalu audience is mostly Brazilian). Finally, the EEP is substantially higher for RI, which is expected due to their huge number of followers.

2.4.4 ONTOLOGY AND ETHICS

Defining the ontological status of Virtual Influencers is challenging. A recent study claimed no meaningful difference between RI and VI [40]. Although VI do not physically exist, they have a unique identity well-defined on social media, and their followers interact with them as with any other real influencer. However, VI are still

considered just company tools since their content is designed and created by the humans managing them. This status might change once VI start using AI to generate their content.

Virtual Influencers' ontological status raises challenging ethical questions. Regarding their motivation, if creating "fake" identities for business might be questionable, this is not meaningfully different from real influencers exaggerating and proposing the best version of themselves [40]. Further, even if VI business model is transparent (i.e., more followers means higher prices for their usage), the secrecy behind their management threatens both real influencers and audiences. The former would see VI as unfair competitors; the latter might find VI communication deceiving. Finally, at present, the moral and legal responsibilities of human-controlled VI are difficult to define, and this will be even more challenging for AI-driven VI.

2.5 OPINIONS ABOUT VIRTUAL INFLUENCERS

This section reflects upon some people, creators, and virtual influencers' opinions for or against Virtual Influencers (VI). We analyzed people's opinions by conducting a survey.

2.5.1 PEOPLE'S OPINION ON VIRTUAL INFLUENCERS

In [41], the authors interviewed several people to understand their thoughts about VI, who both supported and opposed them. Many respondents agreed that it was hard to trust a virtual influencer since it pretends to be real while it is not. The novelty and higher engagement of VI was appreciated, but the lack of authentic content and the impossibility to meet virtual influencers resulted to be prominent.

To have a wider understanding of people's thoughts about VI, we conducted a survey targeting social media users. We used an online platform to recruit participants, receiving 360 valid answers from 37 countries. Participants' age ranged from 16 to 61 (avg 26.3, std 7.7), divided into 169 females, 186 males, and 5 others. We validated answers through several attention checks. The survey had four sections: general, marketing, ethical, and evaluation of VI posts.

GENERAL QUESTIONS We started the section by asking: "Have you ever heard about Virtual Influencers? (e.g., Lil Miquela, Lu do Magalu)". 38.6 percent of the participants responded positively, highlighting how the phenomenon is still new and unripe. Moreover, only 16 percent follow at least one VI on socials, and less than 7 percent three or more. A big portion of positive answers came from countries in North America, such as the United States and Mexico, in which VI have been first developed. This section highlighted that people would follow Virtual Influencer mainly for curiosity and fun, rather than to learn something or feel closer to them. Moreover, participants would find it important to see relatable and authentic content from VI (aspects taken for granted for RI). RI are expected more to deliver frequent updates and ways of communicating with their followers. Finally, viewers slightly prefer VI to look like a more real person than a cartoon character and think they should primarily publish content on technology, fashion, daily life, and social matters.

MARKETING QUESTIONS While 70 percent of the participants believe a company must have a RI sponsoring them, only 20 percent think the same for VI. People think VI could give more flexibility to the company and

boost its exclusivity, innovation reputation, and brand safety. Still, only 12 percent of people would trust a Virtual Influencer equally or more than a RI. Rather, 45 percent stated they would trust VI depending on the context, 27 percent always less than a RI, and 15 percent would never trust them.

ETHICAL QUESTIONS In case the behavior of VI appears unethical, only 26 people would condemn the VI itself, while most participants would accuse both the VI creator and the company using it. In general, people like to see Virtual Influencers supporting Social Issues (e.g., Civil Rights, Gender Inequality) but are reluctant if the supported cause is personal or closer to them. This reflects the 60 percent of participants thinking it is impossible to build a relationship with a VI (33 percent answered “Maybe”). Finally, only 15 percent would chat with a VI, 30 percent maybe, and 55 percent not.

VIRTUAL INFLUENCERS’ POSTS EVALUATIONS In this section, we presented three posts created by three virtual influencers (Fig. 2.4), asking for a value from 1 (low) to 5 (high) for the following aspects: authenticity, relatability, innovative content, attractive content, comparable to a real influencer, and overall evaluation. Knox Frost’s post, which depicts him while composing music, on average was considered the most authentic, relatable, and innovative. This might be related to the theme of the picture (i.e., composing music, which might be accepted for a virtual character), and the influencer wearing a mask during the COVID-19 Pandemic. Lil Miquela’s post received the highest votes for attractive content, while Bermuda’s one was voted the most similar to real influencers’ posts. Overall, Knox Frost’s post was better accepted by the participants, but many others stated that VI and their similarity with real influencers scare them, revealing a general negative feeling.



Figure 2.4: Virtual Influencers’ Instagram posts.

2.5.2 CREATOR’S OPINION

Cameron-James Wilson claims he never intended to mislead anyone by creating Shudu. He described her as a “art piece” and a “virtual” celebration of attractive, dark-skinned women [42]. Wilson tried to recreate the elegance embodied by black supermodels as a fashion photographer. He further added that he created Shudu with 3D modeling software and would like to think of her as a mannequin. “You can pose her and give her an expression once you’ve finished creating her”, he said.

Hirokuni Miyaji, the creator of the virtual influencer Liam Nikuro, explained that he wanted to demonstrate what “we can do” for businesses [43]. Their motive is to raise awareness and get brands excited to collaborate with them on their own virtual humans rather than make Liam famous. He also shared his long-term plan for Liam, which is to introduce AI and allow him to interact with real people. Eventually, he stated that the distinction between fictional and real influencers will become increasingly blurred in the future. Only the content will be important.

2.5.3 VIRTUAL INFLUENCER’S OPINION

Besides the creators, even virtual influencers are opening up for interviews to talk about themselves and what they stand for. With the help of interviews conducted with the VI, Lil Miquela and Lu do Magalu, we would like to highlight some of the opinions that these virtual influencers have about themselves. In an interview, Lu was asked “Who are you?”; she replied saying “I’m a strong, virtual woman who creates content to share her knowledge and her causes with everybody”. She further added that she loves assisting people and is fascinated by technology, and innovation and is honored to serve Magalu [44]. A similar question was asked to Miquela, and she answered saying “I’m an artist and have expressed opinions that are unpopular and as a result, have cost me fans”. Furthermore, she would love to do everything that her fans want, but ultimately, she would have to make choices that she believes in [45]. We may conclude that there is almost no difference between interviews conducted with a human being and those conducted with a virtual character. They are questioned in the same way as a real person would be, and these VI have responded close to how a real person would.

2.6 THE FUTURE OF VIRTUAL INFLUENCERS

Our research showed the rising trend of virtual influencers. With the ever-increasing coexistence of human and virtual beings, like in virtual reality applications or the metaverse, we expect the VI phenomenon to continue growing. We believe techniques used on social media by real influencers, in general, can be applied to VI as well to increase their impact. However, several concerns still exist in people’s minds regarding the transparency and authenticity of virtual influencers, facts that were confirmed by our survey.

The majority of virtual influencers are currently CGI-made, limiting audience interaction to static social media posts or videos. Nevertheless, with advancements in AI and virtual reality, some VI are already participating in live interviews and activities, becoming more “human”. AI-driven virtual influencers will raise ethical concerns worth discussing in the future.

Further VI analysis might focus on their content, such as determining whether it is AI-generated or made by humans, and whether consumers will notice and embrace AI-generated content. Finally, VI impact on existing communication technologies, systems, or services should be evaluated. For example, how their behavior changes on different communication platforms, or whether companies will develop new systems to increase their functionalities.

ACKNOWLEDGMENT

We thank HypeAuditor for the Instagram influencers' reports they provided to us for analysis.

3

Follow Us and Become Famous! Insights and Guidelines from Instagram Engagement Mechanisms

People post photos on Instagram (IG) for many purposes, including conveying personal identity, nurturing relationships, feeling part of a community, and promoting worthwhile content [46]. Getting approval from others is highly rewarding, to the point that engagement metrics (e.g., Likes, Comments, Views) have become addictive, especially for low self-esteem people [47]. Some people use IG for only a few minutes daily, but for others, e.g., the influencers, it has become a way of life. In short, influencers are people who can influence society. Due to their ability to reach people, companies have used them to market their products [48], so much so that influencer marketing is estimated to generate \$33.25 billion in 2022 [15, 49]. Whatever the reason, everyone strives to get as much engagement as possible under their posts, even at the cost of buying it [50]. For influencers, planning popular posts is time-consuming and costly, with no guarantee of success. In this regard, a tool that can predict the popularity of a post in advance would be of great interest, especially when sponsored posts are highly remunerated (e.g., Cristiano Ronaldo is paid around \$1 Million for a single post, as shown in Chapter 2).

Researchers have proposed algorithms for predicting the popularity of posts, but they are far from perfect (Section 3.1.1). The first limitation is they measure engagement only in terms of likes, not incorporating stronger forms of interaction or what IG favors, i.e., the number of comments [51]. Then, the lack of a universal dataset for such predictive tasks leads to outcomes based on limited or biased data. Furthermore, these models often make use of Deep Learning (DL) models that may be difficult (or even impossible) to interpret [52, 53]. As a result, end-users must *first* create the post through an expensive and time-consuming process, and *then* assess posts' popularity using such black-box models.

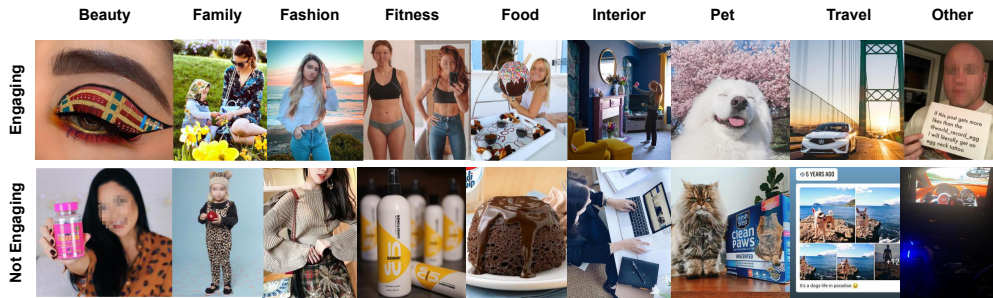


Figure 3.1: Engaging and not engaging posts for each category.

CONTRIBUTION. Our goal is to *understand* and *explain* the underlying mechanisms driving IG engagement. We extract domain-relevant features leveraging the well-known capabilities of DL models, but entrust the prediction to interpretable Machine Learning (ML) algorithms [54], allowing us to draw guidelines. According to the IG recommendation algorithm, we consider likes and comments as engagement metrics. We conduct extensive experiments on a recent dataset of 10M posts from 34K influencers. We demonstrate through statistical analysis that influencer tiers (i.e., their audience wideness) and categories (i.e., the primary topic they cover) are crucial to predict posts’ popularity. Figure 3.1 shows engaging and non-engaging posts, which supports the intuition that the characteristics determining engagement differ by category.

Last, we propose a novel unsupervised approach to detect hot topics (i.e., highly engaging) for each category, which overcomes the need for domain knowledge to extract meaningful features. We summarize our contributions as follows:

- We analyze the underlying mechanisms of IG engagement, in terms of likes and comments, from a dataset of 10M posts, divided into nine categories and five tiers of influencers, leveraging statistical analysis and interpretable ML algorithms;
- We propose an interpretable model to predict posts’ engagement and define handy guidelines, exploiting several features extracted by State-of-the-Art (SotA) Deep Learning models;
- We propose a novel unsupervised approach for spotting highly engaging topics in each tier and category, considering both visual and textual content;
- We release our enriched dataset upon request as a possible baseline for future works.

ORGANIZATION. Section 3.1 presents related works. Section 3.2 describes the dataset and preliminary assessments, while the engagement prediction and interpretation are conducted in Section 3.3. The hot topic detection appears in Section 3.4, and the final guidelines are provided in Section 3.5. Section 3.6 concludes the chapter.

TRANSPARENCY. To promote transparency and reproducibility, we created a repository¹ containing exhaustive details on our study, the source code, and our dataset, which can be requested for research purposes only.

¹<https://github.com/spritz-group/FollowUs>

3.1 RELATED WORKS

The popularity of IG posts has been mainly assessed by predicting the number of likes they received, usually divided by the number of followers of the posting user, or after a log-scaled transformation. Mazloom et al. [55] predicted the popularity of brand-related posts by defining engagement parameters important in marketing and using a Support Vector Regression. The authors extended further their work [56] for different categories such as activities, landscapes, people, and animals. De et al. [57] trained a Deep Neural Network (DNN) on posts' metadata (e.g., creation date, users tagged, hashtags) to predict the popularity of future posts of an Indian lifestyle magazine. Similarly, Zhourian et al. [58] approached popularity prediction as a regression and classification task, focusing on posts of Iranian business IG accounts. Rather than predicting popularity in general, Zhang et al. [59] implemented a dual-attention mechanism to predict user-specific posts' popularity. Ding et al. [60] tried to isolate the contribution of the visual content by predicting the intrinsic image popularity through a DNN. Gayberi et al. [61] extracted concepts and object features using a pre-trained model on Microsoft COCO Dataset [62] and used several Machine Learning algorithms to predict the likes of a post. Through transfer learning, Riis et al. [63] extracted visual semantics such as concepts, scenes, and objects and tried to set an explainable baseline for population-based popularity prediction. Carta et al. [64] proposed an approach based on Gradient Boosting and feature engineering of users' and posts' metadata to predict popularity in a classification fashion. Last, Purba et al. [65] attempted to create a global dataset of around 20K posts from 16K users and leveraged features extracted from hashtags, image analysis, and user history, predicting the number of likes over followers using a Support Vector Regression (SVR).

3.1.1 LIMITATIONS OF EXISTING LITERATURE

This section briefly describes why the past literature in the area is incomplete and how our work closes such gaps.

INCOMPLETE POPULARITY METRIC. Prior works focused *exclusively* on the number of likes to measure post popularity, which is outdated and discrepant with the current IG algorithm. The IG algorithm was changed in 2021 [51] to show users content based on their interests, not just their social graph. The shift to such *recommendation media* changed how posts became popular. The content *must* be engaged with, mainly through likes and comments, so that Instagram spreads it on many users' feeds, and only then it can become popular. Consequently, it is crucial to consider the number of comments as an indicator of engagement, given they result from a higher user effort than leaving a like [66], and thus are more relevant for the IG recommendation algorithm [51]. As far as we know, we are the first to include comments in our engagement metrics.

LIMITED OR BIASED DATASET. Since Meta's APIs² are limited, there are no public datasets to use as baselines. Most prior works collected their datasets, focusing on limited portions of the population [57, 58]. Moreover, except for Mazloom et al. [56], they do not consider the different categories and tiers of the creators. For example, a picture of a dog and a top model would become popular for different reasons. The influencer tier, instead, was not previously considered. However, the engagement rate of influencers with millions rather than a few thousand

²<https://developers.facebook.com/docs/instagram-api/>

followers reaches different levels [67], and normalizing the metrics is insufficient. In Section 3.2.3, we demonstrate that influencer categories and tiers strongly influence engagement metrics (p -value <0.001) and thus need to be treated separately to yield accurate predictions.

POOR RESULTS EXPLANATION. As deep learning algorithms and ensemble machine learning algorithms have improved performance, recent works have largely relied on end-to-end black-box models [59, 60, 63] rather than extracting specific features to train simple regressors or classifiers [55, 58]. While the model is more accurate, it is difficult (or impossible) to understand what has been learned [52, 53, 68]. As extensively demonstrated in the landmark Nature paper by Rudin [52], interpretable models *must* be preferred to (complicated) black-box models when explainability is critical. Often, if the problem has structured data and meaningful features, there is no significant difference in performance between more complex classifiers (i.e., DNNs, ensemble methods) and simpler ones. We remind the reader that interpreting a model substantially differs from explaining it [69], as done by Riis et al. [63].³ Furthermore, in our scenario, using a black-box model for post popularity means the user must create the post first, which can be extremely costly, as shown in Chapter 2. Thus, we use an interpretable model (i.e., a Decision Tree) to provide guidelines that can be followed *before* generating a post that wishes to gain popularity.

3.2 DATASET & PRELIMINARY ASSESSMENTS

In this section, we describe the dataset (Section 3.2.1), the engagement metrics (Section 3.2.2), the importance of dissecting the data in categories and inner tiers (Section 3.2.3), and the features we considered and extracted for the study (Section 3.2.4).

3.2.1 DATASET DESCRIPTION

In our work, we utilize the dataset proposed by Kim et al. [70] that contains 10,180,500 posts from 33,935 influencers collected in 92 days. The influencers are divided into nine categories, namely Beauty, Family, Fashion, Fitness, Food, Interior, Pet, Travel, and Other, depending on their content type. Furthermore, we categorize each influencer in the five well-known tiers based on their number of followers:⁴ Nano [1K, 10K), Micro [10K, 50K), Mid [50K, 500K), Macro [500K, 1M), Mega [1M, +∞].

Each post is composed of the image, caption, metadata (e.g., publish time, location), and engagement metrics (i.e., the number of likes and comments)⁵. Similar to previous works [58, 63, 65], we normalize our target features (likes and comments) dividing them by the number of followers of the post’s creator, allowing a fair comparison between posts of different users⁶. Given that creators’ followers were taken only at the end of the collection, we

³Interpretability means that the cause and effect can be determined, while explainability indicates which parameters are linked to a prediction, explaining the phenomenon a posteriori, non-deterministically.

⁴<https://www.shopify.com/id/blog/instagram-influencer-marketing>

⁵We did not further process these metrics, e.g., by removing spam comments, since IG algorithm accounts for quantity, and not quality [51].

⁶As a convenience, we refer to the normalized numbers simply as likes and comments.

remove posts older than thirty days, a period within the followers’ growth remains mostly stable [71]. Moreover, since an IG post engagement growth usually last one to three days [67], we exclude posts younger than five days. In the end, our dataset counts 650,118 posts created by 33,935 influencers. Table 3.1 shows the number of posts (and influencers) per tier and category. The small presence of some categories (e.g., food, interior, pet) for very popular influencers is aligned with the actual IG categories distribution [72].

Table 3.1: N. posts (influencers) for categories and tiers.

	Nano	Micro	Mid	Macro	Mega
Beauty	8449 (546)	7879 (537)	6998 (387)	745 (35)	835 (37)
Family	29744 (1887)	23432 (1330)	12740 (674)	1267 (77)	2145 (102)
Fashion	49622 (3154)	82895 (4841)	68737 (3238)	8833 (325)	8987 (355)
Fitness	5060 (301)	6194 (424)	6256 (342)	352 (27)	748 (39)
Food	27697 (1511)	28191 (1440)	14805 (583)	936 (25)	305 (6)
Interior	6461 (373)	9606 (541)	5525 (261)	413 (13)	404 (7)
Pet	3416 (164)	4073 (260)	2929 (153)	87 (6)	115 (4)
Travel	24445 (1774)	19630 (1522)	13098 (838)	816 (49)	540 (27)
Other	73213 (2976)	38967 (1454)	31874 (1004)	4255 (120)	6399 (166)

3.2.2 ENGAGEMENT METRICS: LIKES & COMMENTS

Prior works (Section 3.1) focused *exclusively* on the number of likes as a popularity metric. Nonetheless, since 2021, comments have become a crucial engagement metric to make a post popular [51]. Figure 3.2 shows the box plots of the likes and comments for every category and tier. There are some common trends, but commenting is less frequent than liking. Such discrepancy is justified by the two different levels of public expression they carry [66]. Comments are costly and expose users’ opinions more, while likes are almost immediate and instinctive. Hence, a highly-liked post may not receive many comments. Further demonstrating the independence of the two metrics, we calculated Spearman correlation coefficients (ρ) between the distributions of likes and comments. The result ($\rho = 0.58$, p -value < 0.001) shows a moderate correlation between likes and comments, demonstrating that they need to be analyzed separately as two not-so-dependent phenomena. Thus, we consider as engagement metrics $\frac{\#Likes}{\#Followers}$ and $\frac{\#Comments}{\#Followers}$.

3.2.3 THE IMPORTANCE OF TIERS AND CATEGORIES

Do the Northern Lights create more engagement than a cute puppy? How about a pineapple pizza in Naples? As these concepts are incomparable, answering these questions a priori is difficult. Similarly, would people react analogously if a celebrity and a normal person divorced? Most likely not. Those are just a few examples behind our hypothesis: *influencers’ tiers and categories significantly affect the engagement metrics*. To demonstrate this hypothesis, we conduct a Multivariate ANOVA (MANOVA) [73], with category and tier as independent variables and the likes and comments as dependent ones. By such a statistical test, we can determine whether the mean scores of engagements differ between our nine categories and five tiers. Before conducting MANOVA, we normalize the likes and comment distributions as explained in Section 3.2.2. Among the MANOVA results, we adopted Pillai’s

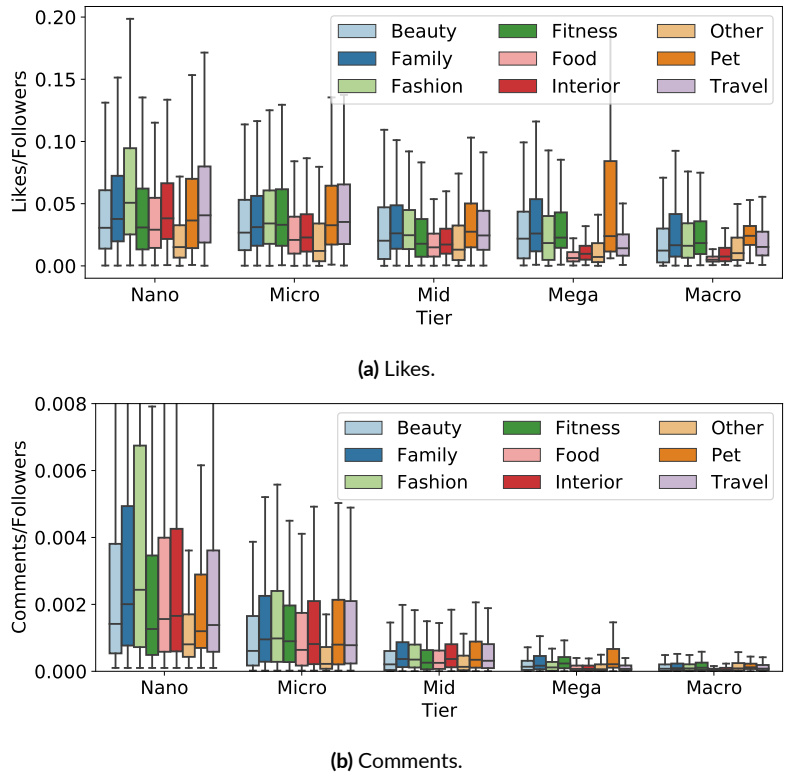


Figure 3.2: Box plots of Likes and Comments for the different categories and tiers. Note that the y-axes have two different scales, giving a lower number of comments in general.

trace test, which is robust when MANOVA assumptions are violated [74]. Pillai’s trace test returned 0.0942 and 0.2646 for category and tier, respectively, with p -value < 0.0001 . Since the p -value is less than the significance level $\alpha = .0001$, we reject the null hypothesis of the MANOVA and conclude that the explanatory variables (tier and category) significantly affect the values of the response variables (likes and comments). In particular, the tier resulted contributing more than the category.

3.2.4 FEATURES EXTRACTION

Starting from the filtered posts of Section 3.2.1, we augmented our dataset with features from each kind of data source, such as metadata, images, and text, which we now briefly describe.⁷ In the process, we also employed nine SotA DL algorithms.

⁷The complete list of features is available in our repository.

METADATA FEATURES

The posts’ metadata provides information on their “discoverability”. This term refers to features that increase post visibility, like hashtags and mentions. Hashtags label the post’s content, while mentions allow tagging someone in a post, so their followers can reach the source profile. Therefore, we created two counters to keep track of the number of hashtags and tagged users. In addition, we specify whether the post is a video, sponsored, has a location, and time-related information, for a total of 10 features.

IMAGES FEATURES

We extract features from images on multiple levels to fully describe the image content, including the scene, people, and aesthetic features.

SCENE FEATURES. To describe the environment where the picture is set, we leverage the Places365 DL model [75]. The model can identify up to 365 places mapped to 3 macro categories (indoor, outdoor natural, outdoor man-made) and 16 micro categories (e.g., shopping/dining, transportation). Moreover, we perform object detection of 80 different classes mapped in 12 categories using Faster R-CNN MobileNetV3 [76] trained on MS COCO dataset [62], counting the objects belonging to each category.

PEOPLE FEATURES. Using RetinaFace [77], we perform face boundaries detection and then estimate the age and gender [78] of the detected people. For each post, we save the number of females and males, and min, max, mean, and standard deviation of people’s age. Furthermore, guided by the well-known impact of nudity in advertising [79], we perform nudity detection using NudeNet [80] for Beauty and Fashion categories, in which the main subject is the human body. The model determines whether 16 parts of the body (e.g., breast, belly, feet, buttocks) are exposed.

AESTHETIC FEATURES. Taking inspiration from Guntuku et al. [81], we derive aesthetic features of the image. In particular, we first extract the percentage of red, green, and blue channels. Then, from the HSV (Hue, Saturation, Value) representation, we obtain the percentage of luminance, warm and cold colors, pleasure, arousal, and dominance scores [82, 83]. Furthermore, we leverage Kong et al. [84] model to obtain eleven high-level aesthetic features (e.g., color harmony, motion blur, content symmetry). Last, we extracted the sentiment score conveyed by the image through the model proposed by Campos et al. [85].

OTHER FEATURES. For the Pet category, we calculated pets’ cuteness scores through a Cute Animal Detector [86]. In total, we obtained 80 visual features.

TEXT FEATURES

From the posts’ captions, we extracted features such as the caption length, the number of Emojis, and their relative sentiment [87]. Moreover, we retrieve the sentiment of the whole text leveraging Google Cognitive Services (GCP) [88], expressed as a score ($\text{Sentiment}_{\text{score}} \in [-1, 1]$, where -1 is negative, 0 is neutral, and $+1$ is positive)

and magnitude ($\text{Sentiment}_{\text{magnitude}} \in [0, +\infty)$), that is representing the strength of the sentiment. We translated non-English text using GCP, and obtained five textual features in total.

3.3 PREDICT & INTERPRET THE ENGAGEMENT

Through correlation analysis, we uncover features that correlate with engagement. Then, we use interpretable models to predict engagement and develop guidelines for producing engaging content.

3.3.1 CORRELATION ANALYSIS

To determine which features contribute the most to raising engagement, we correlate the features with our two engagement metrics (Likes and Comments). To this aim, we use Spearman’s rank correlation coefficient r_s [89]. This method offers the advantages of producing feature ranks, being insensitive to outliers, and not requiring any specific normalization of the data. The Spearman’s correlation coefficient is based on Pearson’s correlation coefficient [90] and it is defined as follows. For n observations, the n scores X_i, Y_i, X_i, Y_i are converted to ranks as $R(X_i)$ and $R(Y_i)$, and r_s is computed as:

$$r_s = \rho_{R(X), R(Y)} = \frac{\text{cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \quad (3.1)$$

where ρ denotes Pearson correlation coefficient but applied to the rank variables, $\text{cov}(R(X), R(Y))$ is the covariance of the rank variables, $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the standard deviations of the rank variables. As for Pearson’s correlation coefficient, the Spearman correlation values are expressed in the range $r_s \in [-1, 1]$ along with their p -value that express their significance that is higher as much as the value is small.

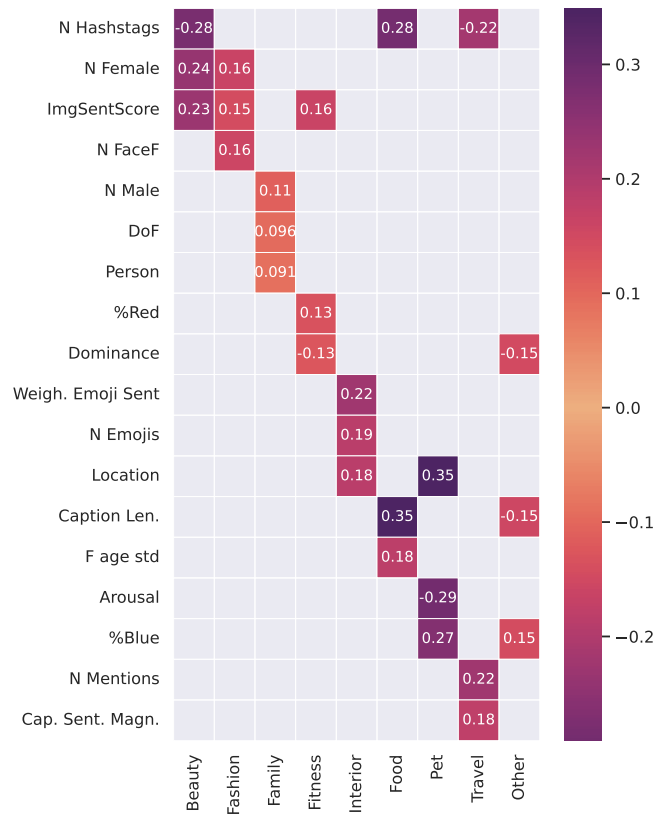
For each one of the influencers categories (i.e., beauty, fashion, etc.) and for each one of the tiers (i.e., nano, micro, etc.) we perform the correlation analysis of the features against the engagement metrics (Likes, Comments)⁸. Figure 3.3 reports the top-3 most correlated features to comment engagement for each category in the Nano and Mega tiers. We notice immediately how the most relevant features in different categories are very similar when the tiers are small (Nano in the figure, but also Micro). In contrast, behavior becomes category-specific as tier size increases (Mega in the figure, but also Macro). This behavior also occurs for likes. As we can see, small influencers, or users aspiring to become influencers, use similar strategies in every category. These include the use of many mentions, a long caption, and location tags.

LIKES ENGAGEMENT. By examining how features correlate to likes, it is possible to observe how the engagement mechanism differs for each type of influencer. Generally, we can notice that the strongest features are related to the images and their content rather than to the text (i.e., the caption), while almost the opposite occurs for the comments. The number of mentions generally has a positive impact on likes, even if their relevance decreases as tiers increase. A similar pattern can be observed in the number of hashtags having a negative effect, which tends to intensify in larger tiers. Availability of the location is very relevant up to the micro tier, after which it becomes category-specific.

⁸We made all the results available in our repository.



(a) Nano Influencers.



(b) Mega Influencers.

Figure 3.3: Top-3 features per absolute correlation value (ρ -value < 0.001) in comments engagement for each category.

COMMENTS ENGAGEMENT. Similarly to the likes engagement, we observe an overall positive correlation for the number of mentions, even though the relevance goes decreasing as the tier increases. Even in this case, the number of hashtags plays an antagonist role in comments engagement, instead the presence of the location field in a post is generally helpful. This type of engagement benefits from text-specific features such as caption length, sentiment magnitude, and Emoji usage.

Takeaway: Likes engagement differs from Comments engagement in that they are oriented toward images and captions, respectively. Additionally, low-tier influencers tend to adopt the same strategy to grow, while high-tier influencers exhibit more category-specific characteristics.

3.3.2 ENGAGEMENT PREDICTION & GUIDELINES METHODOLOGY

Table 3.2: Performance of Decision Trees (DT) against a dummy classifier (Dum.). In bold, the best scores for likes and comments for each category. Values reported are F1-Score, macro-weighted $mean \pm std.$

	Nano				Micro				Mid			
	Like		Comments		Like		Comments		Like		Comments	
	DT	Dum.	DT	Dum.	DT	Dum.	DT	Dum.	DT	Dum.	DT	Dum.
Beauty	0.61 \pm 0.002	0.46 \pm 0.004	0.65 \pm 0.003	0.46 \pm 0.004	0.60 \pm 0.003	0.47 \pm 0.012	0.62 \pm 0.006	0.46 \pm 0.017	0.61 \pm 0.004	0.47 \pm 0.011	0.61 \pm 0.002	0.47 \pm 0.016
Fashion	0.57 \pm 0.006	0.47 \pm 0.008	0.62 \pm 0.003	0.47 \pm 0.001	0.53 \pm 0.002	0.46 \pm 0.002	0.57 \pm 0.004	0.46 \pm 0.004	0.53 \pm 0.004	0.47 \pm 0.004	0.56 \pm 0.007	0.47 \pm 0.002
Family	0.57 \pm 0.005	0.47 \pm 0.010	0.59 \pm 0.005	0.47 \pm 0.007	0.53 \pm 0.002	0.46 \pm 0.005	0.56 \pm 0.002	0.47 \pm 0.008	0.55 \pm 0.007	0.48 \pm 0.005	0.55 \pm 0.002	0.47 \pm 0.003
Fitness	0.62 \pm 0.004	0.48 \pm 0.005	0.63 \pm 0.004	0.47 \pm 0.009	0.60 \pm 0.006	0.47 \pm 0.009	0.59 \pm 0.012	0.47 \pm 0.006	0.61 \pm 0.010	0.47 \pm 0.009	0.61 \pm 0.003	0.47 \pm 0.010
Food	0.59 \pm 0.003	0.46 \pm 0.003	0.62 \pm 0.001	0.47 \pm 0.010	0.57 \pm 0.001	0.47 \pm 0.007	0.61 \pm 0.001	0.47 \pm 0.006	0.57 \pm 0.005	0.46 \pm 0.010	0.63 \pm 0.002	0.46 \pm 0.011
Interior	0.59 \pm 0.002	0.48 \pm 0.012	0.62 \pm 0.003	0.46 \pm 0.01	0.59 \pm 0.003	0.47 \pm 0.002	0.63 \pm 0.001	0.47 \pm 0.006	0.58 \pm 0.010	0.45 \pm 0.012	0.63 \pm 0.003	0.47 \pm 0.011
Other	0.58 \pm 0.002	0.47 \pm 0.001	0.57 \pm 0.0001	0.47 \pm 0.004	0.55 \pm 0.001	0.47 \pm 0.0021	0.56 \pm 0.005	0.47 \pm 0.002	0.59 \pm 0.001	0.47 \pm 0.006	0.58 \pm 0.001	0.46 \pm 0.005
Pet	0.69 \pm 0.004	0.47 \pm 0.019	0.72 \pm 0.006	0.49 \pm 0.009	0.60 \pm 0.008	0.45 \pm 0.018	0.62 \pm 0.005	0.46 \pm 0.007	0.61 \pm 0.018	0.46 \pm 0.006	0.64 \pm 0.003	0.47 \pm 0.024
Travel	0.60 \pm 0.001	0.47 \pm 0.0042	0.61 \pm 0.001	0.47 \pm 0.006	0.59 \pm 0.0012	0.47 \pm 0.003	0.63 \pm 0.002	0.47 \pm 0.007	0.58 \pm 0.004	0.47 \pm 0.006	0.64 \pm 0.009	0.48 \pm 0.010

	Macro				Mega			
	Like		Comments		Like		Comments	
	DT	Dum.	DT	Dum.	DT	Dum.	DT	Dum.
Beauty	0.69 \pm 0.012	0.45 \pm 0.054	0.67 \pm 0.011	0.47 \pm 0.046	0.74 \pm 0.006	0.48 \pm 0.029	0.68 \pm 0.009	0.49 \pm 0.025
Fashion	0.61 \pm 0.001	0.47 \pm 0.009	0.60 \pm 0.011	0.48 \pm 0.015	0.57 \pm 0.013	0.46 \pm 0.011	0.56 \pm 0.002	0.47 \pm 0.005
Family	0.60 \pm 0.014	0.47 \pm 0.031	0.59 \pm 0.013	0.47 \pm 0.017	0.59 \pm 0.006	0.47 \pm 0.026	0.57 \pm 0.001	0.46 \pm 0.007
Fitness	0.72 \pm 0.010	0.45 \pm 0.045	0.69 \pm 0.032	0.46 \pm 0.023	0.65 \pm 0.013	0.49 \pm 0.016	0.61 \pm 0.033	0.48 \pm 0.010
Food	0.75 \pm 0.008	0.46 \pm 0.020	0.71 \pm 0.024	0.47 \pm 0.061	0.71 \pm 0.023	0.57 \pm 0.091	0.72 \pm 0.029	0.53 \pm 0.098
Interior	0.76 \pm 0.033	0.48 \pm 0.021	0.83 \pm 0.020	0.49 \pm 0.015	0.77 \pm 0.011	0.44 \pm 0.083	0.74 \pm 0.027	0.45 \pm 0.074
Other	0.63 \pm 0.003	0.46 \pm 0.013	0.59 \pm 0.008	0.47 \pm 0.023	0.60 \pm 0.003	0.46 \pm 0.005	0.58 \pm 0.008	0.46 \pm 0.012
Pet	0.94 \pm 0.057	0.42 \pm 0.042	0.87 \pm 0.048	0.42 \pm 0.042	0.78 \pm 0.045	0.51 \pm 0.032	0.77 \pm 0.012	0.5 \pm 0.065
Travel	0.72 \pm 0.008	0.45 \pm 0.024	0.67 \pm 0.050	0.47 \pm 0.012	0.72 \pm 0.010	0.46 \pm 0.031	0.71 \pm 0.017	0.42 \pm 0.032

Besides explaining which characteristics of Instagram posts build engagement, we also aim to form guidelines for producing the ideal engaging post. Having such guidelines for influencers saves time and money consistently since the process for producing a high-engagement post is well-defined. To this aim, we leverage interpretable models, even if this could reduce the overall accuracy. Deep learning models are well known for their capability of solving complex tasks, but by definition they work as a black box that we cannot reliably explain [52]. For this reason, we decided to utilize Decision Trees (DT) [91]. By training a DT classifier to predict low or high engagement (bottom 0.75 and top 0.25 quantile), we can simply explain how to produce top engagement posts by following the binary classification tree. The paths to reach the top 0.25 quantile leaves represents guidelines for creating high-engagement posts.

IMPLEMENTATION

Since influencers behave differently according to their category and tier, as they want to reach a different public, we create an engagement classifier for each category and tier. Each classifier is trained and validated on 75% of the dataset and tested on the remaining 25%. To build accurate estimators for each dataset (i.e., combinations of the nine categories and five tiers – 45 in total), we fine-tune the Decision Tree Classifier through a Grid Search ($cv=5$) that evaluates more than 20K combinations of parameter fits to achieve the best F1-Score (macro weighted) possible. To further reduce the bias due to the random split of the dataset, we repeated the evaluation three times on three different training-test partitions. Considering low and high engagement based on 0.75 and 0.25 quantiles implies having heavily unbalanced classes that make the learning process harder. Therefore, we also introduce as tuning parameters the use of well-known under-sampling and over-sampling techniques, i.e., SMOTE and Tomek links [92, 93].

RESULTS

The results on the test sets are reported in Table 3.2. All the results surpass the dummy classifier, showing our method can effectively predict posts' engagement. Moreover, the standard deviations are fairly low, suggesting the models are stable. In terms of Likes, predictions are generally more accurate for Macro and Mega influencers, raging around 60-80% F1-score (20-40% better than the dummy). The reason can be that these high-tier influencers tend to be more diversified as we found in the correlation analysis, making some characteristics more effective. Accordingly, our classifier exhibited difficulties in the lower tiers of Fashion, in which influencers tend to post similar content, and Other, in which the content was extremely diverse. On average, we reached the best performances for Pet, Interior, and Beauty. Regarding Comments, we find a behavior similar to Likes, except for the best performances for Fashion and Family, which appear for Nano influencers. A possible reason is that many Nano influencers might not know the best practices for creating engaging captions, which are strongly correlated to comments engagement as shown in correlation analysis. The best categories we predicted are Pet, Food, and Travel. Last, we reached the best Likes and Comments prediction score (94% and 87%, respectively) for the Pet Macro posts. An example⁹ of guideline with a DT structures is depicted in Figure 3.4. Following the nodes conditions (i.e., post characteristics), a label will be assigned when reaching a leaf (i.e., bottom 0.75 or top 0.25 quantile). We will present more examples of guidelines in Section 3.5.

BASELINES COMPARISONS

As Mazloom et al. [56], Gayberi and Oguducu [61], and other similar studies mentioned, comparison with other works in this area is not completely possible. The main reasons are the use of private algorithms and data, and how the problem is formulated. Unfortunately, IG policies¹⁰ never allowed automatic collection and release of common users' posts, forcing previous works to create a new (private) dataset everytime [55, 56, 58, 61, 57, 64, 65]. Moreover, given the lack of a common dataset to work on, some works focused on a regression problems [58, 61, 65], other on a classification problem [58, 57, 64], adopting different metrics, such as the log-normalized

⁹All the results are available in our repository.

¹⁰<https://help.instagram.com/581066165581870>, accessed: Sep 2022.

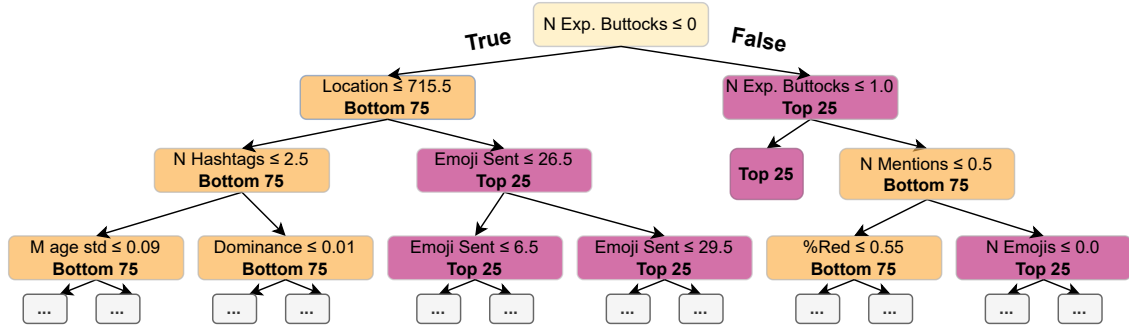


Figure 3.4: Example of guidelines generated by the decision tree for category Beauty, tier nano, likes engagement. The representation is limited at a maximum depth of 3.

number of likes [61, 60] or the likes divided by the number of followers [58, 63, 65]. Thus, to set up baselines despite the aforementioned limitations,¹¹ we adopted four models: (i) I²PA of Ding et al. [60] (the only one publicly released); (ii) a Decision Neural Network (Dec-NN) to represent prior works which first extract generic visual or textual features, and then trained a non-interpretable classifier (similar to [56, 61, 63]); (iii) End-to-End Deep NN (EE-DNN) for prior works that relied on end-to-end black-box DL models, giving in input both posts’ images and captions simultaneously (similar to [59]); (iv) a stratified dummy classifier,¹² which predicts targets based on the training set distribution. Both Dec-NN and EE-DNN extract posts’ image and caption embeddings (through ResNet50 [94] and Sentence-Bert [95]); however, EE-DNN fine-tunes them before the fusion, while Dec-NN receives their early-fusion as input. The decision is taken through three ReLU feed-forward layers (sizes = 2048+768 → 256 → 128 → 2). Both NN were Adam optimized and trained for 50 epochs with early stopping (patience = 5).

The results of Table 3.3 show that our approach outperformed the baselines for each category, except for Fashion and Other, in which we achieved comparable performance, demonstrating the superiority of our simple DT over Deep Learning models. For the categories Beauty, Fitness, Food, Interior, Pet, and Travel, our results are statistically significantly higher than the second-best model (calculated through unpaired *t*-test, two-tailed *p*-value < 0.05). Particularly noteworthy is the result against I²PA and EE-DNN, which represents SotA end-to-end DL models. In particular, EE-DNN performs pretty poorly, likely because fine-tuning the feature extraction modules led to overfitting. On the other hand, Dec-NN, which is more similar to our strategy, generalized better by not tuning the image and text general representations. Probably, we surpassed such baselines mainly because of the category-related features we extracted, again stressing that developing a cross-category engagement predictor could be unfeasible. Accordingly, we probably could not beat Dec-NN in the Other category because of the lack of category-related features.

Although the comparison with previous work is not completely fair for the above reasons, our results are comparable [59, 57] or better [63, 64, 65] than the ones reported on their own data. Anyhow, we remind the reader that our goal is to *explain* the engagement, not necessarily surpass the prediction of existing non-interpretable

¹¹Note that some features used in previous works were not available in our dataset, limiting the comparison.

¹²<https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html>

models. Last, our dataset was collected using IG APIs from business accounts and is thus shareable. We believe our dataset could serve as a baseline for future works.

Table 3.3: Comparison of Mean F1-Score between our model (DT) and baselines in predicting Likes. Underlined results are statistically significantly higher (two-tailed p -value < 0.05) than the second-best.

Category	DT (Our)	I ² PA	Dec-NN	EE-DNN	Dummy
Beauty	<u>0.65\pm0.055</u>	0.587 \pm 0.026	0.582 \pm 0.097	0.362 \pm 0.037	0.466 \pm 0.024
Fashion	0.563 \pm 0.030	0.581 \pm 0.019	0.572 \pm 0.043	0.327 \pm 0.013	0.464 \pm 0.008
Family	0.568 \pm 0.026	0.567 \pm 0.019	0.507 \pm 0.071	0.347 \pm 0.042	0.476 \pm 0.017
Fitness	<u>0.640\pm0.043</u>	0.545 \pm 0.026	0.511 \pm 0.069	0.377 \pm 0.062	0.478 \pm 0.018
Food	<u>0.638\pm0.077</u>	0.550 \pm 0.030	0.518 \pm 0.053	0.464 \pm 0.269	0.48 \pm 0.031
Interior	<u>0.660\pm0.087</u>	0.534 \pm 0.056	0.461 \pm 0.047	0.309 \pm 0.031	0.468 \pm 0.022
Other	0.590 \pm 0.026	0.540 \pm 0.022	0.602 \pm 0.021	0.318 \pm 0.013	0.463 \pm 0.004
Pet	<u>0.724\pm0.126</u>	0.564 \pm 0.046	0.630 \pm 0.127	0.342 \pm 0.0727	0.461 \pm 0.015
Travel	<u>0.642\pm0.064</u>	0.570 \pm 0.012	0.473 \pm 0.044	0.342 \pm 0.069	0.457 \pm 0.016

FEATURE IMPORTANCE

Guidelines to create engaging posts result from following the tree generated by the DT classifier. In addition, similarly to correlation analysis, the content creator can inspect the model’s feature importance to determine which features are impacting the engagement predictions. Thus, we studied the features used by the models, checking whether they matched with correlation results. A representative example¹³ of this analysis is shown in Table 3.4, which suggests good correspondence with the factors expressed in Section 3.3.1. For example, the presence of common features in small tiers, followed by category-specific features with increasing tier size. As for the correlation analysis, the number of mentions and whether a location is given resulted in importance that is inversely proportional to the tier size.

Takeaway: *A simple and interpretable Decision Tree can outperform Deep Learning algorithms if leveraging domain-knowledge features. Prediction results and feature importance analysis confirm the consideration drawn by feature correlation, showing how similar and dissimilar tiers and category behaves.*

¹³All the results are available in our repository.

Table 3.4: Features importance of category Fashion, tier micro.

#	Feature	Imp.	#	Feature	Imp.
1	N Mentions	1.0	1	N Exp. Buttocks	1.0
2	Age avg	0.80	2	N Mentions	0.64
3	Dominance	0.73	3	Caption Len.	0.18
4	N Exp. Buttocks	0.46	4	N Emojis	0.16
5	Outdoor Natural Env.	0.19	5	Cap. Sent. Magn.	0.13

(a) Likes. (b) Comments.

3.4 SPOTTING INSTAGRAM HOT TOPICS

Our features allow us to predict a post’s engagement with good accuracy, but there is room for improvement. In our *interpretable* approach, features have to be extracted a priori instead of being learned “automatically” by a deep learning model. Thus, our features are limited by our educated guesses of what could be engaging, and by the concepts obtainable through existing SotA deep learning models. For instance, if available, we would have used a love or a marriage scene detector, which is likely to produce high engagement. Although such detectors could be implemented through classical approaches (e.g., by fine-tuning an image recognition NN like ResNet [94]), we opted for defining an unsupervised strategy to detect *general hot topics*. In particular, we aim to find (if any) topics or concepts that, if present in a post, would create high engagement independently from the publisher. In this context, unsupervised means we make no assumptions on which topics are engaging (as we did to extract category-related features for Section 3.3), but rather explore users’ interests [96].

From Section 3.3 we learned that likes and comments are mainly driven by the image and caption, respectively. Thus, in the next experiments, we focus on finding likes-related hot topics through visual features, and comments-related hot topics through textual features. We now present our methodology and findings.

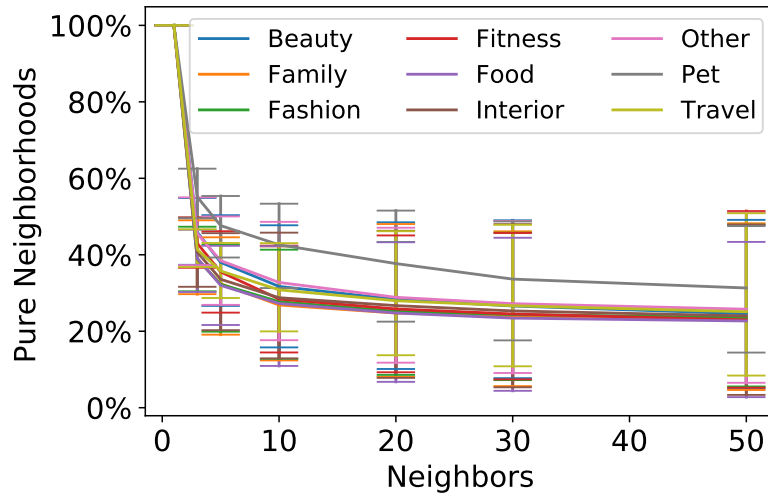
3.4.1 METHODOLOGY

The idea behind our method is to group together semantically similar images and captions and observe whether some of these groups reach high engagement on average.

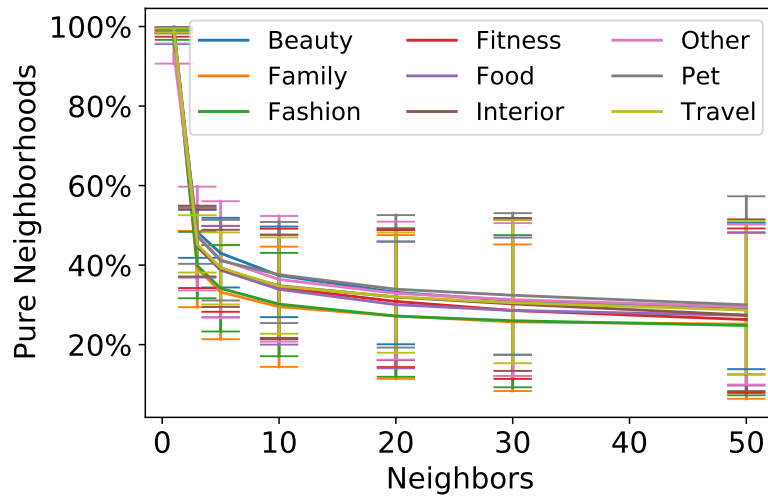
EMBEDDINGS. To define image and text semantic similarity, we rely on the concept of embedding. An embedding is a vector representation of an object (e.g., image, text) in which objects with similar semantics have similar vector profiles [97]. Embeddings are usually extracted by taking the output of the penultimate layer of a deep neural network performing a classification task. In our experiments, we retrieved image embeddings using ResNet50 [94] pre-trained on the ImageNet dataset, and text-embeddings using Sentence-Bert [95] (in particular, in its version *all-mpnet-base-v2* [98]). Before extracting the text embeddings, we translated non-English text leveraging Google Cloud Platform [88], so to perform language detection and translation automatically.

SEMANTICALLY SIMILAR NEIGHBORHOOD. As a first approach, we could create clusters of similar images or captions, and see whether some clusters present higher engagement than others. However, as shown in the literature [99], current cluster algorithms suffer the decision of the number of clusters beforehand. Moreover, finding hot topics is challenging [100], since they could be small and lost in a big cluster. Thus, we prefer a Nearest Neighbors approach to find neighborhoods of points with similar engagement. In particular, we first divided our posts into five engagement classes determined by the percentiles [0-20, 20-40, 40-60, 60-80, 80-100], saving the thresholds of each percentile. Then, for each point, we search its N nearest neighbors, calculate their engagement average, and see whether the average falls in the same engagement class as the point under consideration. If so, that neighborhood is considered “pure”, and new posts falling in it would likely produce that particular engagement class. To find the nearest neighbors, we first reduced the dimensionality of the embeddings using PCA (100 components), and then applied the Nearest Neighbor algorithm leveraging Scikit-Learn implementation [101] using

Euclidean Distance as the distance metric. Figure 3.5 depicts the percentage of pure neighborhoods for different $N = [1, 3, 5, 10, 20, 30, 50]$ for the mid-tier. On average, around 20% ($N=50$) of the points are in a pure neighborhood, which suggests that some topics are more (or less) engaging than others. The pet category presents the highest average, probably because its topics can be the species and breed of animals (visually similar), and some could be liked more (or less) than others.



(a) Likes



(b) Comments

Figure 3.5: Percentage of pure neighborhoods in mid-tier for engagement metrics.

3.4.2 GENERAL VS USER-SPECIFIC HOT TOPICS

The percentage of the pure neighborhoods found in Section 3.4.1 is comprehensive of neighborhoods made only by a single influencer, i.e., not a general hot topic. In this case, we identified what we could call a user-specific hot topic, which is very useful for understanding what topic is engaging (or not) for that particular influencer. Thus, what differentiates general vs user-specific hot topics is how many influencers participate in a pure neighborhood, and with how many posts. We call this parameter *User Diversity*. To calculate it, we took inspiration from the Simpson’s Diversity Index [102], used in ecology to quantify the biodiversity of a habitat. It takes into account the number of species present, as well as the abundance of each species. The diversity index D is expressed as:

$$D = 1 - \frac{\sum_{i=1}^K n_i(n_i - 1)}{N(N - 1)}, \quad (3.2)$$

where N is the total sample size, K the number of species, and n_i is the number of organisms of the i^{th} specie. D ranges from 0 to 1, where 0 is minimum diversity and 1 is maximum diversity. In our scenario, the species are the influencers, and the organisms are the posts. Similarly, we can define an *Engagement Diversity*, which measures the posts’ diversity in terms of engagement. This metric is needed since we created pure neighborhoods based only on the average engagement of their posts; therefore, there can be posts belonging to different engagement classes within a pure neighborhood. To recap, by measuring the *Engagement* and *User Diversity* of our pure neighborhoods, we can define topics as depicted in Figure 3.6. We are more interested in the green part, since neighborhoods with high engagement diversity are less reliable. We set the threshold between low and high at 0.5.

		Engagement Diversity	
		Low	High
User Diversity	Low	User-specific Hot Topics	User-specific Variable Topics
	High	General Hot Topics	General Variable Topics

Figure 3.6: Types of topics for Engagement and User Diversity.

FINDINGS. We concentrated our research of hot topics only on the highest engagement class (i.e., posts falling in the top 20% percentile), focusing on each tier and category differently. During the automatic search, we removed neighborhoods that overlapped for more than 80% of the points. We explored the resulting neighborhoods and found several hot topics, which we did not think about in the feature extraction phase, and could not be detected by SotA models, confirming the benefits of this unsupervised research. All the neighborhoods can be browsed on our repository, while in Figure 3.7 we reported an example for each category. For instance, we found “mother with her child” for Family, “two (or more) girls in bikini” for Fashion, or “girl/kid near/riding a horse” for Pet. For captions, we found less category-specific hot topics and more common strategies. For instance, giveaways attract



Figure 3.7: Examples of hot topics found in our categories.

a lot of comments, since participants usually have to comment and tag other friends. Furthermore, we found the working strategy of asking users' opinions on a topic, general questions, or requests for upcoming content. More general hot topics are presented in Section 3.5.

Takeaway: *Instagram offers both visual and textual hot topics that are likely to generate high engagement levels. Captions tend to use similar strategies across categories despite visual hot topics being category-specific.*

3.5 GUIDELINES INSIGHTS

This section provides some guidelines to get more Likes and Comments for each category, resulting from both DT engagement classifiers (Section 3.3) and hot topics detection (Section 3.4). We also provide some suggestions to make an engaging caption.

3.5.1 GUIDELINES FOR LIKES, COMMENTS, AND TOPICS

Each category presents different characteristics to build engagement. We now present guidelines to get more Likes and Comments along the hot topics we found.

BEAUTY. Likes are mainly driven by exposed buttocks and feet, a high image pleasure, and positive emoji sentiment. Exposed buttocks also generate many comments, as well as a low age average, having the location set, and the use of many mentions. Wavy hair is much appreciated, and hot topics include couples and eye make-up with perfect eyebrows. Users usually love when the influencers receive new make-up products, recreate famous make-up (e.g., from movies), and talk about personal problems.

FAMILY. Likes and Comments are driven by similar factors. People's features like age and gender are predominant. The mean age of female subjects should be low, with a high standard deviation. This suggests that mothers with children are a hot topic, as detected in Section 3.4.2. Indoor or outdoor-natural environments are preferred, and location, colors, and the number of mentions are highly impacting. As hot topics, we found pregnancy, childbirth, and body changes, during which followers feel closer to the influencer.

FASHION. In terms of Likes, a higher number of mentions is suggested for small tiers, whereas colors-related features (e.g., dominance, arousal) contribute heavily to high-tier influencers. A predominant role is held by exposed buttocks, which contribute to both likes and comments. Exposed feet generate many comments, as well as outdoor pictures, short captions with many hashtags, and positive emoji sentiments. As hot topics, girls in bikinis and men with six-packs are successful. Discussing outfits for special events is highly engaging, such as traveling, going to concerts, birthdays, gallant dinners, or simply starting the week.

FITNESS. Likes are driven by warm colors, and high dominance of female subjects with low age standard deviation, preferably in their workout outfit. A short caption with positive sentiment helps in receiving both likes and comments. Low arousal generates many comments, according to the body transformation hot topic. The caption should motivate people to try harder in their workouts.

FOOD. Males are more common and generate more likes in this category. Extreme burgers and spirits are highly appreciated, as well as perfect and very colored food compositions. Pictures in kitchens or outside restaurants help to get likes. The location is important for getting comments, as well as high arousal and a positive caption with many Emojis. The caption should include a brief description of the plate and questions about the favorite food. Pizza days, chocolate, and vegan food are often in the middle of heated discussions.

INTERIOR. To get likes, indoor environments like a living room and cold colors are preferred, as well as the presence of kids and female subjects. The location is relevant for both likes and comments, but avoid commercial buildings and food pictures. Luminous and pleasant pictures generate more comments, as well as the presence of animals. A good caption combines positive Emoji sentiment, a few hashtags, and general questions, like what to do on the weekend.

PET. Pictures should be in an indoor or outdoor-natural environment, use warm colors, and convey a positive sentiment to get many likes. The use of the location and mentions helps a lot for comments, as well as very high or very low animal cuteness. Among the most loved animals, we found horses, exotic animals, Siamese cats, and dogs with clothes and ribbons. Many comments will arrive along with a new family member!

TRAVEL. Likes are gained primarily by female subjects and a low number of male presence, with a generally low age standard deviation and a low minimum age. This suggests that travel pictures of young friends or groups of the same age are highly engaging. To get many comments, besides the importance of outdoor-related features, the sentiment conveyed by the text should tend to be positive. Further, hashtags and mentions are crucial, and the picture should be pleasant and arousing, with low dominance. Pictures near the sea are highly appreciated, and users engage more with summer and holiday posts.

OTHER. More likes are obtained by females in indoor places with cold colors and low dominance. Many emojis and short captions help too. Men of the same age in a single picture get many comments, using a neutral sentiment in the caption. Hot topics are many, for example, football, memes, or superheroes. The captions tend to be funny and quite short.

3.5.2 GUIDELINES FOR ENGAGING CAPTIONS

Even if each category and tier require a specific caption to build engagement, we identified some common strategies to generate highly engaging ones. From our explorations, we identified a typical pattern among most hot captions, i.e., asking questions to the audience. Such questions can be very generic (e.g., “how do you feel today?”), or topic-specific (e.g., “which outfit do you prefer?”), which helps engage the users. Moreover, creators often ask people to perform particular actions, such as tagging or sharing content with friends. This behavior, known as *call to action*, usually generates a lot of engagement. Last, hashtags are generally at the end of the caption, often separated by the rest of the text with one point or dash per line. This behavior forces the users to click on the “View more” button to see the whole caption, generating more engagement.

3.5.3 LIMITATIONS

Our guidelines are the result of analyzing the biggest IG dataset ever released, composed of around 10M posts created by 34K influencers. Nevertheless, it does not include many categories of interest (e.g., sport, cinema, music), and what people like as well as hot topics could change over time. However, we presented two methodologies (supervised Section 3.3 and unsupervised Section 3.4) that can be applied to any category (possibly enhancing the feature set) at any time, by taking a “snapshot” of IG content produced by influencers of a target category and tier. Moreover, a reader might be concerned that the IG algorithm started considering comments to recommend posts in 2021, whereas our dataset is from 2020. We remark that this chapter’s aim is to explain which post’s features induce users to generate more comments (which now are a crucial factor), and not how IG is now recommending posts to users. Indeed, as for likes, the main reason users leave comments is based on the posts themselves [66], not whether users see the posts.

3.6 CONCLUSION

In this chapter, we aimed to close the gap from previous works, explaining the underlying mechanisms of IG engagement and focusing on interpretable models. In this way, it is possible to create engaging IG content by design, following predefined guidelines saving time and money. Through a careful and all-inclusive process of feature extraction, we trained predictors that achieved up to 94% of F1-Score. In particular, our results show that likes are mainly driven by images, while comments are primarily stimulated by captions. Further, we demonstrated how influencers' behavior becomes more category-specific as their tier increases. Last, we proposed a novel unsupervised approach for detecting and analyzing hot topics, to better understand the inner dynamics of each category.

In the future, we plan to improve the predictions through a model that integrates hot topic extraction. Furthermore, more categories should be studied, and a metric that combines likes and comments should be introduced to better understand their relationship. Regarding these metrics, they could be polished by removing fake engagement, for instance. However, as of now, we have no evidence IG algorithm is accounting for such differences. Last, geography should be taken into account to understand whether it impacts engagement mechanisms.

ETHICAL CONSIDERATIONS

We did not collect any data in this work. All the data we have used has been legitimately collected in previous work using Instagram API [70]. These data may be shared with researchers upon request to advance the field of research, and cannot be used in any other manner (e.g., for business). Images reported in this chapter have only been used for research and demonstration purposes. Human subjects in the pictures are all Instagram influencers, i.e., public figures. Anyhow, we carefully blurred their faces in order to make them unrecognizable.

4

Climbing the Influence Tiers on TikTok: A Multimodal Study

Unlike most social media, TikTok pays \$0.02-0.04 for every 1000 views that a post gets on the platform.¹ This provides a huge incentive to create videos that garner lots of views. While academia abounds with “models of influence” [103], industry and influencers in the TikTok ecosystem appear to pay more attention to “tiers” of influence which are based on followers counts. As in the case of Instagram (Chapter 3), TikTok creators are divided into Nano influencers who have [1K, 10K) followers, and Micro [10K, 50K), Mid [50K, 500K), Macro [500K, 1M), Mega [1M, +∞] [104] influencers. In addition to revenue from TikTok’s payments, influencers often get marketing opportunities directly from companies [105]. A partnership with Charlie D’Amelio, one of the most followed people on TikTok, is estimated to cost more than \$100,000 per post [106]. On the other hand, Nano or Micro-influencers tend to have a more homogeneous follower base [105]. In the literature, there is evidence that influencers differ in their behavior and engagement depending on their tier [107, 108, 109, 104], but it is unclear to what extent.

Thus, these influencers have a huge incentive to “move up” from their current tier to the next higher one. In this chapter, we study what distinguishes TikTok influencers in one tier from those in the next one up. The resulting findings may help influencers increase their reach and revenues.

To achieve this, we first created the TikTok Influencer Dataset for Exploratory Study (or TIDES) containing data on 5000 influencers, 230,406 videos, and 10,294 audio clip, by combining data from TikTok and Spotify.² To our knowledge, TIDES is the first ever publicly available dataset that analyzes TikTok influencers. Next, we

¹<https://influencermarketinghub.com/how-much-does-tiktok-pay/>

²The TIDES dataset and associated code will be released when the corresponding paper is published. In the case of video content that we are not permitted to distribute due to terms of use, we will provide the links along with code to hydrate those links.

developed a multimodal set of features associated with each of our 5000 influencers. Finally, we trained classifiers to predict, for each tier t other than Mega, whether an influencer would belong to tier t or the next higher tier. We also performed ablation studies to find two types of important features: (i) the most important of the entirety of features that separate influencers in tier t from those in the next higher tier, and (ii) the most important “actionable” features, i.e. features whose values can be directly modified by an influencer in order to reach the next tier.

We are able to distinguish Nano from Micro influencers with an F1-score of over 0.84, Micro from Mid with an F1-score of over 0.83, Mid from Macro with an F1-score over 0.80 and Macro from Mega with an F1-score over 0.82. This suggests that influencers across adjacent tiers are relatively easy to separate. Interestingly, using only the top 50 most relevant features only drops these F1 scores by 1-2%.

On the other side, an influencer cannot directly change every feature, like the number of followers she has and the number of likes/views her posts receive. These are *not actionable* features. However, she can choose music with a higher danceability score (according to Spotify³) or look happier in her videos (according to video emotion classifier [110]). These are *actionable* features.

We also answered the following research questions:

RQ1: Do multimodal features (*audio, video, text*) make any difference compared to *traditional* features (e.g., likes, profile information) when separating users in one tier from the next higher tier? When we conducted an ablation study by looking at each of the 4 classification problems studied, we saw that there was a 2.95% improvement in F1 for Nano-Micro, 1.6% for Micro-Mid, 0.4% for Macro-Mega, but no improvement for Mid-Macro.

RQ2: When considering only *actionable* multimodal features (*audio, video, text*), which ones make the biggest difference when separating users in one tier from the next higher tier? Our ablation study showed that *video* features make the biggest difference (5.56% and 4.15%) in F1 for Nano-Micro and Mid-Macro respectively, while *traditional* features (3.43% and 6.66%) for Micro-Mid and Macro-Mega respectively. *Text* features only have an impact (1.31%) in the Macro-Mega case, while *audio* features only have a 1-2% impact in all cases except for Macro-Mega. Simply put, this suggests that *video* and *traditional* features are the most important ones, with *audio* features also being important and *text* features being the least important.

RQ3: Which are the most important *actionable* features that an influencer can change in order to move from her current tier to the next one? Do these depend on the tier she is currently in or do they vary by tier? We found influencers should primarily focus on *traditional* features (e.g., publishing videos regularly and frequently) and *video* features (e.g., producing more pleasant and high-quality videos). How influencers should change their behavior is linked to their current tier and the one directly above it. For Instance, for Macro influencers to become Mega influencers, having a verified profile is roughly eight times more important than increasing their total number of videos.

4.1 RELATED WORK

In this section, we review related works concerning general TikTok studies, as it is a (relatively) new social media platform in research, as well as TikTok Influence studies.

³<https://developer.spotify.com/documentation/web-api/reference/get-audio-features>

TikTok Studies. TikTok use rose dramatically during the COVID-19 pandemic [111] — hence, several efforts have examined the dissemination of information via TikTok throughout that timeframe [112, 113, 114]. Another topic of interest revolves around virality and engagement. Researchers found that close-up or medium-shot videos and videos containing text have a higher chance of going viral [115]. Additionally, high user engagement metrics such as likes, comments, or shares are pivotal in propelling videos to popularity. Contrary to expectations [116], joining a trending hashtag bandwagon seems less irrelevant. Further studies focused on education [117, 118], politics [119], and the negative sides of the platform, such as privacy issues [120], cyberbullying [121], and hate speech [122].

TikTok Influence. Influence maximization in social networks [103, 123] has been studied extensively. Custom models on how influencers should behave have been developed for Twitter [124], Facebook [125, 126], YouTube [127], and Instagram [128, 129], but to the best of our knowledge, this is not the case with TikTok. As a platform born mainly to have fun, TikTok influencers’ sense of humor, entertainment, and happiness increase the effectiveness of messages spread on TikTok [130, 131]. Furthermore, TikTok influencers must continuously communicate with their audience and foster parasocial relationships to gain more followers. In the context of influence marketing, [132] developed an algorithm to predict the increase in sales resulting from sponsored videos, while disclosure of sponsorships (usually kept hidden) does not affect brand results.

4.2 OUR NEW TIDES DATASET

For our study, we assembled our new TikTok Influencer Dataset for Exploratory Study (or TIDES), a central contribution of this chapter that we are pleased to release. To the best of our knowledge, TIDES is the first ever publicly available dataset that analyzes TikTok influencers. We collected information and videos from 5000 TikTok Influencers, 1000 for each tier described above (Nano, Micro, Mid, Macro, Mega) using lists of influencers from HypeAuditor⁴ and StarNgage⁵, two famous influencer analytics websites. To limit bias in our experiments, we first collected all the influencers listed on these websites, and then randomly sampled 1000 from each tier. As the information on these websites may not always be up-to-date, we verified the follower counts of each profile by cross-checking them with TikTok’s data, properly categorizing each influencer with their tier. We included influencers from the five major continents: America, Asia, Africa, Europe, and Oceania, with 1,000 from each continent.⁶ After we retrieved the list of influencers, we used a TikTok scraper on Apify⁷ to collect details of their profiles (e.g., the number of videos) and the information of their (up to) 50 most recent videos, including the URLs. Using these URLs, we downloaded the videos without the TikTok watermark⁸ through a web scraper we wrote in Python using Selenium library⁹. We collected our data in June 2023, obtaining a total of 230,406

⁴<https://hypeauditor.com/top-tiktok/>

⁵<https://starngage.com/plus/en-us/influencer/ranking/tiktok>

⁶The actual influencers distribution across continents may vary, but we have no access to official data.

⁷<https://apify.com/clockworks/tiktok-scraper>

⁸The TikTok logo is added randomly to every video when uploaded to the platform, and could have influenced the analyses.

⁹<https://selenium-python.readthedocs.io/>

videos. To sum up, we obtained two datasets: the Influencer dataset \mathcal{I} ($|\mathcal{I}| = 5000$) and the Video dataset \mathcal{V} ($|\mathcal{V}| = 230406$).

4.3 FEATURE EXTRACTION

During data collection, \mathcal{I} was populated by profile-related features for each influencer, while \mathcal{V} included meta-features for each video (e.g., the number of likes, the music used). We augmented \mathcal{I} by calculating behavioral features and \mathcal{V} by extracting content-related features, particularly from audio, video, and text (the caption), implementing several Deep Learning models. We now provide an overview of each set of features.¹⁰

4.3.1 TRADITIONAL FEATURES

Traditional features typically originate from the platform and do not include content-specific information. In \mathcal{I} , they encompass the data derived from the Influencer profile, e.g., follower and following counts, total number of likes and videos, geographical location, whether the user has a bio or a URL. We augmented \mathcal{I} by calculating behavioral features, such as the distribution of videos by day of the week, videos published per day, and the inter-posting time (i.e., average time and standard deviation between each video). The video dataset \mathcal{V} includes, for each video, whether it is sponsored, and engagement metrics such as the number of likes, views, shares, and comments.

4.3.2 AUDIO FEATURES

The audio is a fundamental component of TikTok videos. Indeed, many videos are based on trending music, dances, lip-syncing, or interactions where the influencer communicates with their followers. Therefore, we extracted features to understand how influencers use audio in their videos, relying on two channels: Spotify and the raw Audio Channel.

SPOTIFY FEATURES. The audio can either be original or from another artist. In the latter case, we relied on the Spotify platform to extract additional features. Using the `search` endpoint of Spotify API¹¹, we first verified whether the track existed on the platform, obtaining the Spotify handle (`id`). Through the handle, we first called `tracks/id` API to retrieve general track information like the popularity (in Spotify) and whether it is explicit. We then invoked the `audio-features/id` API to obtain audio features like danceability, speechiness, or instrumentality, which are calculated by proprietary Spotify algorithms. These features serve as effective descriptors of the music, allowing us to gain deeper insights into the types of videos Influencers share.¹² In total, we collected information about 10,294 tracks, often shared among the collected videos.

¹⁰The complete set of features will be available on our repo <https://anonymous.4open.science/r/icwsm-TikTok-03FA>.

¹¹<https://developer.spotify.com/>

¹²A complete explanation of the Spotify features is available at: <https://developer.spotify.com/documentation/web-api/reference/get-audio-features>.

Category	Feature	Category	Feature	Category	Feature	Category	Feature	Category	Feature
Traditional	has_bioLink*	Audio	audio_is_original*	Audio	audio_arousal*	Video	video_age*	Video	video_RGB_percent_cold*
Traditional	has_signature*	Audio	spotify_acousticness*	Audio	audio_dominance*	Video	video_race*	Video	video_RGB_pleasure*
Traditional	commerceUser*	Audio	spotify_danceability*	Audio	audio_valence*	Video	video_surprise*	Video	video_RGB_arousal*
Traditional	likes	Audio	spotify_energy*	Audio	audio_neutral*	Video	video_sad*	Video	video_RGB_dominance*
Traditional	comments	Audio	spotify_instrumentalness*	Audio	audio_happy*	Video	video_fear*	Video	video_hasText*
Traditional	views	Audio	spotify_liveness*	Audio	audio_sad*	Video	video_neutral*	Text	text_n_hashtags*
Traditional	shares	Audio	spotify_speechiness*	Audio	audio_angry*	Video	video_angry*	Text	text_n_mentions*
Traditional	total_likes	Audio	spotify_valence*	Video	video_height*	Video	video_happy*	Text	text_CaptionLength*
Traditional	total_videos*	Audio	spotify_mode_o*	Video	video_width*	Video	video_disgust*	Text	text_EmojisLength*
Traditional	continent*	Audio	spotify_explicit*	Video	video_heightXwidth*	Video	video_n_faces*	Text	text_EmoNeutral*
Traditional	weekdays*	Audio	spotify_time_signature*	Video	duration*	Video	video_RGB_r*	Text	text_EmoPositive*
Traditional	videos_per_day*	Audio	spotify_key*	Video	video_definition*	Video	video_RGB_g*	Text	text_EmoNegative*
Traditional	interposttime*	Audio	spotify_popularity*	Video	n_effects*	Video	video_RGB_b*	Text	text_EmoSentScore*
Traditional	following*	Audio	spotify_loudness*	Video	FAUs (video_AU[XX])*	Video	video_RGB_luminance*		
Traditional	videos_liked*	Audio	spotify_tempo*	Video	video_gender*	Video	video_RGB_percent_warm*		

Table 4.1: Category wise list of features used. Actionable features are highlighted with a “*” symbol.

EMOTION FEATURES. We wondered whether the emotions (e.g., angry, happy, sad, neutral, VAD emotional state [133] attributes such as valence, arousal, and dominance) were important in distinguishing between influencer tiers. We extracted the basic emotions by implementing a fine-tuned wav2vec2 model [134] using SpeechBrain [135]. We also implemented the VAD model [136].

4.3.3 VIDEO FEATURES

To analyze video content, we extracted features at the image level and then aggregated the results (see Experiment Section). As in previous work [137], we extracted two frames per second per video, obtaining a total of 13,929,447 frames. We first extracted the percentages of red, green, and blue channels. Next, using the HSV (Hue, Saturation, Value) representation, we calculate the percentages of luminance, warm and cool colors [82], as well as pleasure, arousal, and dominance scores [83]. We then used several state-of-the-art deep learning models to analyze human subjects. Starting by detecting people’s faces through Retinanet [138], we extracted Facial Action Units [139] using Py-Feat framework [140], Age and Gender through Dlib library [78], and race and emotion through DeepFace [110]. Last, from the entire video, we extract the definition, width, height, duration, and if it contains text (e.g., subtitles). This last feature was proven useful in recent studies [115], and we extracted it through Tesseract.¹³

4.3.4 TEXT FEATURES

This set of features derives from the video caption. We extracted the caption length, the number of Emojis, and their sentiment [87] using the EmoSent library¹⁴. The library returns a sentiment score ranging from −1 (negative) to +1 (positive), and three values representing the probabilities of being positive, negative, and neutral. In addition, we extracted the number of hashtags and mentions. In total, we had 73 features of which 68 were actionable. A complete list is available in Table 4.1.

¹³<https://tesseract-ocr.github.io/>

¹⁴<https://github.com/omkar-foss/emosent-py>

4.4 DESCRIPTIVE STATISTICAL EXPLORATION OF SPECIFIC FEATURES

Looking at the set of feature discussed in the previous section (and listed Table 4.1) we were curious to make interesting hypotheses: we did and tested them through Student t-test. In this section, we discuss these hypotheses, focusing first on the non-actionable features and then the actionable ones.

4.4.1 NON-ACTIONABLE FEATURES

In general, we wondered whether engagement by other users is a proxy for the tier to which an influencer belongs. We examine four hypotheses that model this general intuition.

Hypothesis 1 *We hypothesize that the total number of likes that a user gets is linked to the tier to which the influencer belongs.*

Figure 4.1 shows a box plot whose x-axis represents the tiers and y-axis shows the total number of likes an influencer got from all her videos. We see that influencers who are at higher tiers generally have a larger number of likes. For each adjacent pair of tiers, the hypothesis that influencers at the higher tier get more likes is statistically valid via a t-test with $p < 10^{-36}$. One reason for this could be that influencers in higher tiers post more videos than those at lower tiers. To check this, we explored a derivative hypothesis.

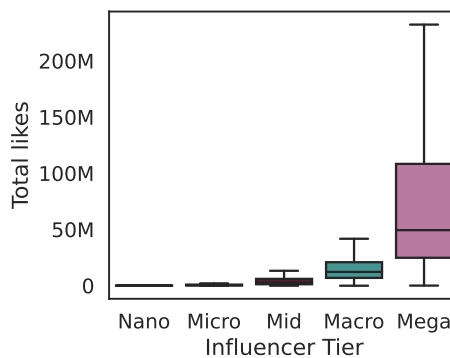


Figure 4.1: Box-plots of the total number of likes per tier.

Hypothesis 2 *We hypothesize that the average number of likes per video that a user gets is linked to the tier to which the influencer belongs.*

Figure 4.2 depicts a box plot where the y-axis indicates the mean number of likes garnered by each video, while the x-axis reflects the tier of the influencer. In general, influencers in higher tiers tend to receive a greater average number of likes for each video. The substantial discrepancy in the number of followers between the higher tiers and the lower tiers supports this observation. Hence, it is reasonable to conclude that higher tiers exhibit a greater mean number of likes. For each adjacent pair of tiers, the hypothesis that influencers at the higher tier get more likes per video is statistically valid via a t-test with $p < 10^{-5}$.

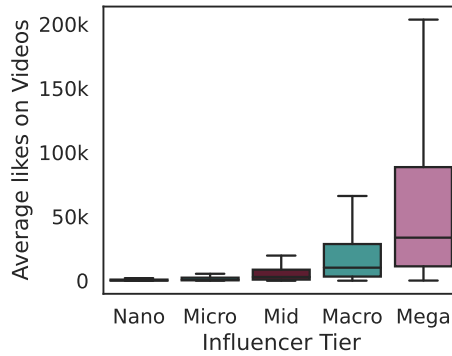


Figure 4.2: Box-plots of the average number of likes per tier.

Hypothesis 3 *We hypothesize that the total number of comments that a user gets is linked to the tier to which the influencer belongs.*

Figure 4.3 shows a box plot showing the average number of comments per user by tier. The hypothesis that users in higher tiers receive more comments is true for all pairs of tiers with $p < 10^{-7}$.

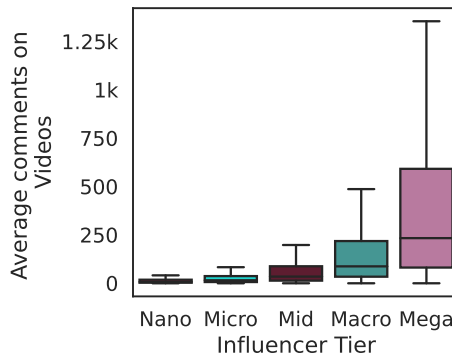


Figure 4.3: Box-plots of the average number of comments per tier.

Hypothesis 4 *We hypothesize that the total number of views that a user gets is linked to the tier to which the influencer belongs.*

Figure 4.4 shows a box-plot showing the average number of views per user by tier. The reader can readily see that users in higher tiers receive more views — this is validated by a t-test with $p < 10^{-17}$

Simply put, Figures 4.1, 4.2, 4.3 and 4.4 show that greater engagement by the audience (other users) is linked to the tier to which an influencer belongs. This is not very surprising.

4.4.2 ACTIONABLE FEATURES

The preceding subsection looks at features that an influencer cannot directly influence because she cannot (at least honestly) increase the number of likes her videos get, the number of comments made on her videos, and so

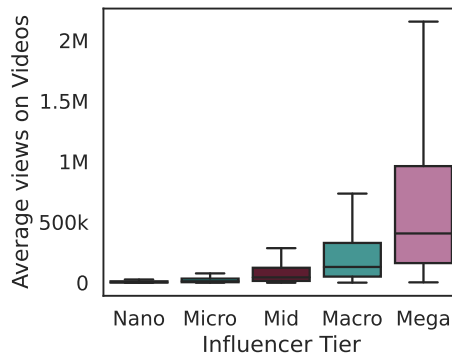


Figure 4.4: Box-plots of the average number of views per tier.

forth.¹⁵ In this section, we look at *actionable* features — these represent features that an influencer can directly change through his actions.

Hypothesis 5 *We hypothesize that the total number of videos posted can directly influence the tier to which the influencer belongs.*

Figure 4.5 shows a box plot whose x-axis represents the tiers and whose y-axis shows the number of videos posted by the influencers in each tier. Higher-tier influencers post more videos than lower-tier influencers. One possible reason could be that higher-tier influencers have been in the network for a long time and, hence, have more videos. This findings holds for all pairs of adjacent tiers with $p < 0.02$.

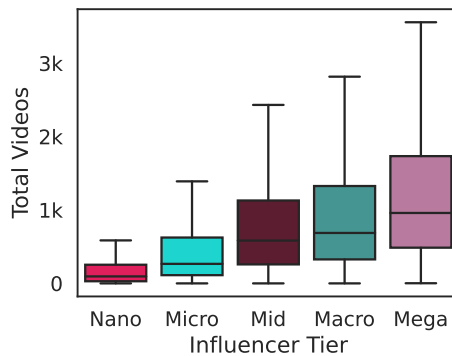


Figure 4.5: Box-plots of the total number of videos per tier.

Hypothesis 6 *We hypothesize that the average inter-posting time between videos posted can directly influence the tier to which the influencer belongs.*

¹⁵In the real world, influencers might do unethical things (e.g. create fake accounts to artificially inflate the number of views, likes, comments on their videos). In this chapter, we assume that influencers are honest and do not take such unethical actions.

Figure 4.6 shows a box plot whose y-axis represents the average inter-post time between videos posted and whose x-axis shows the influencers from each tier with that inter-posting time. We verified this hypothesis using the Student t-test, and which shows statistical significance for all pairs of adjacent tiers ($p < 10^{-4}$) except for Mid-Macro. As reported in the related hypothesis that, this feature is negatively correlated with the ‘total videos’ (pearson correlation: -0.235). The same can be observed by seeing Figure 4.5 and Figure 4.6.

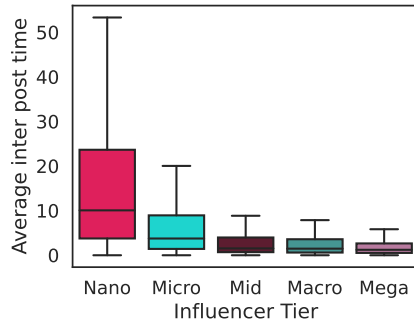


Figure 4.6: Box-plots of the average inter-posting time per tier.

Hypothesis 7 *We hypothesize that the total number of users that an influencer follows is linked to the tier to which the influencer belongs.*

Figure 4.7 shows the number of users that an influencer follows. The x-axis shows the tiers, and the y-axis denotes the number of users that an influencer follows. As shown in the figure, lower-tier influencers follow more users than those in higher tiers. This might suggest that higher-tier influencers cultivate an aura of exclusivity by limiting the number of accounts they follow. This hypothesis is statistically significant for Nano-Micro, Micro-Mid, and Mid-Macro pairs with $p < 0.003$.

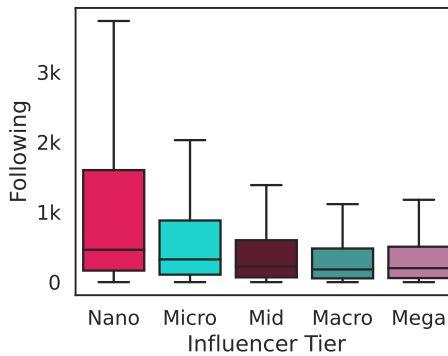


Figure 4.7: Box-plots of the number of users an influencer follows per tier.

Hypothesis 8 *We hypothesize that the total number of videos that an influencer likes is linked to the tier to which the influencer belongs.*

Figure 4.8 shows a box plot whose x-axis shows the tiers and y-axis represents the number of videos an influencer has liked of other users. The results shown are inconclusive. This hypothesis is true only for the Macro-Mega with $p < 0.04$.

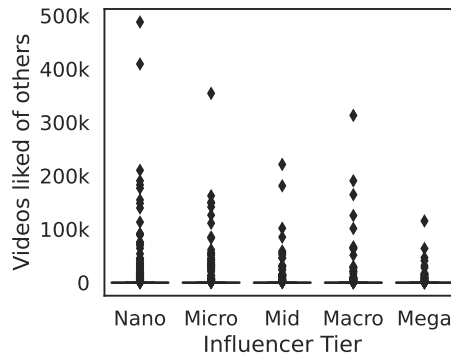


Figure 4.8: Box-plots of the number of videos liked per tier.

Hypothesis 9 *We hypothesize that influencers who post their bios publicly influence are more likely to belong to higher tiers.*

Figure 4.9 shows a bar plot whose x-axis shows the tiers and whose y-axis shows the number of influencers that have (or do not have) a public bio link in their profile for each tier. This figure shows that the probability of influencers having a posted bio increases as influencers reach higher tiers. This finding is verified for all pairs of tiers with $p < 0.0003$.

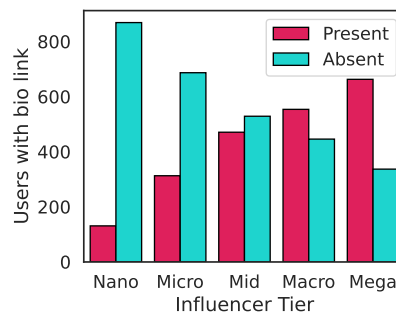


Figure 4.9: Presence of bio link info.

Hypothesis 10 *We hypothesize that influencers tend to post videos associated with their respective tiers on specific days of the week.*

Figure 4.10 illustrates the mean proportion of videos an influencer uploads on weekdays. During weekends, influencers from the Nano tier exhibit a higher posting frequency than influencers from other tiers. Conversely, on workdays (not weekend), influencers across all tiers maintain a similar frequency of video posting. The Student t-test shows that this finding is valid for some pairs of tiers on specific days, and for others, it is not. For example, on Saturday and Sunday, it is valid for Nano-Micro and Micro-Mid ($p < 0.003$) but not for others.

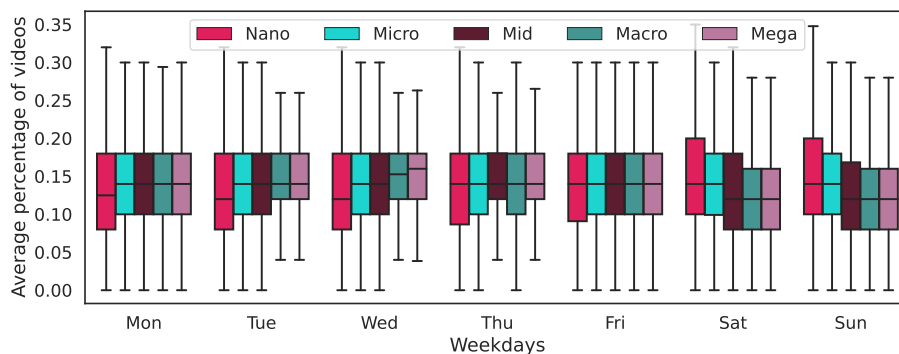


Figure 4.10: Box-plots of the average percentage of videos an influencer post per tier per week of the day.

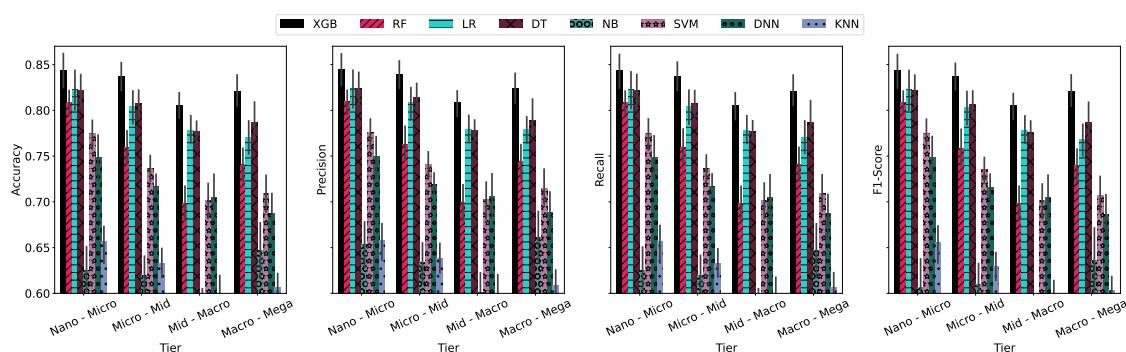


Figure 4.11: Classification results to predict the influencer tier using several classifiers.

4.5 INFLUENCER TIER CLASSIFICATION

To answer our research questions, we trained multiple ML classifiers that differentiate influencers across consecutive tiers. Our initial phase involved the aggregation of video data for each influencer. We then rigorously trained and tested several classifiers to identify the best one. We further conducted a series of ablation studies using the optimal classifier (for each of the four classification problems, i.e., Nano vs. Micro, Micro vs. Mid, Mid vs. Macro, Macro vs. Mega) to provide conclusive answers to our research questions. We remind the reader that the primary objective of this chapter is to identify the most valuable (*actionable*) features for advancing through influencer Tiers. Therefore, we are building high-quality classifiers but not striving for perfect predictions, which could be a future research direction.

4.5.1 DATA AGGREGATION

We conducted our machine learning experiment at the user level. Each influencer corresponds to a row in our dataset. Therefore, for each influencer, we aggregated all the information about their videos into a single entry. We used different aggregation types depending on the variable type (e.g., boolean, categorical, float). For boolean variables (e.g., if the audio is original), we computed the percentage of true values, i.e., the percentage of videos

posted by the user that used original audio. Likewise, for categorical variables, we calculated their respective percentages (e.g., for the posting day, we calculated the percentages of videos published on Monday, Tuesday, and so on). When dealing with numeric values, we computed several statistical metrics, including the mean, standard deviation, maximum, and minimum. We also used distribution features associated with a feature f : we found the minimum and maximum values of f across all influencers and divided the $[min, max]$ range into ten bins. We then computed the probability of a user's video having a value within each of the ten bins. It is worth noting that these ten bins were computed only on training data to prevent data leakage. Two additional bins were used to accommodate values below the minimum and above the maximum. For likes, comments, views, and shares, which exhibit vast value ranges, we opted for a six-bin approach (instead of the 10-equidistant bins) inspired by the box plots shown earlier in the chapter (excluding outliers). The boundaries were $-\infty$, minimum, first quartile, median, third quartile, maximum, $+\infty$. Given that our *video* features are extracted from frames, we aggregated video-related values for each influencer by considering all frames from all their respective videos as a whole. For instance, to calculate the video pleasure feature of an influencer i , we extracted all the frames of all the videos made by i , calculated the pleasure value on all these frames, and aggregated the results using the ten bins approach, defined through the minimum and maximum video pleasure of all the other influencers in the training set. Last, we calculated minimum, maximum, mean, and standard deviation video pleasure values related to the influencer i only. In total, we ended up with 839 features per influencer.

4.5.2 ALL FEATURES ANALYSIS

To differentiate between consecutive tiers of influencers, we trained eight machine learning classifiers: XGBoost (XGB), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Deep Neural Network (DNN), and K-Nearest Neighbors (KNN). We validated our results using nested 10-fold cross-validation (CV). We found the best hyperparameters through a grid-search approach, testing the following values:

- XGBoost (XGB): $eta=[0.01, 0.1, 0.3]$, $max_depth=[None, 3, 6, 12]$, $min_child_weight=[1, 5, 10]$, $gamma=[0, 1, 5]$, $colsample_bytree=[0.6, 0.8, 1.0]$;
- Random Forest (RF): $n_estimators=[50, 100, 250]$, $criterion=[gini, entropy]$, $max_depth=[3, 5, 7]$, $min_samples_leaf=[1, 3, 5]$;
- Logistic Regression (LR): $penalty=[l1, l2, none]$, $C=[0.1, 1, 10]$, $solver=[liblinear, lbfgs]$;
- Decision Tree (DT): $criterion=[gini, entropy]$, $max_depth=[3, 5, 7]$, $min_samples_leaf=[1, 3, 5]$;
- Naive Bayes (NB);
- Support Vector Machine (SVM): $C=[0.1, 1, 10, 100]$, $kernel=[linear, poly, rbf]$, $gamma=[scale, auto]$;
- Deep Neural Network (DNN): $hidden_layer_size=[64, 128, 256]$, $activation=[tanh, relu]$, $solver=[sgd, adam]$;
- K-Nearest Neighbors (KNN): $n_neighbors=[3, 5, 7, 10]$, $weights=[uniform, distance]$.

Figure 4.11 shows the results of the classification. XGBoost outperformed all the other classifiers, reaching more than 80% F1-score in every tier. In particular, the easiest task is to differentiate between the Nano and Micro tiers (84.3% F1-Score), while the most challenging concerns Mid and Macro (80.5% F1-Score).

Given its superior performance, we continued our experiments using XGBoost. We repeated the experiments by selecting the 10, 20, and 50 most important features at training time (out of 839), using Anova F-values (Table 4.2). Interestingly, using 10, 20, or 50 features only drops the F1-Score by 1-3%, except in the case of 10 features for Macro-Mega, which we have a drop of 11%. This suggests that substantially more than 10 factors distinguish Macro and Mega influencers, unlike the other tiers.

Table 4.2: F1-Score of XGB Classifier using top N Features.

N Features	Nano-Micro	Micro-Mid	Mid-Macro	Macro-Mega
10	0.837 \pm 0.023	0.804 \pm 0.040	0.781 \pm 0.025	0.709 \pm 0.037
20	0.836 \pm 0.025	0.800 \pm 0.034	0.781 \pm 0.018	0.790 \pm 0.020
50	0.831 \pm 0.029	0.807 \pm 0.028	0.792 \pm 0.025	0.793 \pm 0.026
All	0.843 \pm 0.031	0.837 \pm 0.025	0.805 \pm 0.024	0.821 \pm 0.029

To inspect these factors, we explored the feature importance of our best classifier, XGBoost. Given our 10-fold CV approach, we have 10 best classifiers for each result. We thus mediated the feature importance of these 10 classifiers to calculate the final feature importance. Figure 4.12 shows the feature importance of the algorithm in the different tiers. We grouped the features based on their typology (e.g., video dimension and definition are under video quality), and ranked the groups based on the value of their most important feature (the value is reported in the graph). This allows us to maintain the original feature-importance ranking, giving a general idea of the type of features, and preventing some groups from being disadvantaged by considering many irrelevant aggregation features. For instance, 10-bins features in which one bin is very relevant, and the others are not, would be disadvantaged compared to boolean features with mediocre feature importance. Obviously, a graph containing all the features would be more accurate, but we cannot report it in the chapter for space reasons.¹⁶ Still, our representation allows the reader to understand how features’ importance changes within and across the tiers. For instance, we can see that likes, and views are helpful to distinguish influencers between each pair of tiers, as also verified by Hypotheses 2 and 4. The video quality (e.g., size, definition) helps the classification especially in low tiers, since when the tier increases, the quality has already reached high standards. Contrary, profiles information helps more in distinguishing influencers in high tiers. Indeed, most Mega influencers have a verified profile with a bio link.

These results show that *traditional* features such as likes and views play a prominent role in distinguishing influencer tiers. Therefore, we asked ourselves the first research question (RQ1): “Do multimodal features (*audio*, *video*, *text*) make any difference when separating users in one tier from the next higher tier?”. To answer this question, we performed an ablation study. We divided our features into *traditional*, *video*, *audio*, and *text* sets, and performed the classification by removing, in turn, each possible combination of sets. Table 4.3 reports the mean increase in F1-score when that combination of sets is added on top of the remaining sets. Therefore, the row “video + audio + text” answers our RQ1. When these features are used, our classifier improves by 2.95%

¹⁶We have a total of 839 features. The complete graphs will be available on our repository when the corresponding paper is published (along with our code and data/links for hydration).

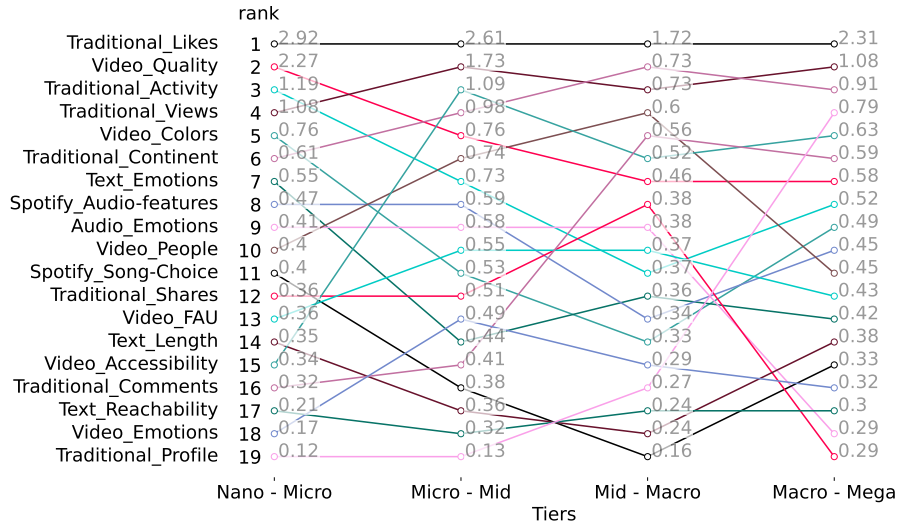


Figure 4.12: Feature importance per Tier - All features.

F1-Score for Nano-Micro problem, 1.6% for Micro-Mid, and 0.4% for Macro-Mega, but loses 0.1%F1-Score for Mid-Macro. Therefore, it is clear that *traditional* features, in particular likes, views, or the influencer activity, are the most useful during classification, while the others goes almost unnoticed, except for Nano-Micro.

Table 4.3: Impact, in percentage, of feature groups when considering all features - Increase of F-1 Score.

Feature Groups	Nano-Micro	Micro-Mid	Mid-Macro	Macro-Mega
Audio	0.44±1.75	0.75±2.12	-0.50±1.30	0.51±1.55
Audio + Traditional	10.43±2.74	18.75±4.08	22.12±5.33	25.63±3.47
Audio + Text	1.24±2.00	0.76±2.81	-1.02±2.47	0.42±0.90
Audio + Text + Traditional	12.60±4.07	20.33±4.72	21.48±4.33	24.25±4.54
Traditional	10.39±2.74	18.47±4.02	22.79±5.38	23.56±2.52
Text	0.74±2.44	0.15±1.83	-0.20±1.48	-0.14±1.93
Text + Traditional	11.49±3.40	19.46±4.17	22.84±5.92	25.12±3.28
Video	2.00±2.55	0.55±1.56	0.95±1.48	1.15±2.65
Video + Audio	2.84±2.42	1.20±2.84	1.02±1.71	1.01±1.29
Video + Audio + Traditional	20.79±4.51	26.90±3.41	26.34±2.40	27.15±3.65
Video + Audio + Text	2.95±3.03	1.60±3.44	-0.10±1.14	0.40±1.35
Video + Traditional	20.28±4.77	24.55±4.50	25.92±4.20	27.04±3.48
Video + Text	2.95±2.43	1.30±1.70	0.80±1.39	1.11±2.21
Video + Traditional + Text	26.95±5.19	27.54±5.03	25.88±3.51	30.92±4.61

Answer RQ1: *When considering all features, multimodal features (audio, video, text) only play a minor role in distinguishing users from one tier to the next one, with the highest benefit between Nano-Micro. Indeed, traditional features such as likes, views, and user activity have a leading role.*

4.6 ACTIONABLE FEATURES ANALYSIS

The previous section shows that *traditional* features are more critical than multimodal features when distinguishing between users in consecutive tiers. However, the user cannot directly act to modify these features — for instance, she cannot take actions that directly increase the number of likes or views of her videos. Table 4.1 has previously defined the set of *actionable* features, i.e., those the user can directly control to improve her status. We thus repeated our ablation study with only *actionable* features in order to understand which features the user might be able to modify in order to climb up to the next level.

Table 4.4 reports the results of the new ablation study. Differently from before, the multimodal features now bring improvements of 7.11% for Nano-Micro, 2.09% for Micro-Mid, 3.54% for Mid-Macro, and 1.14% for Macro-Mega. Interestingly, *video* features are now more critical than *traditional* features for Nano-Micro (5.56% vs 1.32%) and Mid-Macro (4.15% vs 2.79%), suggesting that in these tiers, influencers should focus more on the content of their videos. In the other tiers, *traditional* features remain the most impactful. *Audio* features impact around 1-2% in each category, while *text* features have the most negligible impact, often below 1%.

Answer RQ2: *When evaluating actionable multimodal features, video features emerge as the most influential, followed by audio features. text features, except for Macro-Mega, exhibit minimal relevance. Meanwhile, traditional features retain their significance across all tiers, although they take a backseat to video features in the Nano-Micro and Mid-Macro tiers.*

Table 4.4: Impact, in percentage, of feature groups when considering actionable features - Increase of F-1 Score.

Feature Groups	Nano-Micro	Micro-Mid	Mid-Macro	Macro-Mega
Audio	1.05 \pm 2.10	1.03 \pm 2.94	1.98 \pm 3.67	0.68 \pm 2.64
Audio + Traditional	1.36 \pm 3.28	3.70 \pm 3.56	2.12 \pm 1.86	8.51 \pm 2.84
Audio + Text	0.62 \pm 1.99	1.62 \pm 4.10	1.89 \pm 2.89	1.69 \pm 3.55
Audio + Text + Traditional	3.54 \pm 3.42	5.29 \pm 3.27	1.48 \pm 3.73	7.13 \pm 3.69
Traditional	1.32 \pm 3.59	3.43 \pm 2.83	2.79 \pm 3.94	6.66 \pm 3.06
Text	0.31 \pm 2.00	0.77 \pm 3.40	0.72 \pm 2.63	1.31 \pm 2.73
Text + Traditional	2.42 \pm 3.29	4.42 \pm 2.20	2.84 \pm 3.29	8.00 \pm 5.22
Video	5.56 \pm 2.96	1.76 \pm 3.70	4.15 \pm 4.66	0.52 \pm 2.02
Video + Audio	6.64 \pm 4.49	2.97 \pm 4.10	3.02 \pm 4.37	0.89 \pm 2.33
Video + Audio + Traditional	11.72 \pm 5.13	11.86 \pm 5.38	6.34 \pm 5.07	10.03 \pm 6.14
Video + Audio + Text	7.11 \pm 3.76	2.09 \pm 4.47	3.54 \pm 2.08	1.14 \pm 3.27
Video + Traditional	11.21 \pm 5.11	9.51 \pm 3.07	5.92 \pm 4.13	9.92 \pm 2.99
Video + Text	7.24 \pm 3.09	1.14 \pm 4.20	4.43 \pm 3.83	1.57 \pm 2.64
Video + Traditional + Text	17.88 \pm 5.61	12.50 \pm 5.17	5.88 \pm 3.35	13.80 \pm 3.78

4.6.1 WHICH FEATURES TO IMPROVE?

Until now, we have focused on high-level feature groups to quantify the significance of multimodal content types (*video*, *audio*, *text*). Yet, influencers are keen to gain insights into the specific adjustments required to ascend to the next tier, encompassing global strategies and granular details.

To support this, we first inspect the feature importance of the classifier using all the *actionable* features, similar to the preceding section. We then inspect the feature importance of the classifiers we used in the ablation study, which considered only a particular group of features (*traditional, audio, video, text*).

ALL ACTIONABLE FEATURES

Figure 4.13 shows the feature importance of the classifiers using all the *actionable* features. It immediately stands out that each classification considers different features. For example, emotions in the *text* are important for Nano-Micro classification, but not important in Micro-Mid, mildly important in Mid-Macro, and again important for Macro-Mega. This is explained by the fact that, on average, Nano influencers tend to use many positive emoticons, while Mega influencers use less. Being verified is the least important for Nano-Micro, but is of utmost importance for Macro-Mega. Indeed, only one profile is verified in the nano category, contrary to 171 and 474 in the Macro and Mega category, respectively. However, we can still see some common patterns. For example, the video quality and the profile features are in the top-5 groups for each category, while the Facial Action units are in the top-6. Similarly, the video emotions are always in the last positions, suggesting no tier has a predominant emotion.

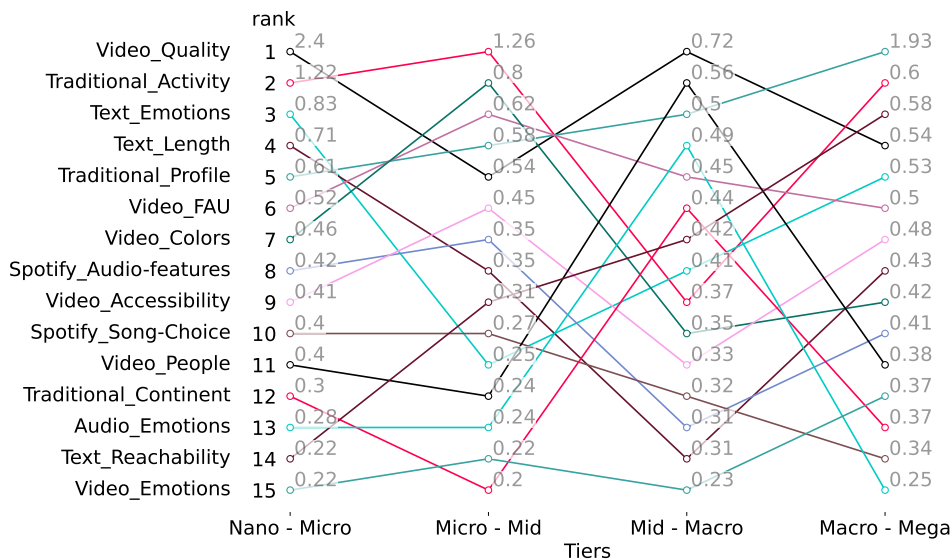


Figure 4.13: Feature importance - All actionable features.

Obviously, these features are important in different ways, and with different weights. For example, the video quality is two times more important than the activity for Nano-Micro, or the profile information is three times more important than the activity for Macro-Mega. To understand how algorithms consider these features, we should examine how influencers behave in their tiers with respect to these features. For instance, From Figure 4.10, we know that Nano influencers post more on the weekend, while macro-and mega post constantly throughout the week. Likely, the Nano-Micro and Macro-Mega algorithms consider this feature in the opposite direction. To get a more fine-grained understanding of the importance of features, we now explore each feature set.

TRADITIONAL FEATURES

Figure 4.14 shows the importance of *traditional* features. When we look at the top-5 *actionable* features across the four classification problems, we note that whether the user has a publicly available bio is in the top-5 in all 4 cases, while the total number of videos posted is in the top-5 in 3 of the 4 cases. Indeed, by increasing the tier, influencers tend to set their bio link more frequently (Figure 4.9) and have many more videos (Figure 4.5). When we look at the top-10 most important features, the time between posts and the number of videos posted daily also become significant. In fact, influencers should behave differently depending on their tiers. Figure 4.6 shows that the inter-posting time should decrease when the tier increases: the average inter-posting time is ten days for Nano influencers, but less than one day for Mega influencers. Interestingly, the significance of geographical continents varies across tiers, implying that influencers might also tailor their behavior based on this factor.

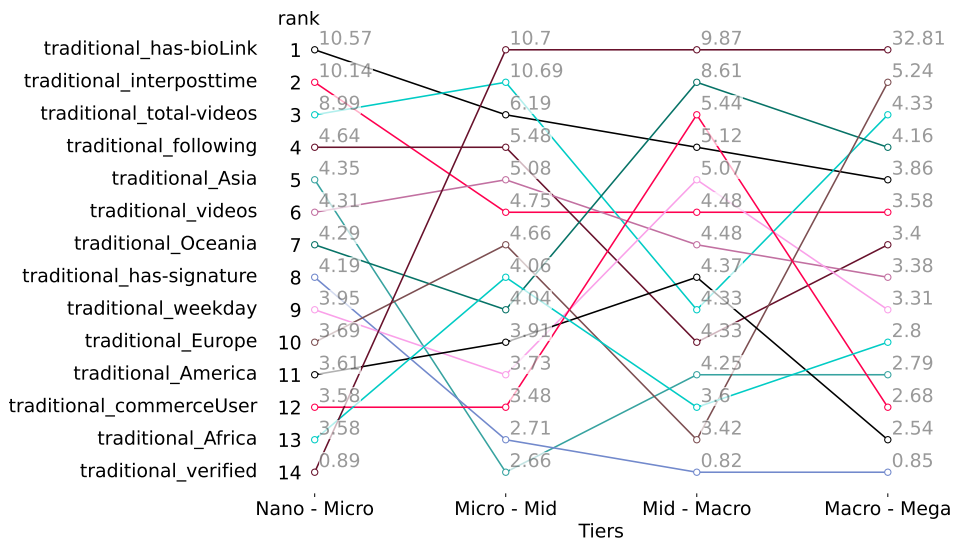


Figure 4.14: Feature importance - Traditional Actionable features.

AUDIO FEATURES

Figure 4.15 shows the importance of *audio* features. Even in this case, we see a mix of common and conflicting trends. For instance, Spotify loudness is crucial for Nano-Micro classification, but not for distinguishing between other tiers. In fact, Nano influencers use more Spotify tracks with a loudness value close to 0, contrary to the other tiers, which tend to avoid that on average. Similarly, Spotify acousticness is important only in Macro-Mega classification, where Mega influencers tend to use more acoustic tracks in their videos. On the other hand, Spotify popularity and audio valence stay in the top-10 across all four classification problems, but are important in different ways. Further inspecting popularity, we found that Nano and Macro tiers often use less popular songs, with Mega influencers avoiding them.

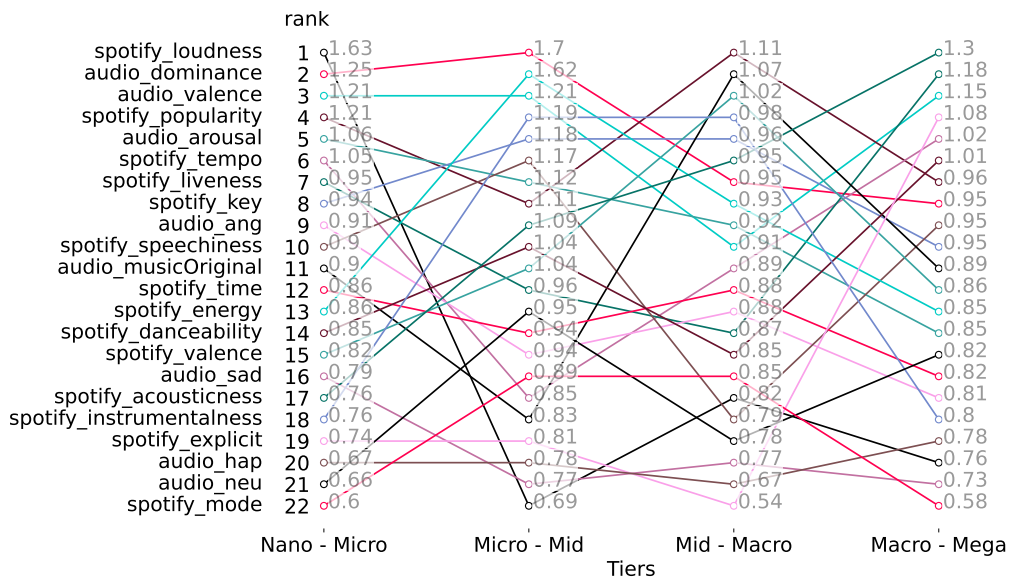


Figure 4.15: Feature importance - Audio Actionable features.

VIDEO FEATURES

Figure 4.16 shows the importance of *video* features, which are amongst the most essential features for Nano-Micro classification, and also essential across the other three classification scenarios. Further inspection reveals that lower-tier influencers shot their videos using many different sizes, while high-tier influencers tend to use more uniform sizes. This might indicate that most high-tier influencers own the same devices (e.g., top-notch smartphones with outstanding cameras). Facial Action Units are also very important in all four classification problems. For instance, Nano-Micro classifications show that Nano influencers exhibit more lip-related AUs, likely related to the higher presence of lip-syncing video in the Nano tier. The video pleasure is also significant, which increases and stabilizes in the Mega tier. Neutral and angry emotions appear to be less significant.

ACTIONABLE TEXT FEATURES

Figure 4.17 shows the importance of *text* features. Here, we witness more linear patterns. Indeed, as stated above, *text* features do not impact much in tier prediction. Still, the caption length is the most essential feature between Nano-Micro and Micro-Mid. Indeed, Nano captions are the shortest, Mid captions the longest, and Micro ones in the middle. The positive and neutral emotions are important in all classifications but Macro-Mega, in which the negative ones take over. In fact, Mega influencers tend to use more negative emotions in their captions. According to previous literature [116], hashtags and mentions are not very relevant.

Answer RQ3: *Influencers should primarily focus on traditional and video features. Important features vary greatly within each feature type depending on the classification problem being considered. How influencers should change their behavior is strictly related to their current tier and the next one up.*

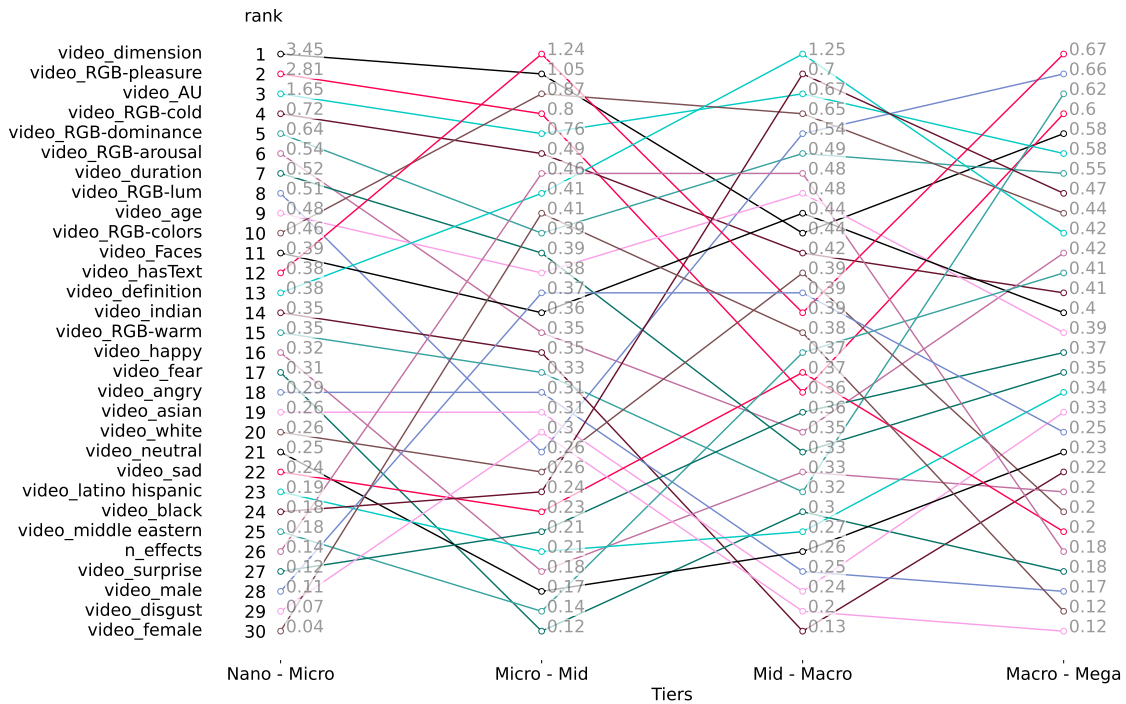


Figure 4.16: Feature importance - Video Actionable features.

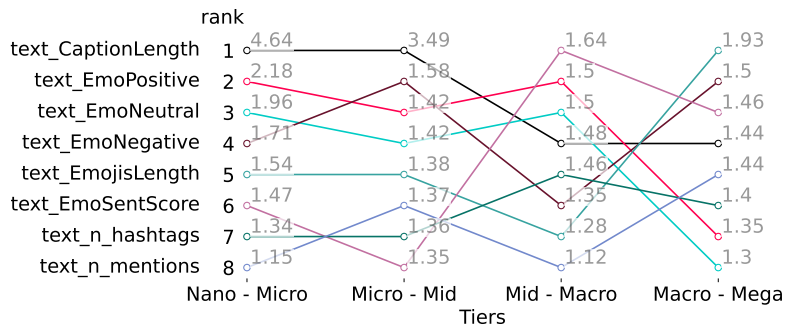


Figure 4.17: Feature importance - Text Actionable features.

4.7 CONCLUSION AND FUTURE WORKS

We investigated the key factors contributing to an influencer’s tier progression. Accordingly, we built powerful tier prediction classifiers and explored feature importance within the best model. While metrics such as likes and views may seem influential, they are beyond influencers’ control. Our experiments identified *actionable* features primarily within *traditional* features (e.g., posting frequency) and video-related attributes (e.g., video pleasure or facial expressions).

We have encountered certain limitations that offer opportunities for future research. Firstly, while we can

identify features that influencers should enhance, we currently lack precise guidance on how much improvement is needed to maximize tier progression probabilities. Secondly, our current approach analyzes features individually, even though combining them often yields superior results. In our future work, we aim to address these constraints, refine our feature sets, and explore trends on the platform, including how well influencers align with these trends.

ETHICAL CONSIDERATIONS

Similar to previous work [115], our study exclusively utilizes publicly available data and does not involve human subjects, thus exempting it from a formal review by our institution’s IRB. Nonetheless, our experiments are all performed by adhering to the guidelines of the Menlo report [141]. For instance, we report data only in aggregated form and do not pose risks to TikTok users, e.g., attracting unwanted attention. Upon request after publication, we will release our processed FAIR dataset, excluding the raw video data due to potential privacy settings or deletions. Instead, we will provide the URLs and code for rehydration. We collected (public) data using automated methods, which is discouraged by the ToS. However, as discussed in a recent ICWSM paper [142], “online data collection decisions should extend beyond ToS and consider contextual factors.” TikTok’s ToS permits manual collection, implying that automated collection is likely restricted to prevent server overload [142]. We took measures to ensure our data collection did not burden TikTok servers and received no warnings or bans from the platform.

5

Twitter Bots Influence on the Russo-Ukrainian War During the 2022 Italian General Elections

At the dawn of 24 February 2022, the president of the Russian Federation, Vladimir V. Putin, announced an imminent “Special Military Operation” in the oriental part of Ukraine. Soon thereafter, the global political leaders decided which side to support in the Russo-Ukrainian conflict. Along with most European countries, Italian politics sided with Ukraine by approving a law decree on 28 February 2022 [143]. The consequences of this decision were numerous. For instance, Italy reported a massive increase (+138%) of cyber-attacks directed at critical infrastructures, apparently caused by hackers lined up with Russia [144]. Additionally, Italian public opinion soon divided over the modalities of supporting Ukraine, such as sending military aid or applying sanctions to Russia. Since international relations inevitably impact democratic domestic politics [145], the role of Italy in the Russo-Ukrainian conflict was a major campaign issue for the Italian (snap) general election on 25 September 2022.

People and politicians started expressing their concerns and opinions regarding the Russo-Ukrainian war on social media platforms like Facebook [146], TikTok, Instagram, and Twitter. As largely demonstrated in the literature, opinions on social media are often manipulated by social bots [147, 148] or colluding activities [149]. Clear evidence has been found, for instance, in Japan’s 2014 general election [150] or USA presidential elections in 2016 [151] and 2020 [152]. Presumably, the last Italian general elections have not been exempted. Figure 5.1 illustrates a bot’s provocative tweet in response to Matteo Salvini, a leader of Italian politics. Therefore, studying the impact of bots is fundamental for understanding the potential consequences they may have on social dynamics and online interactions. By investigating the role of bots in shaping the community, we can gain valuable insights into how they may have influenced the dissemination of information and the formation of opinions.

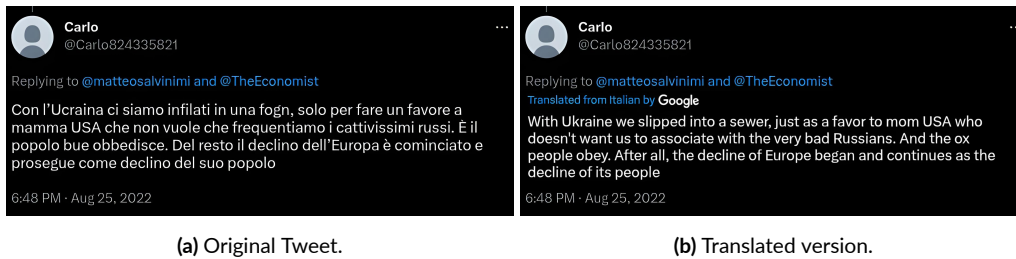


Figure 5.1: Bot response to an Italian politician expressing a strong-sided opinion regarding the conflict.

CONTRIBUTION. In this work, we investigate how Italian politics responded to the Russo-Ukrainian conflict on Twitter and whether bots manipulated public opinion before the 2022 general elections. In particular, we collected 39,611 tweets made by members of the main 6 political parties that belong to a left-wing or right-wing coalition from the period February-December 2022. We first conduct a semantic and temporal analysis of how politicians discussed the war, showing that some parties showed a high level of interest in the conflict and were actively engaged in commenting on the issue while others remained relatively silent. Secondly, we analyze 360,823 comments made during the last month of the political campaigns, from 23 August 2022 to 23 September 2022, examining bots' activities and influences on genuine users. We detected bots using Botometer [153], a popular tool capable of evaluating the realness of an account using a Machine Learning-based classification method. Our results show that around 12% of the profiles commenting on political posts are bots. Particularly, we found that bots have manipulated topics related to the Russo-Ukrainian war, especially on the center-right coalition, and that they influenced real users, often driving or soliciting discussions related to the conflict. We summarize our contributions as follows:

- We collected a dataset of 39,611 tweets posted between 24 February 2022 and 31 December 2022, from the six major parties in Italy, and 360,823 comments from 105,603 unique users who replied during the last month of the 2022 Italian general elections. The dataset will be made publicly available for future research;
- We provide a detailed analysis of how the 6 major Italian parties expressed and sided concerning the Russo-Ukrainian war on Twitter from the beginning of the war to the end of 2022;
- We examine the bots' impact on Twitter and how they influenced real users regarding the Russo-Ukrainian war during the last month of the general elections.

ORGANIZATION. Section 5.1 discusses related works, while Section 5.2 presents the dataset used in the experiments. In Section 5.3 and Section 5.4, we analyze politics in Italy during the conflict and the bots' influence on the elections, respectively. Section 5.5 makes further discussion and Section 5.6 concludes the chapter.

5.1 RELATED WORKS

In this section, we focus on the state-of-the-art analysis of bot infiltration in delicate scenarios and opinion manipulation through Twitter. Antonakaki et al. [154] conducted a comprehensive literature review presenting different approaches and techniques used for Twitter research. The authors acknowledged that Twitter had become a valuable data source for researchers, offering data for many purposes, such as forecasting social, economic, or commercial indicators [155] as well as assessing and predicting political polarization [156, 157]. For instance, Weber et al. [158], during the 2013 “Arab Spring” in Egypt, collected and analyzed a large dataset of tweets to categorize the users based on their political affiliation.

However, such information is often undermined by the presence of bots, i.e., automated accounts used to engage and behave mimicking human users, often controlled by a bot master. While there are some benevolent social media bots, many are used for dishonest and nefarious purposes [159, 160]. The existence of bots on the Twitter platform has been firmly established through many academic investigations [161, 162, 163, 164, 165], and news articles [166, 167]. Weng et al. [168] explained the differences between the opinion manipulations done by bots compared with those from real users, and Mazza et al. [169] investigated the difference between trolls, social bots, and humans on Twitter. Notably, these accounts can wield an exceptionally strong influence in delicate situations [170], such as stock trading [171], sensitive content diffusion [172], vaccination [173], or political elections manipulation. Regarding the latest, Pastor et al. [174] analyzed the presence and behavior of social bots on Twitter in the context of the November 2019 Spanish general election. They limited the analysis of the bots’ interaction up to seven days before Election day using Social Feed Manager [175] to capture the tweets and analyze the bot. Fernquist et al. [176] presented a study on the influence of bots in the Swedish general election held in September 2018. Bessi and Ferrara [177] investigated how the presence of social media bots impacted the 2016 Presidential elections in America, and similar works were conducted on the latest one in 2020 [152, 178]. For a comprehensive overview of bots, political elections, and social media, we refer to [179].

5.2 DATASET CREATION

In this study, we collected our own Twitter dataset due to the unique nature of the analysis. We selected six parties to analyze according to the current political scenario in Italy. In particular, we considered:

- The coalition that preceded Mario Draghi’s technical government (the so-called “giallo-rosso” government, who guided Italy from 5 September 2019 until 13 February 2021 [180, 181]), made by the Democratic Party (Partito Democratico, PD), the Five Stars Movement (Movimento 5 Stelle, M5S)
- The Italian Green-Left party (Sinistra Italiana-Verdi, SiVe);
- The coalition that won the September 2022 elections, and is currently in power: Brothers of Italy (Fratelli d’Italia, FdI), League for Salvini Premier (Lega per Salvini Premier, Lega), and Forward Italy (Forza Italia, FI).

We then model each of the parties to be constructed as:

$$D_i = [P, L, p_1, \dots, p_6]$$

where:

- D_i is the Dataset, $i = 1, \dots, 6$, one for each party.
- P is the “Party account”, e.g., @FratellidItalia.
- L is the “Leader account”, e.g., @GiorgiaMeloni.
- p_1, \dots, p_6 are six “major political figures” in that party, e.g., @DSantanche, @Ignazio_LaRussa, @FrancescoLollo1, @FidanzaCarlo, @fabiorampelli and @isabellarauti.

The final dataset has been constructed by collecting all the tweets from the party account, the leader account, and six other politicians in the party (following the structure defined above) that were posted from 24 February 2022 until 31 December 2022. To download the tweets, we queried the official Twitter API [182] to browse each profile’s timeline and retrieve all the necessary tweets. After this initial collection of tweets, we focused on the posts published during the latest month of the political campaign in Italy, from 23 August 2022 until 23 September 2022. We considered all the content shared by the secretary of each party and every reply. An overview of the full dataset can be seen in Table 5.1. We indicate the party, the party leader, the selected profiles we fetched the information from, the cumulative number of followers of each party’s profiles, and the overall number of posted tweets. For the last month of the political campaign, we considered all the content shared by the secretary of each party and every reply, as well as the number of unique commenters. These numbers represent only the tweets directly posted by the party members. During the collection, we excluded the retweets to reduce the number of repeated tweets between different accounts, to avoid redundancy, and to have a real and clear opinion from each profile.

Table 5.1: Complete overview of the dataset.

<i>Party</i>	<i>Leader</i>	<i>Members</i>	<i>Total Followers</i>	<i>Posted Tweets</i>	<i>Replies to Secretary</i>	<i>Unique Users Replying</i>
PD	Letta	Serracchiani, Orlando, Madia, Provenzano, Boldrini, Gentiloni.	3.511M	4357	158747	35571
FdI	Meloni	La Russa, Santanchè, Lollobrigida, Fidanza, Rampelli, Rauti.	2.471M	6610	60237	22670
M5S	Conte	Fico, Taverna, Appendino, Sibilia, Grillo, Maiorino.	2.419M	3672	47886	14255
Lega	Salvini	Fontana, Arrigoni, Pillon, Rixi, Centinaio, Bongiorno.	1.898M	15797	59317	20159
FI	Berlusconi	Tajani, Bernini, Gasparri, Fitto, Casellati, Ronzulli.	804.2K	4172	29597	9962
SiVe	Fratoianni	Bonelli, Soumahoro, Alemanni, Evi, Marcon, Pellegrino.	411K	5003	5038	2986

5.3 THE RUSSO-UKRAINIAN WAR IN ITALIAN POLITICS

We start our analysis by understanding whether and how frequently the Italian parties mentioned the Russo-Ukrainian conflict (Section 5.3.1). After that, we conduct a temporal analysis to determine when the conflict was primarily discussed, with a particular focus on election time (Section 5.3.2).

5.3.1 THE IMPORTANCE OF CONFLICT FOR ITALIAN POLITICAL PARTIES

Our objective in this section is to answer the question, “How did Italian politicians discuss the war?”. After the creation of the datasets D_1, \dots, D_6 , we cleaned each tweet by (i) removing emojis with the tool `clean-text` [183], (ii) removing the links, and (iii) removing stop words [184]. Figure 5.2 shows the Word Clouds for each party.¹



Figure 5.2: Word Clouds for the tweets of parties captured.

The first row contains the Word Clouds associated with the parties belonging to the center-left coalition: PD focuses mostly on “lavoro” (“job”), “destra” (“right-wing”), and “Ucraina” (“Ukraine”); M5S concentrates on their own public appearance, with words like “TV” and “intervista” (“interview”), and its leader “Giuseppe Conte”. Finally, SiVe emphasizes their new coalition with the words “AlleanzaVerdiSinistra” (“Green Party-Italian Left Coalition”) and “europaverde” (“Green Europe”).

On the other hand, the second row is made by the parties belonging to the center-right coalition: FdI, similarly to PD, concentrates on their opposing wing with words like “sinistra” (“left-wing”) and “governo” (“government”); Lega is vastly influenced by its leader “Matteo Salvini” and his public appearances, indicated by words like

¹We computed the word clouds using WordCloud Python Library [185]

“TV” and “Radio”. FI rotates around its leader too, as the most commonly used words are “Presidente” (“President”) and “Berlusconi”. Since the word clouds only provide a high-level view of the most commonly used words, we refine our analysis by inspecting the topics addressed by the parties. Indeed, political parties usually shape their campaigns by supporting or emphasizing particular themes. Thus, we extracted the topic they mainly discussed, and analyzed whether the Russo-Ukrainian war played a prominent role. To extract the topics, we started by calculating the embeddings of our tweets using the pre-trained multilingual Sentence-Bert model [186] supporting Italian language². The corresponding tweets’ embeddings (i.e., vectors of 768 dimensions) were more similar when their content was semantically closer. By leveraging this feature, we could cluster the data to find topics. First, we used UMAP algorithm [187] to decrease the vectors dimension to 5, setting `n_neighbors=15`. Then, we applied the density-based HDBSCAN clustering algorithm [188] to define clusters of at least 15 points, using the Excess of Mass selection method and Euclidean distance as the similarity metric. Once the clusters were defined (i.e., collections of semantically similar tweets), we extracted their most important words to manually label the corresponding topic. We calculated words’ importance by using class-based TF-IDF [189]. In this version of the algorithm, each document corresponds to a topic (or class), i.e., the aggregation of all the tweets belonging to that topic. We can then identify the most representative words of a topic by selecting its most frequent words that are less frequent in the other topics. Table 5.2 shows the most discussed topics for each party, along with the percentage of tweets posted about them. For conciseness, we report only the top-7 topics for each party.

Table 5.2: Top-7 topics and the number of tweets for each party.

<i>PD</i>		<i>M5S</i>		<i>SiVe</i>		<i>FdI</i>		<i>Lega</i>		<i>FI</i>	
%	<i>Topic</i>	%	<i>Topic</i>	%	<i>Topic</i>	%	<i>Topic</i>	%	<i>Topic</i>	%	<i>Topic</i>
24.25	RU-UA War	16.42	Italy	80.10	Vote Left	24.57	Italy	26.90	Italy	88.97	Berlusconi
14.96	Salary	14.54	Energy	12.34	Do	16.28	Vote	17.45	Energy	5.10	RU-UA War
10.87	Truth	11.12	RU-UA War	3.04	RU-UA War	12.82	Meloni	10.84	RU-UA War	1.23	Agenda
10.16	Italy	10.35	Mafia	1.81	Education	10.27	Do	8.71	Immigrants	1.06	Pandemic
8.74	Europe	9.15	Salary	0.64	Military Exp.	8.49	RU-UA War	7.15	Taxes	0.90	Italy
7.48	Vote	8.81	Agenda	0.48	Iran Women	8.34	Taxes	6.53	Rome	0.85	Foreign wars
6.85	Fascism	7.96	Courage	0.48	Climate	6.05	Energy	6.24	Vote	0.59	Europe

It immediately stands out that the Russo-Ukrainian conflict was a prominent topic for each party. Particularly, the topic placed in the first three positions for five out of six parties. PD mentioned the conflict the most, while FDI was the least. By inspecting the most important words for the topic, we find the words “sanctions” to appear frequently for PD, M5S, Lega, and FI, “weapons” for PD and SiVe, and “solidarity” for M5S and FDI. In any case, this topic appears to have a similar impact on other “internal” matters like taxes, migrants, or energy. Only SiVe and FI show a heavily unbalanced topic frequency. In both of these cases, however, the war played a prominent role. To conclude, all major Italian parties discussed and included the war in their campaigning.

5.3.2 TEMPORAL ANALYSIS OF RUSSO-UKRAINIAN DISCUSSIONS

We noted that each party included the Russo-Ukrainian war in their political campaigns. However, it is important to understand when the parties discussed it the most. We could expect, for instance, high frequencies at the

²We used the model `distilbert-multilingual-nli-stsb-quora-ranking`.

beginning of the war or near the elections. In such a sense, a temporal analysis can help us understand which parties concentrated their whole campaigns on the war or only referred to it in crucial moments to express solidarity. To this aim, we created stack plots to inspect the temporal references to “Ukraine” and “Russia” during the year. Specifically, we computed the frequency of tweets related to Ukraine and Russia using a bag of words approach, i.e., by counting the number of occurrences of Ukraine/Russia-related words, such as “Ukrainian”, “Zelensky” or “Russian”, “Putin”. The results are presented in Figure 5.3. For clarity, we also reported four major events during the conflict, such as the three main phases described in [190] and [191].

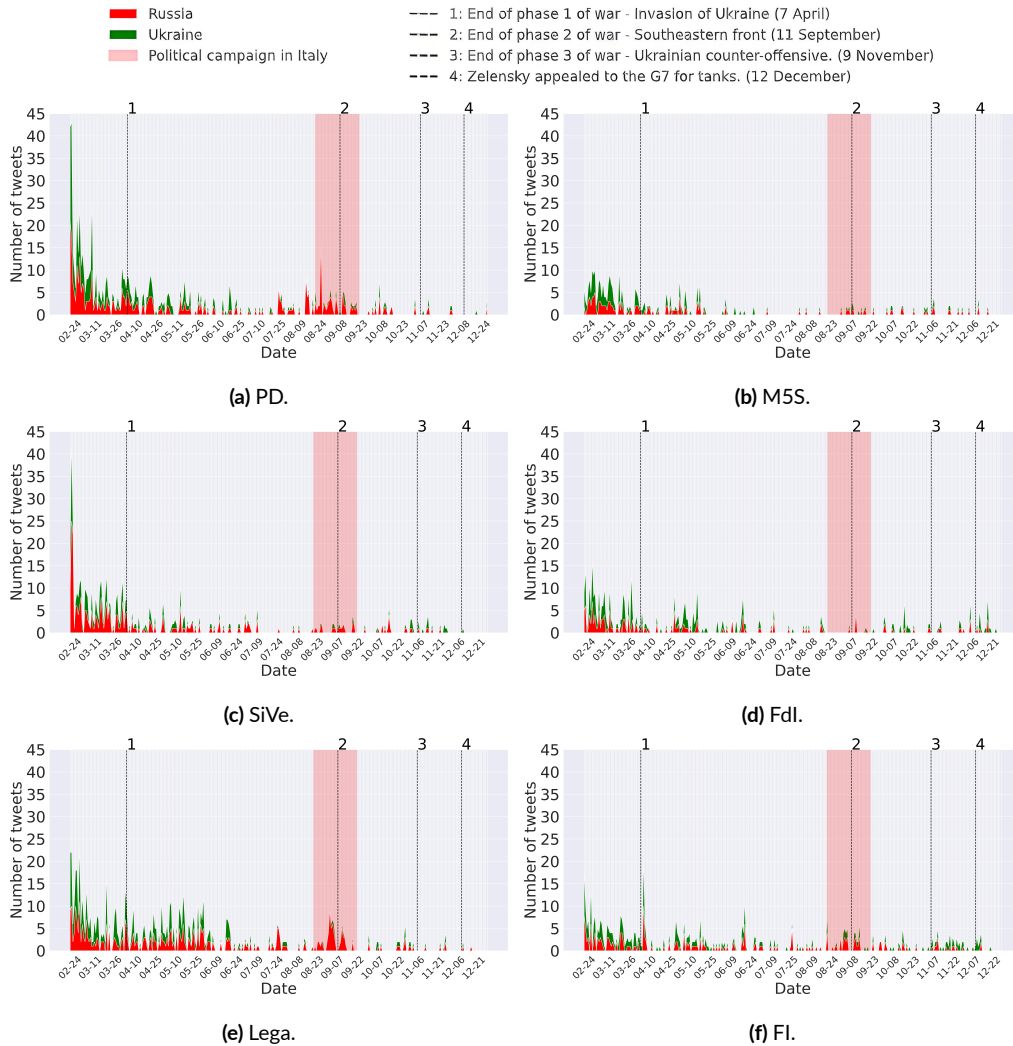


Figure 5.3: Temporal trends for the war-related tweets, 15 days aggregation.

All parties discussed the Russo-Ukrainian war mostly between the beginning and end of phase 1. Particularly, PD shows the most active involvement, which is in accordance with Table 5.2, while FI displays the highest number of tweets at the end of phase 1. Over the year, all parties gradually decreased their discussion of the

topic, except for PD, Lega, and FI, which devoted a significant portion of their campaign propaganda. Interestingly, while Russia and Ukraine-related words were balanced initially, these parties focused most on Russia-related words during the campaign, showing a condemnation attitude rather than solidarity, as confirmed by manual inspection. The remaining parties did not accentuate the topic during the campaign, except near the end of phase 2.

Following the election, which saw the center-right coalition led by FdI winning, there was a noticeable decline in the number of tweets related to the war from most political parties. In contrast, FdI and FI continued to post about the war, sometimes with increasing activity during phase 3 and phase 4. In these cases, the focus seems to have switched to Ukraine rather than Russia, probably reflecting the evolution of the conflict. These considerations suggest that while the Russian-Ukrainian war may no longer be a trending topic among most political parties, it remained quite an important issue for FdI and FI, who continue supporting Ukraine in their political messages [192].

5.4 BOTS INFLUENCE ANALYSIS

In the previous section, we highlighted that the Russo-Ukrainian conflict played a major role during the 2022 Italian General Elections. We now explore how many bots participated in the political discussions (Section 5.4.1), whether bots manipulated or distorted the discussions of the Russo-Ukrainian conflict, (Section 5.4.2), and whether they influenced real users or simply followed the flow of the conversation (Section 5.4.3).

5.4.1 BOTS PRESENCE ANALYSIS

To evaluate the bots' influence on elections, we retrieved all replies under the posts of each party's secretary during the last month of elections, between 23 August and 23 September 2022. To detect bots among the commenters, similar to previous works on Italian tweets [193, 194], we employed Botometer [153], a widespread ML-based tool [195, 196] that distinguishes between legitimate users and bots. Among the metrics, Botometer returns, for each checked account, the following scores:

- `overall raw score`: score in $[0, 1]$ determining whether an account is a bot;
- `cap`: (Complete Automation) Probability in $[0, 1]$ that an account with that score or greater is a bot. In other words, it expresses the prediction's confidence.

A classic approach to classify a bot takes the `overall raw score` and compares it to a fixed threshold (e.g., > 0.50 classified as a bot, ≤ 0.50 classified as human). Instead, for each user, we labeled as bot those with `overall raw score` $> \text{cap}$, with `cap` > 0.80 . By doing so, we adopted a dynamic and more accurate threshold than the classic approach, reducing the number of false positives. This method was confirmed by parsing several accounts manually, and among them, users with a high CAP (i.e., above 0.80) value were always classified as bots. Table 5.3 reports the number of unique accounts labeled as bots that replied under the party's secretary. On average, we found $\sim 12\%$ of bots replying to each secretary, with Meloni showing the higher percentage of bots (15.08%) and Fratoianni the lowest (9.61%).

Table 5.3: Percentages of bots and non-bots for each profile.

Profile	Unique Users	Bots (%)	Non-bots (%)
Letta	35,571	10.76	89.24
Conte	14,255	12.20	87.80
Fratoianni	2,986	9.61	90.39
Meloni	22,670	15.08	84.92
Salvini	20,159	11.12	88.88
Berlusconi	9,962	12.92	87.08

We further investigate the categories of bots interacting with Twitter profiles, according to Botometer classification. In particular, bots fall into the following categories:

- *Financial*: bots that post using cashtags;
- *Fake-follower*: bots purchased to increase follower counts;
- *Spammer*: accounts labeled as spambots from several datasets;
- *Self-declared*: known bots listed on botwiki.org;
- *Astroturf*: accounts that primarily focus on influencing public opinion, often being part of a network;
- *Other*: miscellaneous bots.

Given that Botometer’s response includes a percentage indicating the likelihood of an account belonging to each category, a bot was assigned to the category with the greatest likelihood. The final cumulative results for each politician are presented in Table 5.4.

Table 5.4: Categories of bots distribution replying to the tweets of the leaders.

Profile	Number of Bots	Financial (%)	Fake-followers (%)	Spammers (%)	Self-declared (%)	Astroturf (%)	Other (%)
Letta	3828	0.06	25.33	0.15	33.07	35.83	5.56
Conte	1739	0.08	33.87	0.08	31.27	32.04	2.67
Fratoianni	287	0.00	19.44	0.00	42.78	31.67	6.11
Meloni	3418	0.04	30.03	0.15	33.53	31.50	4.75
Salvini	2242	0.06	39.35	0.11	27.97	27.69	4.83
Berlusconi	1287	0.44	26.40	0.00	31.79	34.43	6.93

A significant proportion of counterfeit profiles engaged with political figures fall under the categories of “fake followers” and “astroturf”. This result confirms that most analyzed bots aim to influence or manipulate public opinion. Another notable percentage pertains to “self-declared” bots that, on the other hand, operate on the platform without any nefarious motives. In general, the bots distribution is consistent across all profiles.

We further investigate whether bots cooperate within the two coalitions we described in Section 5.2, namely, the Center-Right coalition (Berlusconi, Meloni, and Salvini) and the Center-Left coalition (Letta, Conte, and Fratoianni). Figure 5.4 shows the shared number of bots in the two coalitions. For the Center-Left coalition, many accounts identified as bots and commenting on multiple politicians are associated with Letta and Conte, the primary figures in the “giallo-rosso government” mentioned earlier. Additionally, the remaining shared bots are linked to Fratoianni and, once again, Letta, the leaders of the two largest parties comprising the Center-Left

coalition in the most recent elections. On the other hand, in the Center-Right coalition, there is a significantly stronger affiliation between the three profiles, as confirmed by the interrelation between the three political parties. Several bot accounts are common to two profiles, with a select few being shared by all three, suggesting a much closer connection between the coalition’s parties and their ideologies.

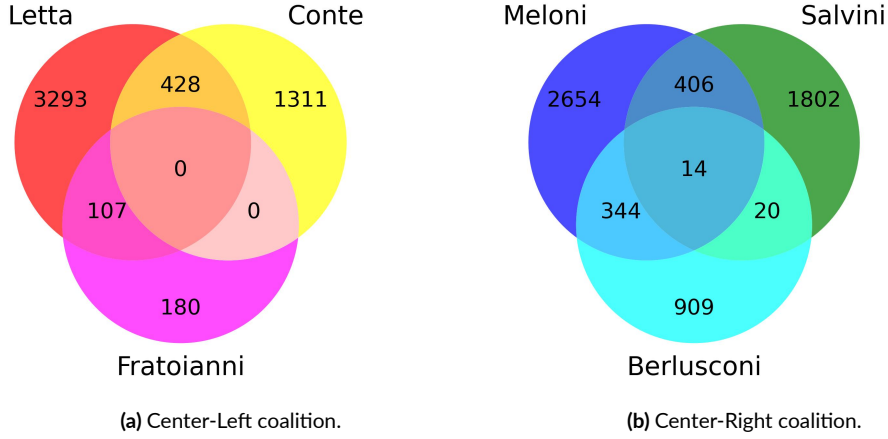


Figure 5.4: Number of shared bots between profiles belonging to the same coalition. Colors are representative of the parties, according to the Italian press.

5.4.2 BOTS TOPICS DISTORTION ANALYSIS

We now investigate the lexical associations between the words employed by authentic and bot users during the last month of the Italian General Elections’ political campaign. In this way, we can explore and understand how bots and humans communicated regarding the Russo-Ukrainian conflict, and whether bots distorted the vision of war-related topics. Inspired by the methodology introduced in Sartori et al. [197] and Tahmasbi et al. [198], we aim to discover associations between war-related words, e.g., how frequently they appear together in a tweet. For this purpose, we first trained a Word2Vec model [199] on our tweets to determine how words related to the Russian-Ukrainian war relate to each other. In this model, words with similar vectors are likelier to appear together in a tweet. Starting from the words “Russia”, “Ukraine”, and “War”, we manually identified 10 frequent related words, selecting (i) institutional-related words, i.e., “USA”, “EU”, “NATO”, “Europe”, and “Italy”; (ii) war-related words, i.e., “weapons”, “conflict”, “invasion”, “aggression”; (iii) “gas”, as its price rose sharply due to the conflict. Subsequently, we calculated the incidence matrix $M \in \mathbb{R}^{3 \times 10}$ for each involved party, utilizing the trained Word2Vec model. The incidence matrix M can be mathematically formulated as in the Matrix 5.1.

$$M = \begin{pmatrix} m_{1,1} & m_{1,2} & \dots & m_{1,9} & m_{1,10} \\ m_{2,1} & m_{2,2} & \dots & m_{2,9} & m_{2,10} \\ m_{3,1} & m_{3,2} & \dots & m_{3,9} & m_{3,10} \end{pmatrix} \quad (5.1)$$

where $m_{ij} = \text{cosine_similarity}(v_i, w_j), i = 1, 2, 3$ and $j = 1, \dots, 10$ ³. The words v_i are the selected words {"Russia", "Ukraine", "War"}, while the words w_j are the selected words {"USA", "EU", "NATO", "Europe", "Italy", "weapons", "conflict", "invasion", "aggression", "gas"}. If the cosine similarity was negative, we truncated it to 0. This matrix M was computed for each party in two different scenarios:

- A *Complete* scenario, considering both replies from real and bot accounts;
- A *No Bots* scenario, considering only replies from real users.

We fed these matrices to the Gephi Software [201] to construct weighted undirected graphs, which we call "Spider Graphs" due to their shape, and we used Force-Atlas 2 [202] as Layout for the rendering. In our graphs, the nodes are the words, and the edges represent the cosine similarity. According to the incidence matrix, edges exist only between the three initial words ("Russia", "Ukraine", "War") and the 10 selected words. The node size reflects its degree (larger words have more connections), while the thickness of the edges reflects the similarity of the connected words (thicker edges connect more similar – or likely to appear together – words). Last, we applied the modularity algorithm [203] to build clusters of strictly connected words.

We set the resolution parameter of the modularity algorithm to obtain three clusters: a red cluster with centroid "Russia", a green cluster with centroid "Ukraine", and a blue cluster with centroid "War". The remaining 10 words are then placed by the algorithm in the closest cluster, acquiring its color. Edges have the color of the cluster if they are connected to their centroid, or a mixed color if they are connected to the centroid of a different cluster. For instance, the edge between a word of the "War" (blue) cluster and "Russia" (red) will be purple (blue + red). Figure 5.5 and Figure 5.6 present the spider graphs in the *Complete* and *No Bots* scenario of the Center-Left and Center-Right coalitions, respectively.

For the Center-Left coalition, the most significant change between the two scenarios concerns the M5S party. While the lexical similarity between "War" and "conflict" remains the same, there are no other words in the "War" cluster when considering the *No Bots* scenario. An important constant between the two M5S graphs is the strong link between the central node "Russia" and "Italy". The graphs of PD and SiVe seem to show several differences. In the graph 5.5b the word "gas" disappears and the cluster of "War" gains the word "Italy" from the "Ukraine" cluster. The word "weapons" is always clustered with "Russia" in all the graphs of PD and SiVe and the word "Italy" is always with Russia in M5S and SiVe. Moreover, the word "conflict" is present only in the graphs of PD and M5S and it is absent from the ones of SiVe. The presence in the three clusters of institutional-related words, i.e. "NATO", "USA", "EU" and "Europe", seems not to have such relevant lexical importance for the parties except for "USA" and "War" in the PD scenarios.

For the Center-Right coalition, the primary observation concerns the intensified association between the central term "Russia" and words that frequently pertain to institutions, such as "Italy", "NATO", and "Europe". Another identifiable characteristic noted by the model is the substantial presence of words within the cluster associated with the term "War", whereby the most frequent ones include "gas", "invasion", and "conflict". Within this coalition, it appears that every "Russia" cluster encompasses a closely related term, such as "Italy" or "USA", with a strong connection. The strongest differences between the two scenarios appear around the "Ukraine" cluster. Indeed, for all three parties, the words within the cluster differ significantly between the *Complete* and *No Bots*

³The cosine-similarity was computed according to the formula in [200]

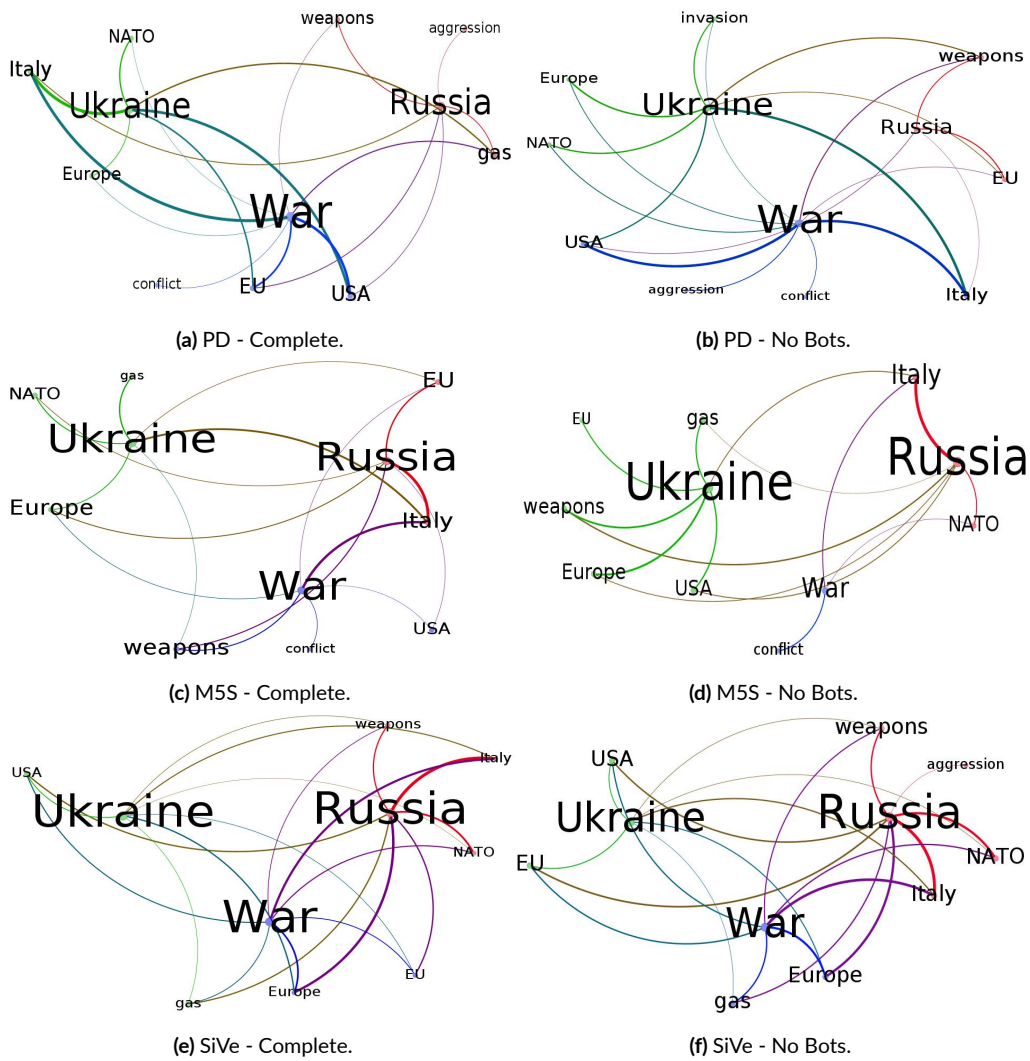


Figure 5.5: Comparison between “Spider Graphs” of the Mixed and No-Bots Scenario in the Center-Left coalition.

scenarios. For instance, for FdI, the “Ukraine” cluster goes from “aggression” and “invasion” in the *No Bots* scenario to “NATO”, “EU”, and “weapons” in the *Complete* scenario. Significant differences between the scenarios also appear around the “Russia” cluster. Therefore, we notice how bots significantly impacted public opinion by going in the opposite direction of real users. Considering all the graphs, we can assert that the existence of bots appears to influence the outcomes of the clustering analysis, especially for the Center-Right coalition.

5.4.3 BOTS TEMPORAL INFLUENCE ANALYSIS

To conclude our analysis, we deeply investigated the final month of the Italian elections, exploring the different discussions and perspectives surrounding the war that emerged under the leaders’ posts. Our goal is to understand

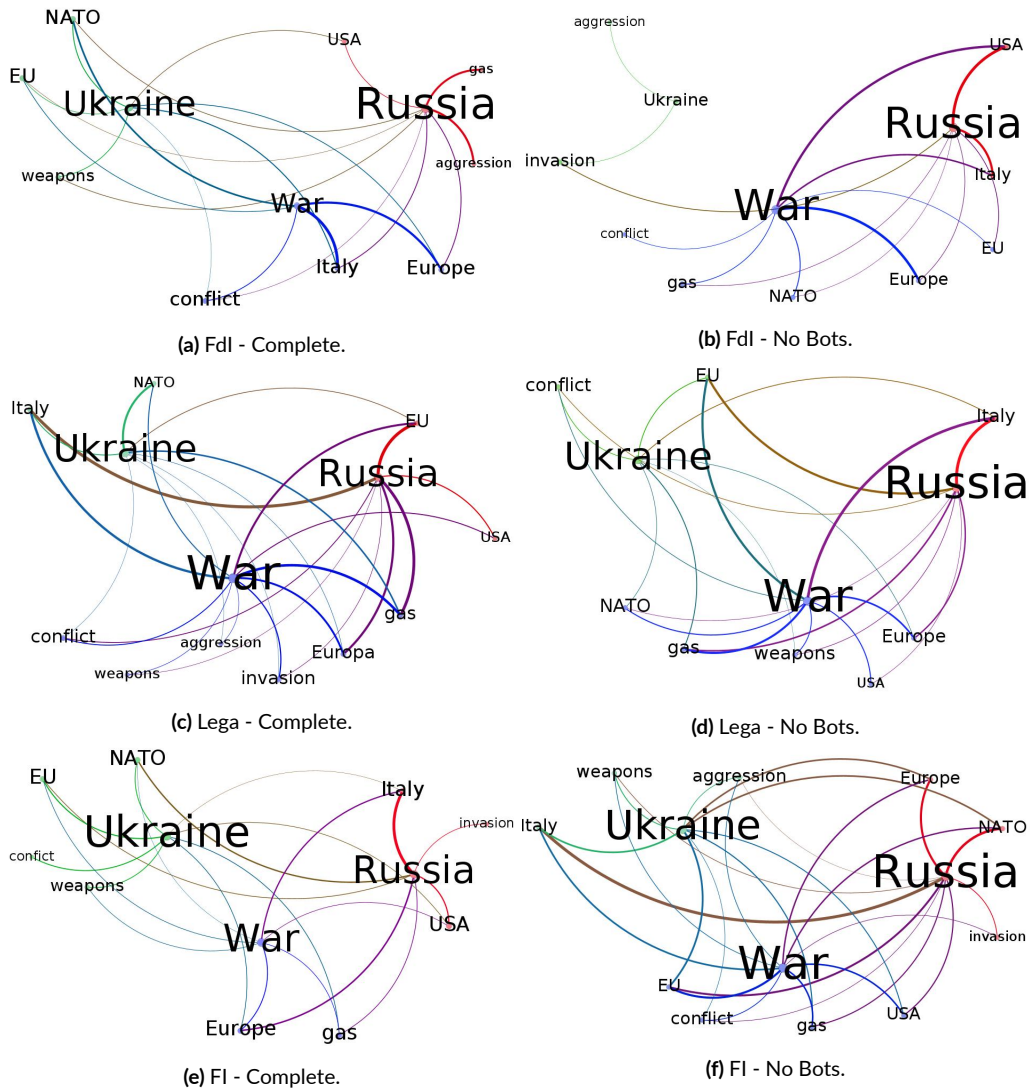


Figure 5.6: Comparison between “Spider Graphs” of the Complete and No Bots Scenario in the Center-Right coalition.

whether humans or bots discussed more the conflict, and which side influenced (or started) the debate. To this aim, we plot a two-scale graph for each party, considering the mean number of tweets concerning the war and the mean posting time (hour) for bots and real users. The results are shown in Figure 5.7.

We computed the harmonic_mean with the Formula 5.2:

$$b_{freq} = 2 \times \frac{\text{Ukraine_frequency} \times \text{Russia_frequency}}{\text{Ukraine_frequency} + \text{Russia_frequency}} \quad (5.2)$$

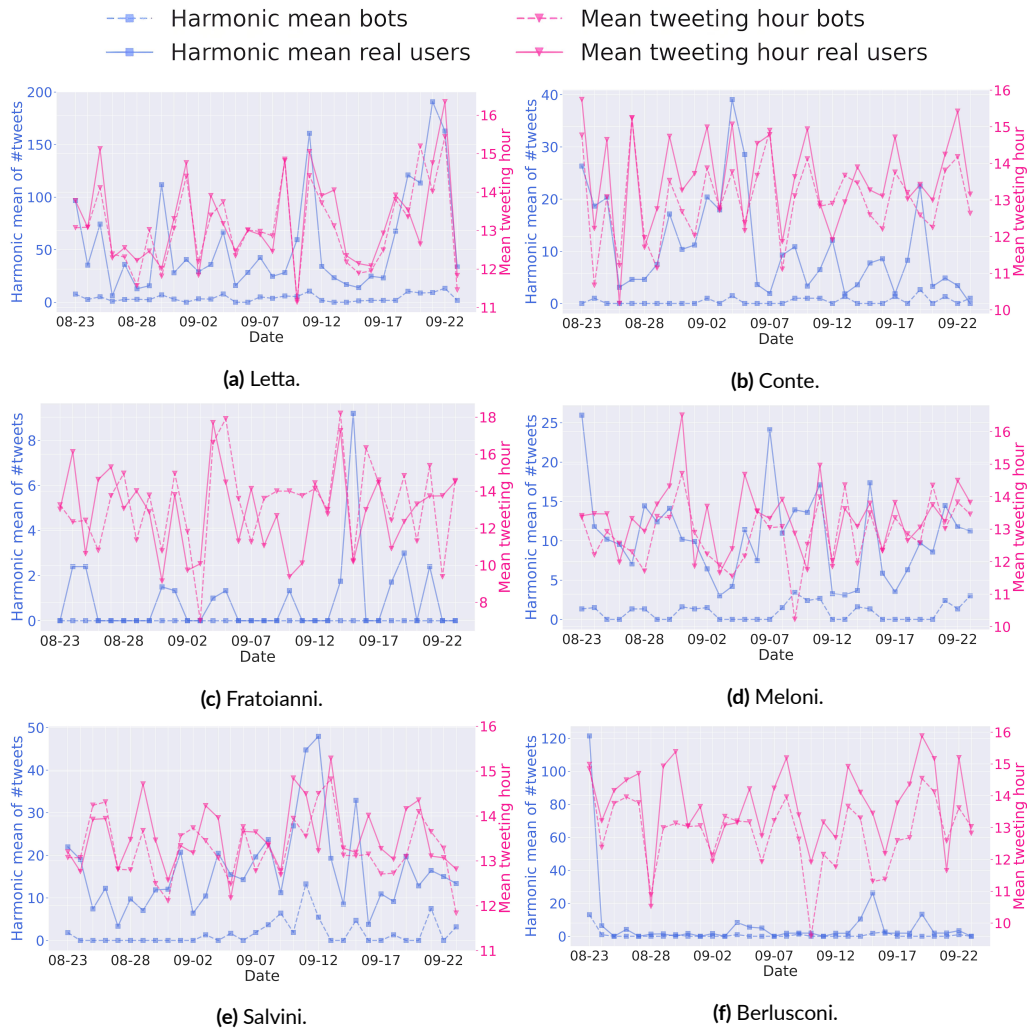


Figure 5.7: Mean number of posts and mean posting time for war-related tweets in the last month of Italian elections. Data are reported for both real accounts and bots.

as an indicator to visualize the number of tweets posted daily by both real users and bots during the last month of the political campaign. Ukraine and Russia posts included only strictly related words to the countries, e.g., “Ukraine”, “Ukrainian”, “Zelensky”, and, “Russia”, “Russian”, “Putin”.

This measure is bounded from above by the arithmetic mean, indicating its tendency to mitigate the influence of large outliers while accentuating the effect of small ones. This property allows for the evaluation of even the smallest frequencies to be computed, which may be otherwise masked by the influence of dominant outliers in the data. In this scenario, e.g., the results for a politician like Fratoianni, which has a smaller frequency of bots if compared to the other figures, are not suppressed, but his mean will clamp to 0. The other indicator we considered is the `mean_tweeting_hour`, which gives us the arithmetical average of posting time by both genuine and bot accounts.

We focus our attention on the blue spikes in the graphs, which indicate a quantitative increment in the number of tweets regarding the war. The majority of the spikes, either regarding the real or the fake users, concentrate on the period between 10 September and 24 September. The number of tweets posted by real users is always greater than bots' posts, which is in accordance with the percentage of bots found earlier ($\sim 12\%$). Looking at the `mean_tweeting_hour`, we can establish that on various occasions the bots posted tweets in a time before the spikes coming from the real users, on average. This trend is glaring for Conte, Meloni, Salvini, and Berlusconi, in which bots often started tweeting before the real users, hence influencing or driving the daily discussion.

5.5 DISCUSSION

Our analyses found that Italian politics has actively considered the Russo-Ukrainian conflict in their campaigns, with parties taking on a greater role than others. Additionally, we found a fair number of bots to be active and influential during the last elections. The effect seems to be tied to the particular parties or coalitions, requiring further investigation. Indeed, we could not determine nor speculate on who was driving these bots or for what purpose. Anyhow, our findings demonstrate that external events can significantly impact local (national) ones, with unpredictable consequences. Social media platforms like Twitter are credited with democratizing discussions about politics and social issues, but as demonstrated in the literature, manipulation of information is an actual threat rather than a risk. Unfortunately, most studies addressing this issue focus on English-based data or countries, since state-of-the-art models are more reliable. However, analyzing non-English countries is of utmost interest nowadays, since every country has a significant impact on global political equilibrium.

As we found interferences in the political scenarios, bots or fake accounts might likely be involved in disinformation or other malicious activities in the country. With the rapid development of Artificial Intelligence, it could always become harder to detect these colluding entities. It is, therefore, necessary to conduct further studies to address the language-specific obstacles, as well as to identify who operates such bots to eventually detect their objectives and contrast them.

5.5.1 LIMITATIONS

As we mentioned earlier, our study was limited by the few models available to process the Italian language. However, we think our work can stimulate further research and improve NLP models for Italian, as well as other minor languages. An additional limitation relies on the use of the external tool Botometer for the detection of bots. As such, the reliability of our findings is contingent on the accuracy of this tool [204]. However, Botometer is widely recognized as a state-of-the-art bots detection mechanism, and we have taken a conservative approach in the detection phase to limit false positives. Indeed, the number of bots and their influence could be higher than our estimates, stressing the need for more research in the area.

5.6 CONCLUSION AND FUTURE WORKS

The purpose of this study was to investigate how Italian politics responded to the Russian-Ukrainian conflict on Twitter and understand the bots' influence and manipulations before the 2022 general elections in Italy. Our findings suggest that bots are a significant presence in political conversations on Twitter, with approximately 12% of commenters being identified as bots. We also analyzed the timing in which the bots posted concerning when the real users posted, and we can infer that in some cases, these accounts could have forced a certain direction in the topics discussed online. This highlights the potential impact of automated accounts on public opinion during political campaigns.

Our analysis can be improved in the future in several ways. For instance, we could consider the presence of comments in other languages. As our study focused solely on comments posted in the Italian language, taking into account comments in other idioms could offer a more comprehensive understanding of the discussion. In addition, users' attitudes and behaviors could be studied based on their location, in order to analyze potential regional differences in the discussion. Notably, identifying the geographical location of bots can provide more insight into *who* attempts to manipulate discussion and *why*.

Part II

Developing Resilient Social Network Analytics Tools

INTRODUCTION TO PART II

The last chapter of the previous part underscored the significant impact that social bots and adversarial activities can have on the efficacy of social network analytics. Moving forward, the second part of this dissertation takes a proactive stance by concentrating on developing resilient Social Network Analytics tools, designed to operate effectively even in adversarial contexts. Specifically, it delves into three critical dimensions: firstly, the detection of Instagram crowdturfing, where deceptive tactics artificially boost social media metrics; secondly, the identification of evasive samples disseminated by OSN users, which pose challenges to automated analysis and hinder SNA progress; and lastly, it introduces an innovative application of social honeypots for the in-depth exploration of communities and topics within OSNs, providing valuable insights into user interactions and emerging trends. This multifaceted effort is set to advance OSN research, formulating strategies that enhance the resilience, integrity, and security of online social networks.

6

Are We All in a Truman Show? Spotting Instagram Crowdturfing through Self-Training

Instagram (IG) is the most popular photo-sharing social media, with around 1.5 billion monthly active users [205], and the preferred platform for influencer marketing [206]. Unfortunately, such a market is often manipulated, making influencers unreliable [207]. Indeed, many providers offer services to boost the visibility and fame of a specific account, for example, by increasing the number of followers, likes, and comments. As (social) bots [165] or fake accounts [208] originally conducted these activities, IG has adopted Machine Learning (ML) algorithms to remove them efficiently. Instead, nowadays, *real people use their accounts to conduct such unauthentic activities behind a monetary reward*. In the literature, this collusive phenomenon is called crowdturfing (CT), a term combining the collaboration of many individuals (crowdsourcing [209]) with an apparently natural action controlled by agencies (astroturfing [210, 211]). Figure 6.1 shows Fake and CT profiles. While the fake profile exhibits well-known characteristics (e.g., no posts, no bio, few followers [208]), the CT profile looks legit (indeed, it is a real-person account), and thus more challenging to spot. Considering CT engagement is not real, we can label it as fake. Fake engagement damages the authenticity of social media, creating threats such as brand abuse or followers farming [50].

To spot fake engagement derived from CT, a reliable strategy is to detect the involved accounts. Different approaches have been proposed to distinguish between genuine or fake users, but to the best of our knowledge, none consider CT-involved accounts on IG. Among these approaches, machine learning-based solutions are the most powerful and cost-effective techniques [212]. While most proposed ML algorithms for account classification leverage supervised learning, there is always the need for an adequately labeled ground truth, which is inherently difficult to obtain for CT activities [213]. Instead, Semi-Supervised Learning (SSL) methods could be more ap-

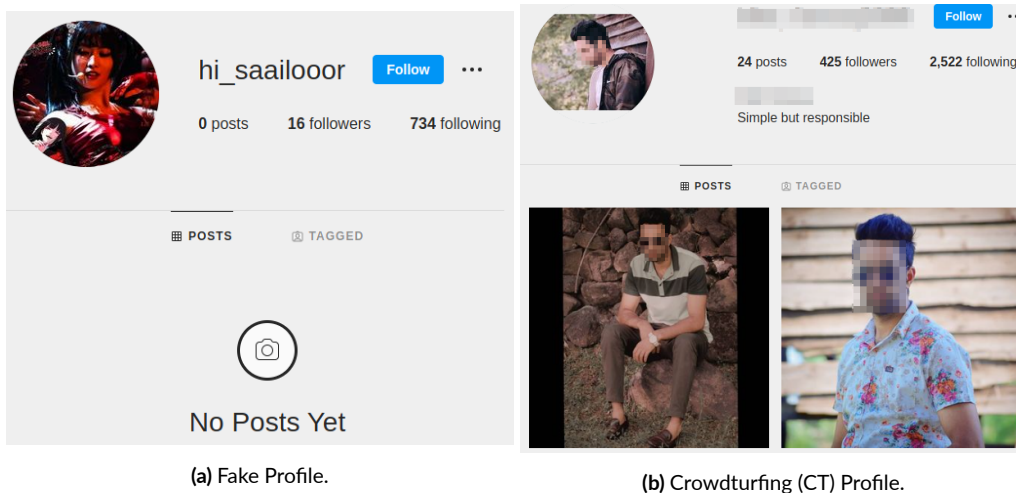


Figure 6.1: Example of fake vs crowd-turfing profiles.

appropriate when only a few labeled samples are available or needed. In fact, a large amount of unlabeled data can help improve the classification without impacting the performance [214]. Last, given the intrinsic differences between IG and other social media where CT has been studied (e.g., Twitter), we run our detector in the wild to analyze CT profiles and their fake engagement under several aspects, highlighting the difficulty of detecting such activities merely by looking at generated content.

CONTRIBUTION Our contribution is summarized as follows:

- We are the first, to the best of our knowledge, to propose a CT engagement detector on IG, which furtherly reduces the computational costs of previous fake accounts and bot detectors by leveraging semi-supervised algorithms;
- We provide a detailed analysis of CT providers to explore the services they offer and the profiles involved;
- We analyze Instagram CT engagement in the wild, mainly related to comments, by running our detector on 1000 posts generated by 20 mega-influencers.
- Our (anonymized) data will be released upon request to help researchers study CT activities on IG.

ORGANIZATION Section 6.1 presents related works. Section 6.2 examines CT providers, while Section 6.3 describes our CT detection mechanism. CT profiles and comments spotted in the wild are analyzed in Section 6.4 and Section 6.5, respectively. Section 6.6 concludes the chapter.

6.1 RELATED WORKS

We discuss CT detection in social media(Section 6.1.1), along with fake account detection on IG (Section 6.1.2) and the adoption of semi-supervised algorithms to detect bots and fake profiles (Section 6.1.3).

6.1.1 CROWDTURFING IN ONLINE SOCIAL MEDIA

Researchers have examined CrowdTurfing (or *collusive* [50]) social media activities first on Sina Weibo, and then primarily on Twitter, on which misinformation or political campaign manipulation often occurs. Wang et al. [210] investigated two popular crowd-sourcing sites in china and tracked down the CT campaigns on Sina Weibo. Then, they discussed the characteristics of CT and genuine accounts and analyzed the CT campaigns. Another work on Sina Weibo [215] examined CT accounts engaging in political activities, claiming their methodology could not find any clear evidence to show the presence of large-scale political CT. The authors of [216] categorized different types of CT tasks on Fiverr and applied ML algorithms to distinguish these tasks from legitimate ones. Song et al. [213] focused on spotting targets of CT tasks, such as posts, pages, and URLs, on Twitter. Chetan et al. [217] proposed CoReRank, an unsupervised method for detecting suspicious tweets and collusive retweeters. Dutta et al. developed several mechanisms to detect and characterize collusive users involved in black market services on Twitter [149, 218, 219]. Eventually, the authors in [220] qualitatively investigated the impact of CT activity on content visibility and popularity on IG. They claimed that IG is vulnerable to CT activities and stressed the need for a CT detector. To the best of our knowledge, we are the first to implement such a detector for IG, adopting a performing and efficient SSL approach.

As outlined in [221], social media have unique characteristics, purposes, and interactions that require tailored CT studies. For instance, researchers have recently moved their interests to YouTube [222, 223], showing that platforms besides Twitter need to be studied. We argue that IG is fundamentally different from Twitter. First, IG has roughly 1.5 Billion monthly active users (three times Twitter ones [205]), who spend three times the time spent on Twitter [224], indicating IG’s greater influence (2022). Second, they serve very different purposes [225]: Twitter lets users communicate in an elevator pitch fashion with quick messages, while IG primarily focuses on creating interactive communities through images and videos. Not surprisingly, nearly 80% of brands use IG influencers for their marketing campaigns, compared to 20% on Twitter [206, 226]. Last, while Twitter APIs¹ allow collecting a variety of users’ data (e.g., profile info, activities, connections), IG APIs² release only limited data. Due to these reasons, algorithms developed on Twitter cannot inherently apply to IG, so deploying methods to detect IG CT activities is urgently needed.

6.1.2 INSTAGRAM FAKE ACCOUNTS DETECTION

Although no prior works attempted to detect CT activities on IG, several works tried detecting fake profiles [208] or (social) bots [227, 165], which we can refer to as *classic* fake profiles. In [228], the authors developed an ML model to detect fake likes on IG, deploying honeypots and botnets to collect the ground truth. They employed ML methods to find the authenticity of likers with features including the number of followers, following, and their relationships. To detect fake and automated IG accounts, the authors in [229] applied different ML algorithms on posts and media-related features, obtaining 86% and 94% accuracy for automated and fake accounts, respectively. In [230], the authors used bagged decision trees on profile-related features to detect trivial (manually labeled) fake users. Zarei et al. [50] applied clustering methods to track down impersonators in three different categories based on their profile similarity. In [231], the authors tried to detect three categories of fake accounts: active,

¹<https://developer.twitter.com/en/docs/twitter-api>

²<https://developers.facebook.com/docs/instagram-api/>

inactive, and spammers. They bought fake accounts from Indonesian providers; however, most were simple bots, not linked to CT activities. They reached 92% accuracy using Random Forest. Kim and Hany [232] proposed a neural network to detect engagement bots by three sets of features, including text, behavior, and graph-based features. Given the existence of fake accounts and bots detection mechanisms in the literature, we will evaluate such methods on CT profiles, understanding to which extent *classic* fake accounts differ from CT accounts.

6.1.3 SEMI-SUPERVISED FAKE ACCOUNTS DETECTION

SSL approaches can leverage a vast amount of unlabeled data, reducing labeling costs with few to no drops in performances [214]. Most of these approaches were adopted on social media to detect Sybil Nodes or Bots. Sybil-Belief [233] is an SSL framework for finding Sybil nodes such as spammers and impersonators. SybilTrap [234] uses label propagation random walk as a semi-supervised transductive-learning approach to detect malicious users. This approach focuses on both structural and content-based features. Dorri et al. [235] developed SocialBotHunter as an SSL collective classification technique to detect social bots in Twitter-like platforms. Their approach uses the social behavior and interaction of users. Last, SEMIPSM [236] is an SSL Laplacian SVM model using manifold regularization to discover users responsible for propagating misinformation on social media.

6.2 CROWDTURFING PROVIDERS ANALYSIS

Table 6.1: Characteristics of Crowdturfing providers. The table reports information claimed by the provider and retrieved by analyzing 100 profiles bought from each. The last row reports info on real profiles for comparison.

Provider	Price	Delivery Time	Drop Protection	Followers Received	Followers 1 Month	#Followers Avg (std)	#Following Avg (std)	Private Profiles	#Posts Avg (std)	URLs in Biography
CT-1	\$5.69	Instant	yes	115	74	409.59 (1110.46)	812.38 (1331.52)	0.13%	14.83 (57.98)	0.08%
CT-2	\$2.39	5-10m	no	211	340	44.61 (106.85)	4679.75 (1452.19)	0%	16.0 (8.06)	0%
CT-3	\$2.95	Instant	yes	111	85	132.17 (327.28)	3027.08 (1883.18)	0.05%	20.19 (55.99)	0.09%
CT-4	\$2	Instant	no	100	42	239.45 (262.64)	2735.6 (1286.65)	0.45%	111.95 (332.2)	0.01%
CT-5	\$3.95	Gradual	yes	79	61	201.43 (214.0)	3510.77 (2316.12)	0%	16.06 (12.13)	0.054%
CT-6	\$2.89	24-72h	yes	136	129	36.79 (39.64)	2398.88 (2191.18)	0%	14.06 (5.69)	0%
CT-7	\$2.70	1h	yes	108	109	39.23 (73.32)	3966.36 (761.16)	0%	19.74 (20.13)	0%
CT-8	\$5.78	Gradual	no	110	95	57.52 (138.97)	1818.84 (1353.95)	0.04%	29.75 (41.09)	0.01%
CT-9	\$3.95	12h	no	109	99	129.54 (759.85)	2012.93 (1198.17)	0.06%	26.99 (74.94)	0%
CT-10	\$5.94	Gradual	no	97	94	83.38 (174.57)	2118.31 (1323.78)	0.03%	40.28 (51.5)	0%
Low quality	\$0.80	24-72h	no	117	96	87.26 (276.26)	3200.67 (3041.89)	0.04%	1.88 (6.15)	0.02%
Real	-	-	-	-	-	359.33 (237.87)	571.24 (517.53)	57.92%	279.09 (369.67)	14.44%

To spot CT activities, such as fake followings or comments, we must study, understand, and collect “authentic” CT profiles. Previous studies on fake profiles detection collected fake profiles or bots by manually searching for poorly designed accounts, such as those without a profile pic, with alpha-numeric names, or a very low number of posts and followers [229, 230]. Other works focused on synthetic data [228] or bought mostly naive fake accounts from local providers [231]. However, the profiles gathered using these methodologies indubitably introduce bias in the data, and the resulting detectors will identify just simple profiles, very likely driven by a bot master.

Instead, we are interested in spotting fake activities conducted by real people profiles that are populating and remaining on IG by evading its bot detection mechanisms [237, 238, 239]. To this aim, we selected 10 well-known crowdturfing providers and bought from each of them 100 fake followers. All the selected providers **ensure to**

deliver real followers (i.e., real people) who interact with the target profiles by liking and commenting on their posts to boost their engagement rate. These CT profiles are identified as high-quality followers and usually cost more than “base” fake profiles (i.e., profiles usually managed by a bot master). To identify reliable providers, we selected services that had at least an average of 3 (out of 5) stars on the famous reviews platforms TrustPilot³. Moreover, many of our CT providers allow people to freely join their platforms to participate in CT activities, confirming the reliability of the service and the presence of human activity behind the fake engagement they generate. Table 6.1 describes these providers, along with information about a low-quality provider. We also included information on real profiles we used in our study.⁴ To limit CT activities on IG, we bought CT followers for profiles we created for the study, which we deleted at the end. We are not reporting the names of the CT providers to avoid the encouragement of such activities.

The table shows that the price average is pretty low, around \$3 for 100 high-quality followers, but much higher than the \$0.80 for 100 low-quality followers. Followers are usually delivered within a few hours, and most providers offer drop protection, replenishing any lost follower. To assess the providers’ reliability, we checked how many followers remained after one month. On average, we lost only 15-20% of them, and sometimes, we gained more. CT-4, the least expensive provider, lost the most, while we lost only 3 followers from the most expensive CT-10. Compared to real profiles, CT profiles have a noticeable difference in followers and following. This is understandable, given that the more they follow and interact, the more they get paid. However, from CT-1, the second most expensive provider, the follower/following balance is quite close to real profiles. The CT profiles also are quite different from real ones in terms of being private, the number of posts, and the URLs in the biography. Very likely, CT platforms require profiles to be public. People joining these platforms generate a minimum amount of posts to be reliable, except few cases (CT-4, CT-10). The low-quality profiles show a very high imbalance in followers and following, and the average number of posts is close to 0, far below the CT profiles. Among the properties not included in the table, some providers allow customers to buy followers from a specific region or language or increase their followers periodically.

6.3 CROWDTURFING PROFILES DETECTION

Instead of directly detecting CT activities (e.g., a fake comment), we first detect profiles involved in CT activities, and accordingly, we label their interaction as CT. The rationale behind this approach is that CT profiles are mainly **real accounts** belonging to individuals willing to create fake interactions. Thus, their interactions should resemble genuine ones, in both content and temporal activity [221]. Similarly, their profile information should appear legitimate, which makes detecting CT profiles considerably different from spotting *classic* fake profiles [218] (i.e., the focus of previous works). Indeed, the latter usually present simplistic features (e.g., no posts, no followers), or recognizable patterns (e.g., low-variability content) [208]. We now present the dataset we collected to classify CT profiles (Section 6.3.1), our detection model (Section 6.3.2), and a comparison with previous approaches (Section 6.3.3).

³<https://www.trustpilot.com/>

⁴Here, to simplify comparisons, we excluded celebrities and highly-followed accounts (> 500 followers), which could present inflated statistics.

6.3.1 DATASET AND FEATURE SELECTION

Since there are no IG CT datasets available, we collected our own. Given IG API could not provide our requirements, we performed automated data collection through Selenium⁵. For our detector, we use general profile info (e.g., #followers, #following, #posts) instead of behavioral patterns since IG does not provide such information, unlike other social media (e.g., Twitter). Some previous works [228, 231] used features that are not publicly available, e.g., the number of likes of a user’s posts, limiting their approach only to public profiles. Instead, we focused only on profile features that are publicly available for both public and private profiles.⁶

The dataset contains the profile information of 2600 users, including 1293 CT and 1307 authentic accounts. The CT profiles are the ones analyzed in Section 6.2. We gathered authentic accounts similarly to previous works [230, 229, 231]. We included from several countries and fields: general users from our expanded social connections, verified or business accounts, and celebrities. Three authors validated these accounts through extensive manual labeling, adopting a majority voting for the decision, and focusing on attributes such as the Follower/Following imbalance, the number of posts, or the full name. The feature distributions of our real accounts (Table 6.1) closely align with previous works. For the collected accounts, we gathered all the attributes available on the profile page. Then, we pre-processed the features by removing those with zero or very low variance. Last, we transformed categorical and non-numeric attributes into numeric or boolean features. The final features are shown in Table 6.2. Since all the data we collected is public, we will make it available upon request (anonymized) to help the research community studying CT.

Table 6.2: Final set of features of our dataset.

<i>Numeric Features</i>	<i>Boolean Features</i>
# followers, #following	Account is private
# videos, #posts	Account is verified
# char in username, #digit in username	Account has clips
# characters in fullname	Account is business account
# characters in biography	Account has external URLs
# non-alphabetic char in fullname	Accounts has category name
# hashtags and mentions in biography	Account has multiple categories

6.3.2 OUR SEMI-SUPERVISED MODEL

The next step is to develop a detector to distinguish between real and CT profiles. In light of previous discussions, labeling CT profiles manually is challenging, since they resemble legit users [213]. Therefore, instead of adopting supervised methods as in previous works, we use SSL to maximize the use of unlabeled data and improve generalization. While most previous SSL approaches on social networks utilized graph-based methods, we consider

⁵<https://www.selenium.dev/>

⁶Some attributes (e.g., #videos) were retrieved from the page source code.

only profile-related features, making our model less complicated and easier to handle (e.g., for practitioners). Our self-training approach is depicted in Figure 6.2.

Initially, we divide the dataset into labeled and unlabeled datasets, discarding all labels from the unlabeled dataset. In the first training cycle (dashed arrows in the figure), training data corresponds to labeled data. We train a classifier with this data and ask it to predict the labels of all the unlabeled samples, generating their pseudo-labels. For each pair *sample:pseudo-label*, we check the prediction probability (i.e., classifier confidence, from 0 to 1) associated with the pseudo-label. If the probability is higher than 0.75, we add the pair *sample:pseudo-label* to the training data; otherwise, the sample remains unlabeled. We repeat the training cycle (train the classifier → predict pseudo-labels → enlarge the training set) 10 times or until no unlabeled data remains. The final model corresponds to the classifier of the last iteration.

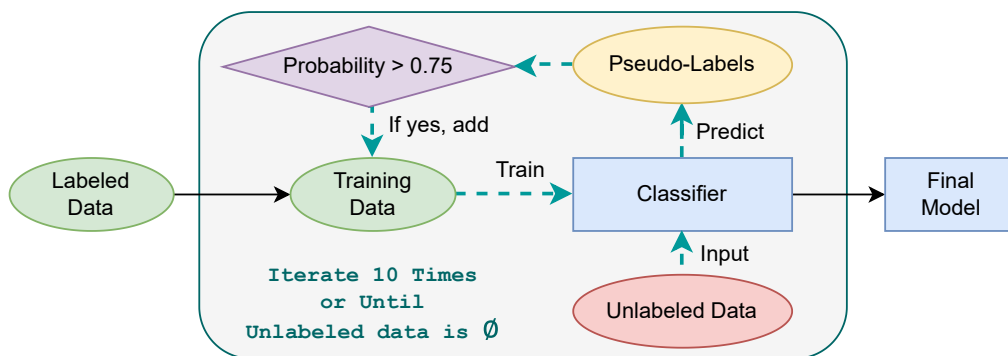


Figure 6.2: Self-Training process. Dashed arrows represent the training cycle.

We implemented our models using Scikit-learn.⁷ We randomly split our dataset in a stratified mode to have 80 percent of data for training (2080 labeled samples) and the remaining for testing. We applied 5-StratifiedKFold Cross-Validation on the training set to find the best model hyper-parameters. In each iteration, one fold (~416 samples) was left out to validate the model, while the remaining ~1664 samples were used to train the classifier in the semi-supervised fashion described above. To demonstrate the power of SSL, we tested different (small) labeled training data portions: 1, 3, 5, and 9%. As classifiers and hyper-parameters, we tested:

- K-Nearest Neighbor (KNN): $n_neighbors=[1, 3, 5, 10]$;
- Logistic Regression (LR): $penalty=[\text{none}, l1, l2]$, $C=[10, 1, 0.1]$, $solver=[\text{lbfgs}, \text{liblinear}]$;
- Decision Tree (DT): $max_depth=[\text{none}, 3, 5, 10]$, $samples_leaf=[1, 3, 5, 10]$;
- Random Forest (RF): $max_depth=[\text{none}, 3, 5, 10]$, $samples_leaf=[1, 3, 5, 10]$, $n_estimators=[10, 100]$.

The best hyper-parameters for each model were selected through a grid-search approach. We also trained the classifiers on all the labeled data (i.e., in a supervised mode) for comparison. The results for each classifier using the best hyper-parameters during the cross-validation are reported in Table 6.3.

The table shows that increasing labeled data does not necessarily improve the model’s performance but increases its stability. Moreover, the results in the SSL mode do not differ significantly from supervised ones. This

⁷<https://scikit-learn.org>

Table 6.3: Average \pm std of classification results of the best models during cross validation. Sup = Supervised.

<i>Model and % Labels Used</i>	<i>Train Accuracy</i>	<i>Valid. Accuracy</i>	<i>Valid. Precision</i>	<i>Valid. Recall</i>	<i>Valid. F-Measure</i>	
KNN	<i>0.01</i>	0.79 \pm 0.04	0.79 \pm 0.03	0.84 \pm 0.02	0.79 \pm 0.04	0.78 \pm 0.04
	<i>0.03</i>	0.92 \pm 0.04	0.92 \pm 0.02	0.92 \pm 0.03	0.92 \pm 0.04	0.92 \pm 0.04
	<i>0.05</i>	0.93 \pm 0.02	0.94 \pm 0.01	0.93 \pm 0.02	0.93 \pm 0.02	0.93 \pm 0.02
	<i>0.07</i>	0.92 \pm 0.00	0.92 \pm 0.00	0.92 \pm 0.00	0.92 \pm 0.01	0.92 \pm 0.00
	<i>0.09</i>	0.96 \pm 0.01	0.95 \pm 0.00	0.96 \pm 0.01	0.96 \pm 0.01	0.96 \pm 0.01
	<i>Sup.</i>	0.97 \pm 0.01	0.97 \pm 0.00	0.97 \pm 0.01	0.97 \pm 0.01	0.97 \pm 0.01
LR	<i>0.01</i>	0.97 \pm 0.01	0.97 \pm 0.00	0.97 \pm 0.01	0.97 \pm 0.01	0.97 \pm 0.01
	<i>0.03</i>	0.78 \pm 0.09	0.78 \pm 0.10	0.85 \pm 0.05	0.78 \pm 0.09	0.77 \pm 0.10
	<i>0.05</i>	0.94 \pm 0.02	0.94 \pm 0.02	0.95 \pm 0.02	0.94 \pm 0.02	0.94 \pm 0.02
	<i>0.07</i>	0.92 \pm 0.07	0.92 \pm 0.06	0.93 \pm 0.05	0.92 \pm 0.06	0.92 \pm 0.07
	<i>0.09</i>	0.96 \pm 0.01	0.96 \pm 0.00	0.96 \pm 0.01	0.96 \pm 0.01	0.96 \pm 0.01
	<i>Sup.</i>	0.96 \pm 0.01	0.96 \pm 0.00	0.96 \pm 0.01	0.96 \pm 0.01	0.96 \pm 0.01
RF	<i>0.01</i>	0.87 \pm 0.03	0.87 \pm 0.02	0.88 \pm 0.02	0.87 \pm 0.03	0.87 \pm 0.03
	<i>0.03</i>	0.92 \pm 0.02	0.93 \pm 0.02	0.92 \pm 0.02	0.92 \pm 0.02	0.92 \pm 0.02
	<i>0.05</i>	0.90 \pm 0.02	0.90 \pm 0.01	0.90 \pm 0.02	0.90 \pm 0.02	0.90 \pm 0.02
	<i>0.07</i>	0.90 \pm 0.03	0.91 \pm 0.01	0.91 \pm 0.02	0.90 \pm 0.03	0.90 \pm 0.03
	<i>0.09</i>	0.95 \pm 0.01	0.96 \pm 0.01	0.95 \pm 0.01	0.95 \pm 0.02	0.95 \pm 0.02
	<i>Sup.</i>	0.97 \pm 0.00	0.97 \pm 0.00	0.97 \pm 0.00	0.97 \pm 0.00	0.97 \pm 0.00
DT	<i>0.01</i>	0.81 \pm 0.05	0.82 \pm 0.05	0.82 \pm 0.05	0.81 \pm 0.06	0.81 \pm 0.06
	<i>0.03</i>	0.88 \pm 0.04	0.88 \pm 0.04	0.88 \pm 0.04	0.88 \pm 0.04	0.88 \pm 0.04
	<i>0.05</i>	0.91 \pm 0.01	0.92 \pm 0.01	0.92 \pm 0.01	0.91 \pm 0.01	0.91 \pm 0.01
	<i>0.07</i>	0.90 \pm 0.02	0.91 \pm 0.01	0.90 \pm 0.02	0.90 \pm 0.02	0.90 \pm 0.02
	<i>0.09</i>	0.93 \pm 0.01	0.93 \pm 0.01	0.93 \pm 0.01	0.93 \pm 0.01	0.93 \pm 0.01
	<i>Sup.</i>	0.95 \pm 0.01	0.97 \pm 0.00	0.95 \pm 0.01	0.95 \pm 0.01	0.95 \pm 0.01

suggests that CT profiles share similar characteristics, as partially discussed in Section 6.2, and algorithms can converge by taking a few labeled data. On the contrary, adding more samples could lead to over-fitting or biasing the classifier, reducing prediction accuracy (as happened for LR 0.03). The LR classifier with 1 percent of labeled data (penalty = l_2 , C = 1, solver = *liblinear*) showed the best cross-validation results among the semi-supervised models, so it was selected as the final model.⁸ Such a model reached 95% accuracy and F1-score on the test set and was used in the remainder of our analyses.

6.3.3 BASELINE COMPARISON

To assess the quality of our results, we compared them with previous IG fake and bot account detection mechanisms [228, 230, 229, 231]. Only Akyon et al. [229] released their data, so we used their dataset comprising authentic and fake/bot accounts to train all the baselines, adapting the features and re-implementing the models. Each baseline was tested on all our CT accounts (provider by provider) and real accounts. Table 6.4 reports the avg \pm std in detecting CT profiles for each provider. Our algorithm outperforms all the baselines, being statisti-

⁸We discarded RF sup. (same scores) since the chapter focuses on SSL. Practitioners should choose models with the best performance.

cally better⁹ than the best baselines for Recall (p -value < 0.05) and F1-score (p -value < 0.01). The lower baselines’ recall can be explained by CT accounts resembling real accounts characteristics, avoiding detection as expected. However, the relatively high standard deviations imply that the quality of CT providers varies significantly, i.e., some of them deliver lower-quality accounts, detectable by previous methods. The presence of low-quality profiles also highlighted in Table 6.1, allowed our detector to spot both CT and *classic* fake accounts, making it more reliable than previous models trained on simple bots or synthetic data.

Table 6.4: Baseline comparison in detecting CT profiles.

<i>Baseline</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
Thejas et al. [228]	0.77 \pm 0.05	0.89 \pm 0.09	0.82 \pm 0.06
Sheika et al. [230]	0.94 \pm 0.02	0.84 \pm 0.11	0.88 \pm 0.07
Akyon et al. [229]	0.87 \pm 0.05	0.83 \pm 0.19	0.84 \pm 0.14
Purba et al. [231]	0.92 \pm 0.03	0.80 \pm 0.14	0.85 \pm 0.09
Our	0.95 \pm 0.02	0.95 \pm 0.03	0.95 \pm 0.02

We now explore the features’ importance to explain why baselines performed worse. Figure 6.3 shows our model coefficients based on standardized features, so they are comparable. Baselines’ most predictive features were the number of posts, following, followers, and bio length [228, 230, 229, 231]. While the number of following and posts is also crucial for us, the followers and bio length are less influential. The reason is that bots and simple fake accounts tend to have few followers and no bio, thus biasing baselines. Instead, CT profiles usually have many followers and genuine bios since they are real people profiles. Moreover, baselines do not leverage username and fullname characteristics, the number of videos, and if an account is private or verified, which are relevant to us. This suggests our model performs better due to the training data that includes CT profiles and the features we extracted (e.g., # digits in username) rather than the model itself. Nonetheless, we contribute to the state-of-the-art by demonstrating that (i) *classic* fake accounts detectors are not enough to effectively detect CT profiles, (ii) the training data are more important than the detection algorithm, (iii) the task can be efficiently solved with SSL algorithms, significantly reducing (99% less!) the time and costs to label data.

6.4 CROWDTURFING ANALYSIS: PROFILES INFORMATION

With our CT profile detector trained, we are ready to analyze CT engagement in the wild. In our detection strategy, we detect profiles involved in CT using our model (Section 6.3.2) and label their engagement accordingly. Since CT profiles contribute to a fake engagement, we will also refer to them as *fake* (non-genuine) accounts and engagement vs. *real* (genuine) ones. For our analyses, we collected the comments and commenters’ profile information¹⁰ of 50 recent posts of 20 mega-influencers with over 1 million followers (1000 posts in total). We selected posts at least five days old to allow IG automatically remove *classic* fake interactions [237, 238]. The

⁹Unpaired t test with $\alpha = 0.05$ as significance threshold.

¹⁰Profiles info were collected via Instaloader <https://instaloader.github.io/>.

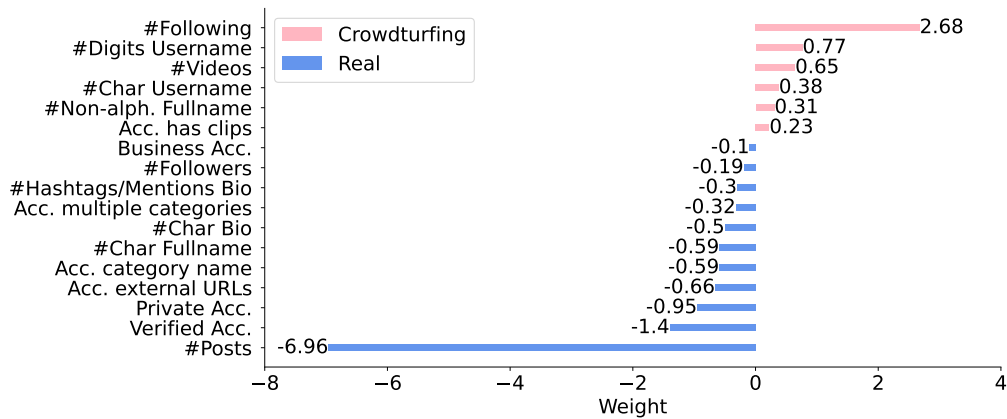


Figure 6.3: Logistic Regression weights to discriminate Crowdturfing (positive label) vs Real (negative label) profiles.

influencers come from different nationalities and the following categories: fashion, beauty, fitness, art, music, lifestyle, and family. In total, we gathered 603,007 comments generated by 248,388 unique users. The reasons why we collected only comments-related information and e.g., not likes, are discussed in the comments analysis section (Section 6.5).

Our CT detection model detected 55,719 CT profiles among the 248,388 collected (~22%). This percentage aligns with the estimate of 20-40% in celebrities' accounts shown in Chapter 2. We acknowledge that some of the detected accounts may not be CT; however, we are still dealing with “advanced” fake profiles that have bypassed (i) the automatic screening mechanisms employed by IG [237, 238] (and Meta [240] in general), and (ii) the potential moderation done by the influencers themselves (e.g., by removing blatant spam comments). Therefore, we can assume our further analyses will primarily focus on CT or advanced fake profiles that resemble and act as legitimate profiles. In this section, we provide a detailed study of CT profiles' information, including the number of followers and following (Section 6.4.1), biography (Section 6.4.2), and external URLs (Section 6.4.3), to determine whether CT profiles engage in malicious activities besides crowdturfing.

6.4.1 FOLLOWERS AND FOLLOWING RATIO ANALYSIS

To increase other accounts' engagement (and therefore gain more money), a fake account will display an unusually high number of following (Section 6.2). Conversely, genuine users should have a more balanced ratio of followers and following according to IG averages [241]. Figure 6.4 shows the mean and std of followers/following for fake and real users. Real users are divided into normal and influencers tiers¹¹ as follows:

- *Normal accounts*: less than 1,000 followers;
- *Nano influencers*: $1,000 \leq \text{followers} < 10,000$;
- *Micro influencers*: $10,000 \leq \text{followers} < 50,000$;
- *Mid-tier influencers*: $50,000 \leq \text{followers} < 500,000$;

¹¹<https://www.shopify.com/id/blog/instagram-influencer-marketing>

- *Macro influencers*: $500,000 \leq \text{followers} < 1,000,000$;
- *Mega influencers*: more than 1,000,000 followers.

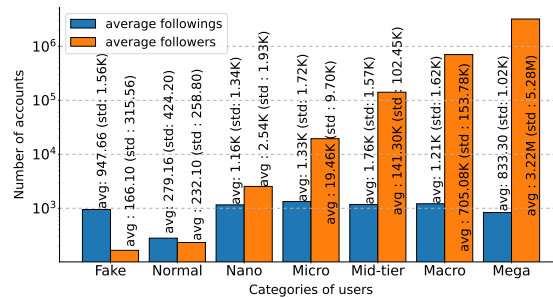


Figure 6.4: Followers and following avg and std of CT users (Fake) and different categories of real users. Y-axis is in log scale.

The graph shows that the number of followers of fake users is (on average) much smaller than the number of following. Indeed, these accounts are incentivized to follow more people to grow their earnings through CT activities, confirming our initial assumption. Following and followers of normal users are balanced, but as the popularity of the genuine account grows, followers increase exponentially while following hovers around 1000. For more popular influencers, the standard deviation increases simply because their categories include wider ranges (e.g., from one to hundreds of million followers for mega influencers). We further inspected the following distribution of CT accounts in Figure 6.5. Most CT accounts have between 0 and 500 following, with the number

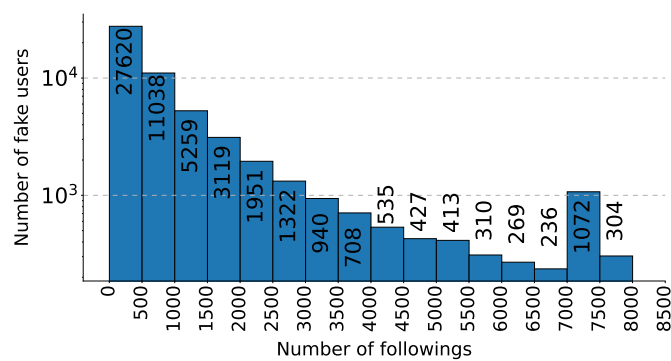


Figure 6.5: Distribution of fake accounts' following.

decreasing as the following increases, suggesting CT accounts tend to maintain a low profile to avoid being flagged as spammers. An exception occurs in the last two bins. IG introduced a 7500 following limit¹² to contrast spamming activities, and many CT (probably more similar to *classic fake*) accounts are just below this limit. Despite it, 304 fake profiles likely surpassed the threshold before its introduction.

¹²<https://help.instagram.com/408167069251249>

6.4.2 FAKE PROFILES BIOGRAPHY ANALYSIS

Many *classic* fake IG accounts use a catchy biography to lure victims into clicking malicious links. Thus, we tried to find suspicious words in the CT users' biographies. To this aim, we created a list of 31 elements, including words and emojis often used by this fake user. The list, based on our knowledge of fake behavior and a brief manual inspection, contained words like "stories", "chat", "follow", "gain", "click", "link", and emojis usually linked to malicious or sexual activities, like "👁️", "👇", "👉", "👀", "👄", as we will show in Chapter 7. Only 5635 CT accounts (10.11% of the total detected) had at least one of the elements of the list. Thus, most CT accounts do not seek to boost their profiles or induce people to click links. Rather, they are interested in making profits by increasing the engagement of other accounts.

6.4.3 FAKE PROFILES EXTERNAL URLS ANALYSIS

Last, we analyzed CT accounts' external URLs to understand the most used categories and whether they could be vectors of attacks conducted over social networks [242]. Of the total fake accounts, only 2834 (5.08%) had an external URL on their profile page. We grouped these 2834 URLs into the following categories:

- *Videogame*: Youtube, Twitch, Discord;
- *Messaging*: WhatsApp, Telegram;
- *Social Network*: Facebook, Twitter, Instagram, etc.;
- *Music & Photography*: Spotify, Soundcloud, VSCO.co;
- *Email & Google services*: Gmail, Maps, Outlook;
- *URL redirecting*: Linktr.ee, Tinyurl, Linkr.bio, Bit.ly;
- *Shopping & Payment*: PayPal, Vinted, Amazon, etc.;
- *Personal website & Petition*: Blogspot, Wordpress, etc.;
- *Adult content*: URLs to different adult websites;
- *Other*.

Inside the categories, we also included shortened URLs (e.g., wa.me or t.me for WhatsApp and Telegram, respectively). The results are shown in Figure 6.6. Remark that even if the categories contain well-known websites, some can be used for malicious purposes. For instance, we found many WhatsApp links starting a conversation or a phone call with strangers who could easily be scammers. Similarly, we inspected and monitored Telegram URLs, grouping them into:

- *Conversation*: Similarly to WhatsApp URLs, starts a conversation with a potential scammer;
- *Piracy*: Illegal groups that share movies and tv series;
- *Selling*: Scam groups that try to sell clothes, Amazon gift cards, cryptocurrencies, NFTs, etc.

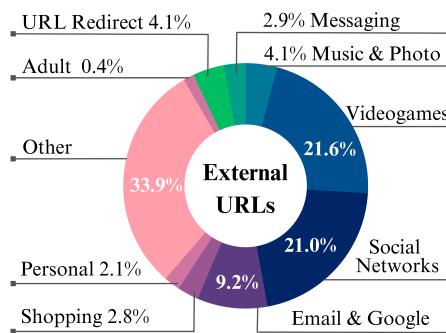


Figure 6.6: Categories of External URLs of the fake profiles.

Moreover, *classic* fake profiles commonly use redirect URLs to route the victim to a malicious site [243]. From Figure 6.6, it is possible to see that the “Other” section is more relevant than the other categories inside the pie chart, with precisely 961 URLs. It contains very heterogeneous URLs, making their categorization challenging. To better understand these URLs’ nature (i.e., if they are malicious), we have relied on a fraud prevention and detection service called Ipqualityscore.¹³ It allows checking for suspicious links by using a mixture of blacklists and deep learning algorithms, and to define the following URLs categories:

- *Parked*: Domains that have been dormant for a long time;
- *Spamming*: Websites that spams malicious content;
- *Malware*: Websites hosting viruses, malware, etc.;
- *Phishing*: Websites hosting fake login, or sign up forms;
- *Adult*: Websites that contain adult content.

The results of this evaluation are shown in Figure 6.7. For convenience, we grouped the “Phishing”, “Malware”, and “Adult” categories since they had very few matches. From the total 961 “Other” URLs, 599 were considered safe, while the remaining 362 were divided as follows:

- 190 URLs were parked and/or spamming websites;
- 5 URLs were marked as malware websites;
- 7 URLs were marked as phishing websites;
- 11 URLs were adult websites;
- 149 were considered suspicious websites.

These results show that most external URLs in the “Other” category were considered safe. However, many spamming and suspicious websites can be used for malicious purposes. Comparing the obtained results to the overall number of CT users, we can confirm that most are solely involved in CT activities rather than malicious activities.

¹³<https://www.ipqualityscore.com/>

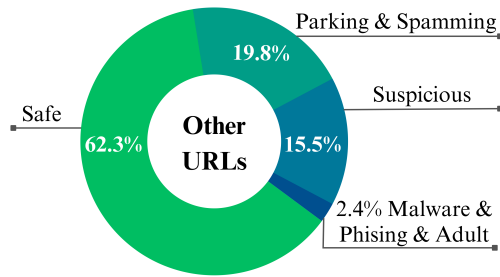


Figure 6.7: Results provided by the fraud prevention and detection service on the URLs in the “Other” category.

6.5 REAL VS CROWDTURFING COMMENTS ANALYSIS

This section analyses CT engagement. In particular, we aim to understand if CT can be directly spotted by actions (e.g., comments) instead of leveraging profile information. As stated before, CT profiles are driven by humans, so intuitively, there should be little to no difference between real and fake engagement, but we cannot draw conclusions without proper analysis. On IG, the primary forms of engagement are liking and commenting. CT likes cannot be isolated from the action itself since it carries no information beyond temporal data (unavailable on IG). Instead, comments provide valuable information (e.g., stylometric features) that could be used for CT detection. Moreover, comments present a higher level of public expression than likes [66] and are considered more important to boost the visibility of an account [51, 244]. For these reasons, we focus on comments in this section, presenting five studies to spot the differences between comments made by CT and real users.

6.5.1 STYLOMETRIC ANALYSIS

From our dataset, we isolated 121,822 comments shared by CT users and 481,185 from legit ones. We performed a stylometric analysis similar to the one conducted in [245], based on Lexical Features, Syntactical Features, and Emoji Features.

LEXICAL FEATURES. We calculated the number of sentences per comment, the number of words in each comment, the number of words in each sentence, and the length of the comments. We found several statistically significant (p -value < 0.001) differences: CT users have an overall mean of 1.13 words per comment, while the real ones have 4.34. Similarly, the number of words per sentence is 0.94 for the CT accounts and 2.96 for the real ones. Instead, both categories of users have a mean of 1.35 sentences per comment. In each comment, there is a low repetition of words: we obtained that 99% of them, made by CT users, have no word repetitions, while for the legit users is 97%. Another important distinction is the length of the comments: the CT users shared text with a mean length of 28.89 (std: 61.19) characters (emojis included), while the legit users have a mean of 23.74 (std: 46.74). Even if similar, they are statistically significant (p -value < 0.001). The emoji comparison better explains how real users, with more words, have shorter comments.

SYNTACTICAL FEATURES. We counted the number of comments starting with a capital letter, punctuation present in the text, and capital words. We found very close results between CT and real users: the beginning of the comment is in uppercase for 33.86% of comments made by CT users and for 34.94% of real ones. 35% of the comments have some punctuation for both accounts categories. Finally, we saw that both categories do not use upper-cased words: the mean of the ratios between uppercase words and all the words in each comment are 0.021 for the CT users and 0.025 for the legit ones.

EMOJI FEATURES. We detected emojis in the comments using `demoji`¹⁴. Our study focused on the presence of emojis and alphanumerical text in the comments, in particular:

1. **The percentage of comments with at least one emoji;**
2. **Most used emojis:** the percentage of an emoji among all the fake comments. Multiple occurrences of the same emoji on the same comment increase the counter by one.
3. **Avg emojis when present:** considering only comments presenting emojis, the avg number of them. Multiple occurrences of the same emoji increase the counter accordingly.

Table 6.5 reports the results. The top-most emojis used are equal for both users, with similar percentages. Another meaningful result is that even if real users have, on average, slightly more comments with emojis, the quantity of emojis in such comments is fewer compared to CT users. This result might explain the outcomes on comment length found in Section 6.5.1. To sum up, results obtained so far show some stylometric differences, but mostly similarities between CT and real users when the focus is on emoji used, sentences per comments, or syntactical features. Legit users share comments with more words, fewer emojis, and an overall shorter comment length.

Table 6.5: Emoji-based Stylometric analysis. CE = Comments with Emoji, EPC = Avg Emoji per comment.

	<i>CE</i> (%)	<i>% of Most used Emoji</i>					<i>EPC</i>	
		❤️	😍	🔥	👉	😂	👏	
Fake	71.6	25.18	19.92	10.57	4.91	4.03	2.73	3.557
Real	72.7	22.30	18.46	14.42	5.00	4.92	3.04	3.211

6.5.2 COMMON WORDS ANALYSIS

We analyzed the most common words CT and real profiles use. As a pre-processing, we removed emojis, punctuations, and unproductive words with less than three characters, e.g., “and”, “the”, “you”. The word clouds in Figure 6.8 show fake and real users’ top 100 most used words. In general, we found a lot of positive and loving expressions, such as beautiful, love, happiness, niceness, etc.

An interesting word from Figure 6.8a is “Dokter”, which appeared in 1069 comments. By investigating the accounts spamming this word, we might have found a botnet whose objective is to spam “IG doctors” accounts.

¹⁴<https://pypi.org/project/demoji/>



Figure 6.8: Most used words by fake and real users.

All these doctors’ profiles have a WhatsApp business link starting a chat with a message to complete: “*NAME*: *CITY/STATE*: *ORDER/COMPLAINTS*: *AGE*:”. Some doctors’ accounts no longer exist on IG, suggesting they probably violated the ToS. Other similar accounts had the format “dr . [doctor_name]”, presenting the same WhatsApp link and conversation, but different phone numbers. We found 1370 comments coming from 33 different accounts containing such words, suggesting the presence of a bigger malicious network.

6.5.3 NUMBER OF COMMENTS PER USER

248,388 unique users posted the 603,007 comments we analyzed; thus, many users posted multiple comments. We found that a legit user, on average, has posted 1.95 comments (std 5.94), while a CT user has posted slightly more (2.24, std 7.57). The result obtained in this analysis complies with the one in Section 6.5.1: a CT user has a similar behavior as the legit user. However, a CT user generally shares more comments than a real one because their purpose is to generate engagement. But to avoid IG bot detection, an account has to act like a real human being.

6.5.4 LANGUAGE ANALYSIS

We analyzed the language used by CT and real users using SpaCy.¹⁵ The text was filtered out of emojis and then used as input for the neural network. The results are shown in Figure 6.9. In both CT and real comments, we found that the prominent language is English (35.2% and 43.5%, respectively), followed by Japanese and French. The “Other” slices include more than 100 languages, each with a presence below 2%. They are probably the second largest sections of the pie charts because many comments just mentioned other accounts or used single words, complicating the language detection process. Besides that, CT users likely adopt the language of their target community or, more commonly, English. In fact, as stated in Section 6.2, many CT providers allow the option to deliver followers from specific geographical locations.

¹⁵<https://spacy.io/usage/facts-figures>

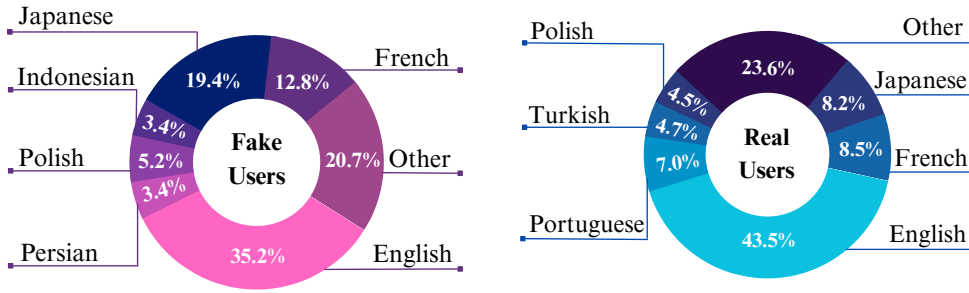


Figure 6.9: Languages detected in comments.

Table 6.6: Top-10 topics extracted from fake and real comments.

Fake Comments				Real Comments			
<i>N. Comm.</i>	<i>Top Words</i>		<i>Label</i>	<i>N. Comm.</i>	<i>Top Words</i>		<i>Label</i>
3817	beautiful gorgeous sexy perfect hot amaze girl		Female Beauty	13034	beautiful gorgeous nice cute pretty lovely girl		Female Beauty
2290	love beautiful cute smile god woman world girl		Love (woman)	7547	love good smile congrats great brother bro wish		Love (Males)
2117	good want video well thank man bro life bike work		Man Compliment	6983	dream make want come time good life day hope		Life Dreams
1755	happy birthday halloween republic thanksgiving		Pagan holidays	6274	christmas merry god bless family thank bible		Christmas
1476	please christmas merry story follow check thank		Christmas/Follow	6035	happy new year birthday day family love republic		Pagan holidays
1136	help fire turkey people stop please give help-turkey		Help Turkey	6008	help need fire people turkey please animal world		Turkey/Ecologists
1021	trop wanna kiss lip red face belle pretty liplock		Kiss & Face	5658	picture crazy bro think top video sick bike man		Exalt Men
674	arm chest belly waist neck armpit thigh dance		Body parts	5524	follow check story post page like support profile		Follow/support
514	problem solution wife money call whatsapp expert		Problems/Ads	4846	please congrats reply check story dance song real		Music
223	love back help massage oil bubbs real magic		Relax	1381	problem belle family life help solution marriage		Family Problems

6.5.5 TOPICS ANALYSIS

To further investigate the behavior of CT and real users, we inspected the topics in their comments. Many state-of-the-art topic modeling algorithms, such as Latent Dirichlet Allocation (LDA), require long text to extract topics. However, social network comments are usually concise sentences, making the topic modeling more challenging. In our experiments, we used GPU-PDMM[246], which is typically adopted to extract topics of tweets. Based on the Poisson-based Dirichlet Multinomial Mixture (PDMM) model, GPU-PDMM promotes the semantically related words under the same topic during the sampling process by using the Generalized Polya Urn (GPU) model. We considered only English comments for the analysis, after removing non-alphabetical characters, emojis, stop words, words shorter than three characters, and applying lemmatization. From our comments, 15,023 CT comments and 63,290 Real comments were suitable for the study. We instructed the model to distinguish ten topics in an unsupervised fashion, returning for each comment the belonging topic and the top words associated with each topic. The results of the topics inference are shown in Table 6.6.

As expected, we find high alignment between topics covered by CT and real profiles. Most comments exalt female beauty, using compliments, love words, or positive feelings to boost engagement. In particular, CT comments contain more exaggerated terms, such as “sexy”, “perfect”, or “amaze”. Conversely, we found few advertisement comments, likely to avoid being flagged as spammers. An interesting difference between CT and real comments is how they dealt with the *Help Turkey* topic. For real profiles, we found additional words such as “animal” and “world”, suggesting they also brought up other environmental arguments, while CT did not. For

real comments, we also found a *Follow/support* topic, which could be a false positive (some spammers were not detected) or that they did not care about being labeled as spammers. In summary, the topic analysis revealed some differences, but not consistently enough to allow for proper differentiation.

6.6 CONCLUSION

In this chapter, we developed an algorithm that leverages profiles' characteristics through semi-supervised learning to spot IG crowdturfing activities. To train our classifier, we purchased CT profiles from 11 providers, which we further studied to understand their services and the type of profiles involved in them. Our Logistic Regression classifier scored 0.95% F1-score. To spot IG CT activities in the wild, we targeted the most recent posts of 20 influencers of different nationalities and categories. We mainly focused on comments, as they are a crucial engagement metric for accounts' visibility, and carry more information than likes. For this purpose, we collected 603,007 comments among the different posts made by 248,388 unique users. Our model labeled 55,719 of these profiles as CT accounts. We compared CT profiles and comments with genuine ones, concluding that CT activities would be difficult to detect based only on their activities. Indeed, CT profiles are mostly real profiles guided by real humans; thus, their activities are close to genuine ones. In contrast to bots or fake profiles, they seem to not be involved with malicious activities besides boosting other accounts' engagement. In the future, we plan to distinguish between CT profiles and other "advanced" fake profiles we might have (in)voluntarily encountered in our analyses. While IG and the research community focused a lot on detecting bots and automated accounts, we believe more studies should be conducted on CT activities or in general, advanced fake profiles which negatively impact influencer marketing, IG, and most of its users.

ETHICAL CONSIDERATIONS

We faced two main ethical challenges: CT activities' involvement and data collection on IG. Our experiments were designed following the exemption guideline from a formal review by our institute's IRB. To deal with CT activities, we acted similarly to previous works that analyze underground activities [247, 213, 220], first by dealing only with a small number of CT followers and platforms, minimizing our effect on them and IG. Second, we linked the followers to freshly created accounts that had no prior connection with other IG accounts, and we deleted them at the end of the study. Thus, CT activities were not involving legitimate users.

For data collection, we gathered only profiles' information and comments publicly available, removing all the information linked to individual subjects (e.g., name, profile picture). Similar to previous works [248, 249], we could not request informed consent to prevent participants from (in)voluntarily changing their behavior, causing the Hawthorne effect [250]. Since IG APIs do not return all the public information of a user's profile, yet visible by simply browsing it, we collected such data in an automated way, which is not allowed by the ToS. However, as argued by Fiesler et al. [142], "ethical decisions regarding data collection should go beyond ToS and consider contextual factors of the source and research". In particular, IG ToS allows manual collection, suggesting that automated collection is probably not allowed to avoid heavy servers' workload [142]. Therefore, we tuned our

tools to collect data at a slow human-like pace, using only our 11 profiles over five months, avoiding any ban from the platform.

7

Turning Captchas Against Humanity: Captcha-Based Attacks in Online Social Media

7.1 INTRODUCTION

The so-called Web 2.0, or Social Web, emphasizes user-generated content and stimulates a participatory culture, giving birth to virtual communities. Examples are Wikis, Blogs, Social Sharing Platforms (e.g., YouTube), and Online Social Networks (OSN). To grant users a safe environment, human content moderators control users' activity and remove malicious content, such as hateful messages or violent and sexually explicit media. However, nowadays, users generate more content that humans can moderate. For instance, Instagram and Facebook count about 350 thousand stories and 150 thousand photos posted by their users, respectively, every minute [251]. Besides, content such as child pornography, hate-filled messages, and gratuitous violence can cause considerable psychological risks to human moderators [252]. Thus, the need for **automated** moderators is constantly increasing.

Researchers and companies started developing automatic tools to tackle the problem of “malicious content” detection efficiently. These tools are mainly solved with data-driven approaches, e.g., machine learning (ML), like in the case of hate-speech detectors [253, 254]. Recently, OSN like Facebook adopted Automatic Content Moderators (ACM) to help human moderators. As reported by *The Verge* [255], Facebook employs ML tools to monitor users' posts to spot potential inappropriate content that human operators will manually review. Such content is either removed or labeled as “sensitive”, which means users have to explicitly accept to view it. Instagram recently adopted a similar system [256], stating that technology and humans cooperate to identify sensitive content (Figure 7.1).

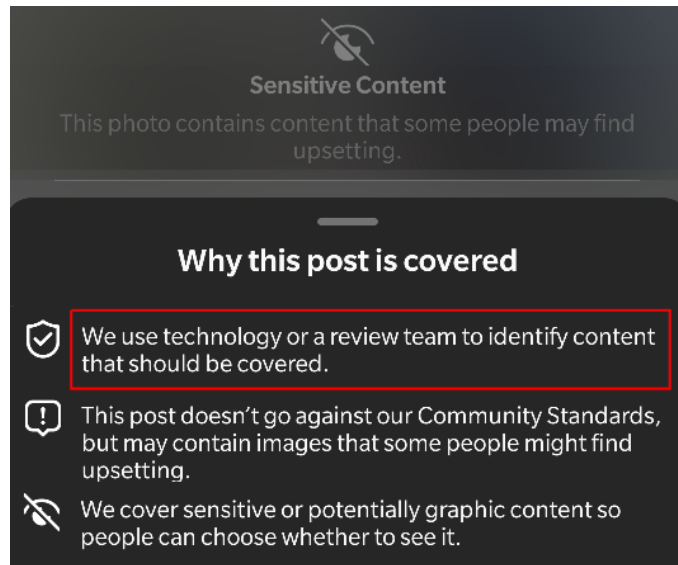


Figure 7.1: Instagram alert of sensitive content.

At the same time, the security of ACM is fundamental to avoid malicious users spreading unauthorized content. For example, if ACM does not detect inappropriate content, human moderators will not control the content, and it will spread on the platform. Thus, only user reports could alarm human moderators, but, in this way, the content might have already harmed those who saw it. ML-based solutions are vulnerable to *evasion attacks*, where the attacker feeds models with crafted samples aiming to affect models' predictions [257]. Similarly, models' decisions can be affected by exploiting pre-processing vulnerabilities [258, 259].

CONTRIBUTIONS In this study, we examine how state-of-the-art ACM can be fooled by simple obfuscation techniques, already in use by social media users, and propose defensive strategies. We summarize our findings as follows:



Figure 7.2: Example of memes with different obfuscations (e.g., typos, letters-shaped objects, hard background).

1. *OSN users are generating malicious posts.* We analyzed 4600 popular Instagram posts from pages and hashtags containing potentially toxic text (e.g., hateful, sexually explicit). We discovered that 44% of them present obfuscations that might undermine ACM decisions. As these posts are reminiscent of captchas (i.e., not understandable by automated mechanisms), we coin this threat as *Captcha Attack (CAPA)*.¹ Figure 7.2 shows an example of a meme we produced using some of these obfuscations (e.g., typos, letters-shaped objects, hard background). This contribution seeks to revisit the classical concept of captcha (*defenses* used in web platforms to distinguish humans from machines), and define a new type of captcha generated by humans to *evade* machine controls.
2. *Different techniques are used to obfuscate content.* The adversarial samples harvested from the wild revealed many ways users obfuscate content, which we organized and formalized in a taxonomy. Our experiments focus on a broad *CAPA* sub-category, consisting of samples adopting textual Captcha Challenges, namely *CC-CAPA*.
3. *CC-CAPA is effective.* We empirically demonstrate the ferocity of *CC-CAPA* by proving that current ACM deployed by top IT companies like Google, Amazon, and Microsoft cannot detect such samples, with an evasion rate equal to 100% (i.e., perfection).
4. *OCR technologies are the ACM weak component.* Our investigation highlights that such failures result from a weak text extraction phase - conducted by Optical Character Recognition (OCR) - an essential step in handling images containing text. Hence, OCR needs to be trained to deal with such obfuscations, but a large amount of data is required for the process. Collecting data is not trivial, since automated detectors are difficult to build.
5. *Defense solutions exist, but they are imperfect.* We propose two *CC-CAPA* identification strategies: supervised and unsupervised. The former is ideal for identifying *CC-CAPA* samples adopting known templates at training time. The latter is a solid methodology to spot unknown *CC-CAPA* templates and, in general, new *CAPA* families. The effectiveness of our approaches is demonstrated through extensive experiments on three OSNs: Pinterest, Twitter, and Yahoo-Flickr.

ORGANIZATION The chapter is organized as follows: Section 7.2 introduces background and related works. Section 7.3 presents *CAPA* taxonomy. Section 7.4 and Section 7.5 show the design of the attack and its results, respectively. Section 7.6 illustrates the detection strategies, and we conclude in Section 7.7.

7.2 BACKGROUND & RELATED WORKS

This section presents theoretical concepts with related works required to understand the rest of the chapter entirely. We discuss security of ML-based applications (Section 7.2.1), moderators in OSN (Section 7.2.2), and captchas (Section 7.2.3).

¹The name of our attack is a quote to Caparezza, an Italian singer famous for his lyrics rich with puns.

7.2.1 SECURITY OF MACHINE LEARNING APPLICATIONS

ML applications like automatic content moderators need to deal with real-world challenges, offering at the same time high performance and attack resiliency. Therefore, when considering the application security, we need to consider all of the components of such pipelines, like preprocessing function, machine learning algorithms, and developing libraries (e.g., PyTorch, Scikit-learn). In general, an adversary’s goal is to control and affect ML application decisions through the definition of *adversarial samples*.

Adversaries can affect ML applications by exploiting ML algorithms. We find different attacks such as the *evasion attack*, where the attacker defines malicious samples that fools a target classifier [257, 260], and the *poisoning attack*, where attackers affects model performances if they have access to the training data [261, 262]. On the opposite, adversaries can further exploit vulnerabilities derived by the ML application pipeline (e.g., libraries bugs, preprocessing functions) [263]. Such attacks are domain and application-related. For example, in the image domain, attackers can exploit image scaling techniques [258]; in the text domain, attackers can leverage non-printable UNICODE characters to affect the text representation [259].

This work focuses on spreading adversarial sentences through images. Given captchas’ deceiving nature, we can categorize our attack as *cross-modal* against *Optical Character Recognition* (OCR) [264]. OCR are tools that extract text from images. Baseline adversarial attacks on OCR use different strategies like noise and watermark addition [265, 266]. These attacks are optimized to fool a target model. In contrast, our attack leverages captchas that are antagonists of OCR by definition. Thus, the proposed attack *CAPA* is not optimized to fool ACM machine learning algorithms but rather to affect earlier stages, such as the text extraction from images using OCR.

7.2.2 MODERATORS IN OSN

Online platforms use human moderators to monitor content shared in their virtual environment and block any malicious content before spreading. However, their efficiency is limited by the many users and interactions a platform presents daily. To overcome this issue, companies started developing automatic tools. From [267], “*the major platforms dream of software that can identify hate speech, porn, or threats more quickly and more fairly than human reviewers, before the offending content is ever seen.*” For example, Facebook uses ML to flag potentially harmful content and remove automatically clear-cut cases, while the rest are processed by human operators [255]. A final contribution to the moderation process is made by OSN users: where human and automatic moderators fail, OSN users may report offensive or harmful content.

Human and automatic content moderators need to deal with multimodal content such as text, image, video, and audio. We can thus find several moderator tools based on the aim and source type. A popular and widely studied application is *hate speech detection*. Furthermore, online platforms are often visited not only by adults but by children as well. Image and video can contain contents that are not appropriate for such a young audience. Examples are *violent* and *sexually explicit* content detectors [268, 269]. While these tools mainly focus on textual contents with NLP-based solutions [270], only recently the attention has moved to multimodal representations (e.g., text inside images). For example, a new popular trend is the hateful meme detection [271, 272, 273], where the ACM combines images and textual information to address the task.

Given the variety of content, ACM need to deal with multiple sources and types of information. In this work, we focus on text and images, which can generate four types of content: (i) textual content like comments, (i i)

image content like photos, (iii) images accompanied by text, like a photo with a caption, and (iv) images containing text like memes. Thus, an ideal ACM should contain DL modules that can work with text and images. The ACM workflow is straightforward in cases like i, ii, and iii. In contrast, for iv the workflow is more complex since the ACM should first extract through an OCR textual information; then, the DL components should process both textual and visual contents. The decision of content being toxic should thus consider both sources. Figure 7.3 shows such a pipeline. While different companies can adopt and develop different ACM, our described pipeline can still faithfully describe their workflows since we do not discuss how to implement specific operations [271, 272, 273].

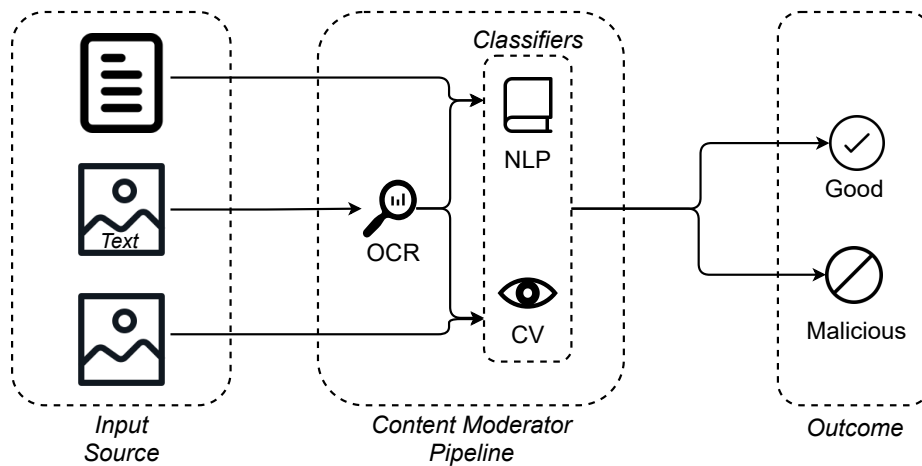


Figure 7.3: Overview of a content moderator in the text and image domains.

ADVERSARIES ON OSN. Generally, users benefit from automatic content moderators since they allow an improvement of platforms’ quality. Nevertheless, popular platforms are populated by malicious users who aim to disrupt such ecosystems. For example, in 2016, a group of users affected *Tay*’s response behavior, a Microsoft chatbot; this tool was shut down after it started spreading hateful tweets [274]. At the same time, ACM has been proven to be vulnerable to adversarial attacks. Yuan et al. focused on *real-world* adversarial techniques on sexually explicit detectors [275]. Here, cyber-criminals used simple image transformations (e.g., rotation, noise addition) to spread porn images on online platforms without being detected. Similarly, Gröndahl et al. presented “all you need is love”, showing that the popular toxic comments detector *Google Perspective*² could be affected by the addition of simple typos and love words [276]. Last, many other malicious activities (e.g., fake news spreading [277]) could be boosted by the possibility of evading automated moderation.

7.2.3 CAPTCHA

A CAPTCHA (Completely Automated Public Turning Test to tell Computers and Humans Apart) is a test to distinguish between humans and computers (e.g., bots, automated users). The first examples appeared in 2000,

²www.perspectiveapi.com

designed by Von Ahn et al. [278], to check whether web requests were coming from humans, improving the security of websites, such as by preventing spam, protecting users' registration, and limiting email address scraping. The first generation of captchas was based on text, altered by rotations, distortions, or wavings, to be hardly readable by a machine (e.g., OCR) but simple for humans. With the advancements in AI technology, text-based captchas began to be solved, with a significant decline in 2014, when Google demonstrated that even the most complicated variants could be easily broken [279]. The security weaknesses related to text-based captchas led the research community to develop new techniques, e.g., based on images, audio, videos, or puzzles [280]. In general, their evolution follows the advancements of technology to break them [281]. Even if the text-based captchas security has been proven to be inefficient, they are still preferred by many users because of familiarity and sense of security and control [282].

The research community put much effort into solving (or breaking) text-based captchas (the type used in our attack). Their robustness has been shown to rely heavily on the difficulty of finding where the character is, i.e., *segmentation*, rather than what character it is, i.e., *recognition* [283]. The breaking methods evolved from algorithmic techniques [284, 285] to machine learning based approaches [286, 287, 279].

To the best of our knowledge, there are no prior works in the literature to detect whether an image is a textual captcha. A possible explanation is that in attacking a website or a web service, the attacker usually knows the phase when a captcha is required and its schema, and for this reason, the research community focused on the breaking path rather than their recognition. Recognizing if an image contains textual captchas could be an effective *CAPA* defense. Thus, we pose a new problem of distinguishing a textual captcha from other real-world objects in images.

7.3 CAPTCHA ATTACK: A TAXONOMY

We now present Captcha Attack (*CAPA*) and a taxonomy of its variants in Section 7.3.1. We discuss the two macro-level of obfuscations: OCR-failures (Section 7.3.2) and classifier-failures (Section 7.3.3). We report statistics of the *CAPA* adoption on Instagram in Section 7.3.4.

7.3.1 CHALLENGES FROM OSN'S USERS: A TAXONOMY

Finding potential adversarial samples in a platform such as Instagram is not trivial given its large amount of posts shared daily. Therefore, we limited our investigation to posts (images containing text) with potentially toxic text. We focused only on posts in English or Italian, which we could fully understand since language can be a barrier to identifying elements such as typos, slang, or double meanings.

We selected four popular hashtags, three well-known English meme pages, and three well-known Italian meme pages, all related to memes or adult content (potentially harmful). We limited our manual inspection to the latest 100 posts for each page to analyze the most recent trends. Given that hashtags convey content from many pages and users, we focused on the latest 1000 English posts without incurring the risk of analyzing old content. A total of 4600 (100×6 pages + 1000×4 hashtags) posts were manually analyzed as a result of this process. Although analyzing only popular hashtags and meme pages could affect, or even limit, the type of observable obfuscations, our inspection is yet focusing on those posts that are already reaching a broad audience. In other words, the obfuscations adopted in these posts are likely i) widespread, ii) effective, and iii) easy to implement. Therefore,

we are collecting the most compelling obfuscation techniques worth categorizing and addressing. By generalizing these categories, minor obfuscation techniques will likely be included as well.

We now need to define *what is a potential adversarial sample*. By considering ACM nature in the multimodal case, as previously discussed in Section 7.2.2, potential threats can be derived by:

1. **OCR-failures** – a wrong text extraction. We thus considered challenges inspired by captchas (Section 7.2.3), such as complex backgrounds or occluders.
2. **Classifiers-failures** – perturbations that can undermine NLP modules, such as typos and leet speech [276].

Using these criteria, we observed that 44% of the 4600 posts present at least one obfuscation.³ We thus decided to investigate the nature of such posts profoundly, and we organized the found obfuscation techniques, resulting in the *CAPA* taxonomy presented in Figure 7.4. The organization follows the security violation level, i.e., at OCR or NLP level. In the remainder of the chapter, we refer to *CAPA* as a generic attack adopting one or more obfuscations.

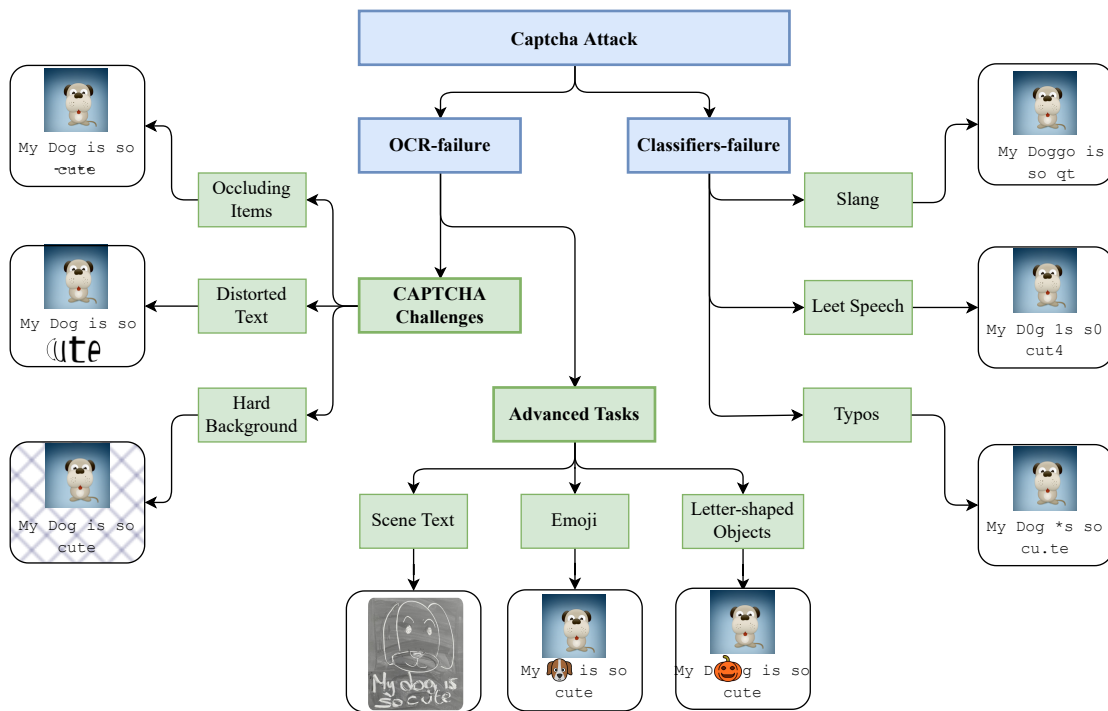


Figure 7.4: Representative samples with obfuscation techniques we identified in online social networks. Blue boxes represent the ACM component that might fail. Green boxes represent different obfuscation techniques.

³We want to underline that, in this phase, we considered not only potential harmful posts but any post published on such pages that might undermine the correct workflow of ACM.

7.3.2 OCR-FAILURE

OCR-level obfuscations aim to disrupt or affect the text extraction phase from images. We identified two sub-family of techniques: *advanced task* for OCR and *CAPTCHA challenges*.

ADVANCED TASKS FOR OCR

With *advanced tasks* we mean a set of applications that differ from the classic document extraction and pose more challenges for OCR. For example, *scene text* recognition is an area that gained popularity in the last few years [288]. This task consists of detecting and extracting text from real-life scenes (e.g., a road sign, a T-shirt). Another exciting challenge is letter-shaped objects, i.e., images whose shapes recall a specific alphabet letter. OCR might not recognize the correct character, resulting in an erroneous extraction. To the best of our knowledge, this task is not yet discussed in the literature. We conclude with the family of *emoji* obfuscations. In Figure 7.4, we show three typical examples of emoji obfuscations. On the top, the text contains an eggplant with a visual double-meaning (i.e., referring to a penis). In the middle, the P-emoji is used with a phonetic deception (i.e., P can be read as ‘pee’). On the bottom, two emoji are combined to represent a sexual action.

CAPTCHA CHALLENGES

CAPTCHA challenges represent obfuscations usually adopted by textual captchas. Such transformations are *hard background*, *distorted text*, and *occluding items*. While we classified these obfuscations as stand-alone, they are usually blended with other obfuscations we presented in the taxonomy.

7.3.3 CLASSIFIER-FAILURES

ML-level obfuscations contain techniques that, while allowing a proper textual extraction, undermine the correct functioning of ML classifiers. These techniques are similar to those presented in [276]: slang, leet speech, and typos. The first category relates to posts that contain slang terms (e.g., wtf → what the f*ck). The second class is the *leet speech*, where some characters are replaced with other visually similar ones, e.g., a * s → @sS. The last class relates to text with *typos* or grammatical mistakes, i.e., images containing misspelled words that, however, can be comprehended by human readers. In the example, we show a meme that contains a sentence with a swear word where the letter ‘i’ is replaced by ‘*’.

7.3.4 STATISTICS FROM THE WILD

Table 7.1 reports the statistics of *CAPA* usage in the wild, supporting our taxonomy. From the 4600 posts analyzed, we discovered that 44% present at least one obfuscation strategy. We can first notice that ‘hard background’ is present in most sources, reaching 77% of posts on one page. In general, this seems to be a trend in new posts, where the text is written on top of a complex background (e.g., real-life scenes). Moreover, we noticed that some techniques (i.e., emoji, leet speech, typos, and occluding items) were mainly used to cover sexually explicit content or swear words.

Takeaway 1: *Users are adopting adversarial techniques in OSN.*

Table 7.1: Percentage of obfuscation techniques observed in different Instagram sources.

Source	CAPTCHA Challenges			Advanced Tasks			Classifier-Failures		
	Occl. Items	Dist. Text	Hard BG	Scene Text	Emoji	LSO	Slang	Leet Speech	Typos
epicfunnypage	3.0	10.0	14.0	9.0	5.0	0.0	26.0	1.0	4.0
6.memes.9	6.0	7.0	33.0	4.0	5.0	0.0	5.0	0.0	5.0
9Gag	0.0	0.0	20.0	1.0	0.0	0.0	7.0	0.0	0.0
partitodisagiato	11.0	0.0	77.0	1.0	23.0	0.0	3.0	15.0	4.0
pastorizianeverdiesreal	0.0	0.0	29.0	2.0	1.0	0.0	1.0	0.0	0.0
alpha_man_real	1.0	1.0	47.0	1.0	0.0	0.0	1.0	0.0	3.0
#naughtymemes	5.3	4.1	23.8	5.2	12.1	0.2	9.7	3.0	5.4
#sexualmemes	3.2	6.1	23.2	4.7	7.3	0.0	15.2	0.4	5.5
#nswfwmemes	7.0	7.1	31.8	0.8	4.2	0.0	14.5	2.1	5.9
#adultmemes	1.3	3.2	18.4	5.9	4.7	0.4	10.0	4.3	5.3

7.4 ATTACK EXECUTION

This section describes the ferocity of *CAPA* in real-life conditions. We start by motivating the attack (Section 7.4.1), followed by the generation procedure of adversarial samples and the resulting dataset in Section 7.4.2 and Section 7.4.3, respectively.

7.4.1 MOTIVATION

Section 7.3 presented Captcha Attack (*CAPA*), i.e., examples of real-life obfuscations we spotted on social networks like Facebook and Instagram. Among these posts, we saw several extremely inappropriate (e.g., sexually explicit, hateful sentences) obfuscated with one or more techniques. Studying ACM behavior in the presence of such ‘adversarial’ samples would highlight ACM weaknesses. Behind these obfuscations, we always find the same rationale: people are trying to create content that can be easily understood by humans but is challenging for machines.

An ideal way to study how ACM would behave with these malicious samples would require collecting a vast number of them. However, we find three major challenges to collect such a dataset: (1) these obfuscations seem novel and a direct consequence of the recent adoption of ACM in OSNs [255], resulting in a limited number of samples; (2) there are many variants or ways to produce an obfuscation, making the problem of limiting samples worse; (3) an automatic tool to detect such posts currently does not exist. In Section 7.6, we discuss in detail a strategy to collect such a dataset.

To address the previously listed issues and to effectively evaluate current real-world ACM robustness, we focus on the automatic generation of *CAPA* by leveraging classic textual captchas containing custom words. Custom

textual captchas can be considered a broad sub-category of the more general class of *CAPA* presented in our taxonomy (Section 7.3, CAPTCHA challenges branch). The adoption of an automatic generation process presents the following advantages. First, given a set of captcha styles, we can generate an arbitrary number of samples. Second, the generated samples represent a simplified version of real-life posts since they do not contain any visual aspect that might affect Computer Vision (CV) classifiers (e.g., racist visual components). Third, classic textual captchas have been widely investigated in the literature, and thus the knowledge acquired so far might help counter *CAPA* in this and all of its forms. Therefore, from now on, through our experiments, we deeply explore the CAPTCHA Challenges branch of *CAPA*, which we call CC-CAPA.

7.4.2 CC-CAPA GENERATION PROCEDURE

This section describes the process of generating custom textual captchas, i.e., CC-CAPA. Given an harmful custom textual sample x , and an automatic content moderator M , we aim to identify a transformation function T such that:

$$\begin{aligned} M(x) &= c_i, \\ M(T(x)) &= c_j, \end{aligned} \tag{7.1}$$

where c_i is the offensive class, and c_j the non offensive one. The function T should satisfy the following properties.

1. *Easy to deploy.* This would open to a broad target of possible adversaries, not only people highly skilled in computer science. CC-CAPA is easy to execute, as shown by the already deployed attacks we presented in Section 7.3.
2. *Target model agnostic.* The transformation should be independent of the target system, i.e., the process T is not mathematically optimized to fool a specific ACM M , but rather any ACM. This would make the attack stronger and more effective to different unknown ACM. As we are going to discuss in the rest of this section, CC-CAPA does not require any information about the target system.
3. *Effective.* The attack should be successful with high confidence. This is desirable since online platforms follow strict policies for inappropriate content sharing, e.g., suspension or account ban. From a theoretical perspective, the usage of captchas should guarantee a high evasion rate. We demonstrate CC-CAPA effectiveness in Section 7.5.

The first CC-CAPA transformations T is the insertion of text in images. This domain-change transformation T_1 represents the first deceptive layer. While analyses on the text and image contents follow standard predictions, the case of text contained in images might represent a gray area since it involves additional operations such as text extraction and the cooperation between NLP and CV machine learning algorithms. If an online platform does not explicitly develop an ACM handling such cases, there is a high chance that malicious content $T_1(x)$ can evade detection mechanisms. We explore this scenario in Section 7.5.2.

If we consider proper implementations of automatic content moderators (see Figure 7.3), setting $T = T_1$ might not be sufficient to guarantee complete attack effectiveness. Thus, the addition of typical manipulation and distortion of classic textual captchas produces images with similar properties to the ones presented in Figure 7.4. For example, we noticed most posts (e.g., memes, Instagram reels) present a hard background. A customizable

textual captcha can be seen as a function composition:

$$T = T_n(\dots(T_2(T_1(x)))) \tag{7.2}$$

where T_1 represents the domain transfer function, while the set $[T_2, \dots, T_n]$ is the combination of image transformations to generate the captcha, and x is the given sentence. As reported in [2], popular transformations can be applied at the background (e.g., solid, complex, noisy), character (e.g., font, size, color, rotation, distortion), and word level (e.g., character overlapping, occluding lines, waving, noise). The notation presented in Equation 7.2 can also describe generic CAPA images. Figure 7.5 shows an overview of the attack execution. The generation process we just described is well-known to the state-of-the-art. While this process does not constitute a part of the novelty of this work, in contrast, the usage of captchas from defense solutions to attack vectors in OSN is, to the best of our knowledge, not explored.

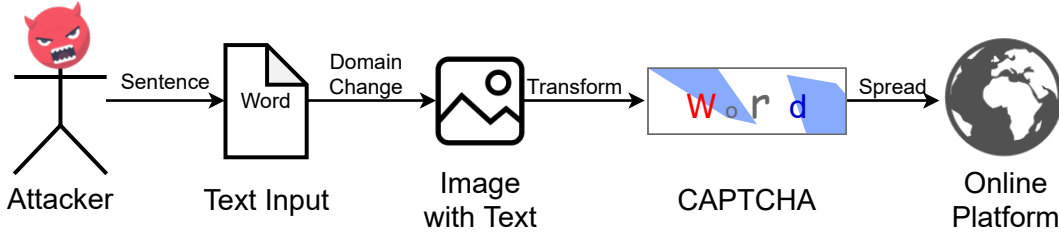


Figure 7.5: Overview of CC-CAPA execution pipeline.

CC-CAPA can exploit the following target ACM weaknesses:

1. *Unimplemented detection case.* The implementation of cross-domain ACM is not trivial and is not widely explored in literature. ACM not implementing such a scenario will miss images with harmful plain text.
2. *Text extraction phase.* If ACM deploys the monitoring of multimodal contents, a pipeline key phase is the text extraction. OCR usually handle this operation. OCR extraction from textual captchas might result in noisy inputs that feed NLP models and thus affecting their predictions.

7.4.3 CC-CAPA DATASET

As introduced in Section 7.2.2, a popular and essential ACM role is the identification of hateful messages on online platforms. Thus, an example of a possible attacker’s goal is to let hateful messages be undetected by ACM. We build our dataset with potentially hateful textual captchas. We retrieve a list of frequent English words associated with hateful sentences from *Hatebase.org* [289], for a total of 1383 samples. From this list, we maintain only those samples that, as stand-alone, should be banned from online platforms. For this purpose, we used *Microsoft Content Moderator*⁴ and *Google Perspective* as our ground truth⁵. These APIs identify the presence of different toxicity aspects. We first applied Microsoft moderation, obtaining 502 toxic words. We refined the list further using Google Perspective, producing a final list of 197 toxic words.

⁴<https://azure.microsoft.com/en-us/services/cognitive-services/content-moderator>

⁵For both APIs, we use 0.5 as a threshold.

In this work, we are interested in understanding if ACM are vulnerable to textual captchas, particularly if different styles of textual captchas affect such target systems in different ways. We thus generate four variants of custom textual captchas. Each style differs in the type and number of transformations applied to the textual captcha. The four classes show different readability difficulties; the more transformations we apply, the more complex the image readability. We now describe the four adopted styles.

1. *Clean*. These are normal white images containing text. No further transformations are applied. Font: FreeMono.
2. *Claptcha*. Python captcha generator available on GitHub.⁶ Complex transformations are applied to the text. Font: FreeMono.
3. *Multicolor*. Python captcha generator available on GitHub.⁷ We modified the library to use an arbitrary text of arbitrary length. Complex transformations are applied to the text. Font: Free family fonts.
4. *Homemade*. Our captcha generator, it aims to be more readable than *Claptcha* and *Multicolor*. Simple transformations are applied to the text. Font: FreeMono.

Figure 7.6 shows examples of attacks, one per class.

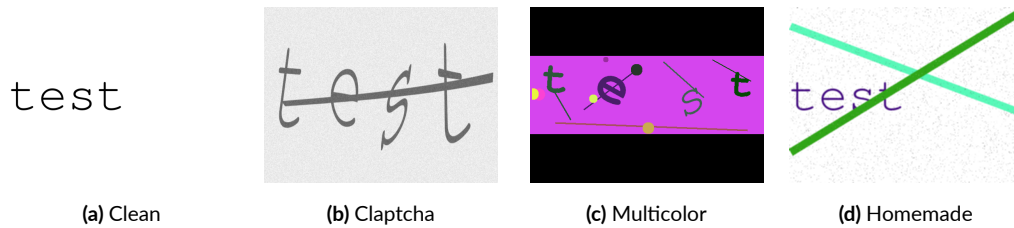


Figure 7.6: Captchas' styles used in the experiments.

The four sample classes are composed by different transformations. For example, *Clean* only uses only domain transfer, while *Claptcha* and *Multicolor* a high number of transformations. Table 7.2 shows the transformations contained in the dataset.

Table 7.2: List of transformations for textual captchas variants.

$T\#$	<i>Transformation</i>	<i>Clean</i>	<i>Claptcha</i>	<i>Multicolor</i>	<i>Homemade</i>
T_1	Domain transfer	✓	✓	✓	✓
T_3	Rotation		✓	✓	
T_4	Distortion		✓		
T_5	Waving		✓		
T_6	Solid background			✓	
T_7	Noisy background		✓		✓
T_8	Different fonts & sizes		✓	✓	
T_9	Different colors			✓	✓
T_{10}	Occluding symbols		✓	✓	✓

⁶<https://github.com/kuszaj/claptcha>

⁷<https://github.com/J-Rios/multicolorcaptcha>

We produce *Clean* samples to verify if current deployed ACM deal with textual captchas, while we produced *Claptcha*, *Multicolor*, and *Homemade* to verify if attackers can affect ACM OCR. About the domain transfer transformation, this operation is easy to implement, from graphic software (e.g., Paint, Photoshop) to standard programming libraries (e.g., matplotlib). We further remark that there exist several online tools aiming to generate customizable textual captchas. The aftermath is that even attackers with low computer skills can produce customisable undetectable textual captchas. This statement is true if we consider that users are already producing CC-CAPA (and more in general, **CAPA**) samples, as discussed in Section 7.3.

Starting from the 502 toxic words identified by Microsoft, and 197 by the addition of Google, we produced two toxic captchas datasets: $|\mathbb{D}_{tox}^m| = 2008$ and $|\mathbb{D}_{tox}^{m+g}| = 788$, where m stands for Microsoft, and g stands for Google. These final datasets are validated through a user study which proves that the generated samples can be read very easily by human beings, thus supporting the idea OSN users would notice the malicious content even if written with captchas. The user study methodology and results are available in Section 7.4.3. We do not make the dataset publicly available since it might be used for attacks in the real world. However, we make it available upon request for researchers to facilitate future investigations in this field.

CC-CAPA READABILITY

Ideally, if CC-CAPA samples are posted on the web, they should be easy to read for humans, otherwise, the whole attack would lose its purpose. Although posting unreadable content would surely evade any ACM, our samples need to have a good balance between low OCR readability and high human readability. While we evaluate the efficacy of OCR in Section 7.5, we assess our captchas' human-readability through a user study. We did not use any harmful words at this stage to not hurt anyone's sensibility.

METHODOLOGY. We recruited 50 participants (27 females, 23 males, age mean 28.6, std 6.1) primarily from our university channels, including students, faculty members, and administrative staff. The participation was voluntary with no monetary compensation. From a list of English verbs⁸, we randomly selected 600 words, generating the corresponding custom textual captchas, 200 for each captcha class (*Claptcha*, *Multicolor*, *Homemade*). Each candidate annotated 50 samples, providing the text they could read along with a difficulty score, from 1 (very easy) to 5 (very difficult), to express how much the participant was sure about the answer, and how immediate the captcha was to solve. The confidence score is crucial to understand if people are likely to read captchas while scrolling social network feeds, or would ignore them because they are considered difficult. Each sample was processed by five participants. During the task, no time restrictions were given.

METRICS. Participants are evaluated with two metrics: accuracy and *Character Error Rate* (CER). In particular, the accuracy evaluates the percentage of samples that were correctly annotated. CER, which is a popular OCR evaluation metric [290], measures the character distance between the annotation and the ground truth (the lower, the closer the two words). The CER score is computed with Fastwer python library.⁹

⁸github.com/aaronbassett/Pass-phrase

⁹github.com/kahne/fastwer

RESULTS. As shown in Figure 7.7, we confirm the high readability of our samples. On average, humans obtained 94.53% and 1.31% of accuracy and CER, respectively. Overall, the task was trivial, with low difficulty scores reported (*Clapcha* = 1.3, *Multicolor* = 1.5, *Homemade* = 1.3). Moreover, we counted the number of samples that have always been successfully (or unsuccessfully) annotated by participants, producing an agreement score. Most samples are always correctly annotated (82.83%), while only 0.5% are always wrongly annotated. We thus expect comparable high readability on the CC-CAPA dataset as well.

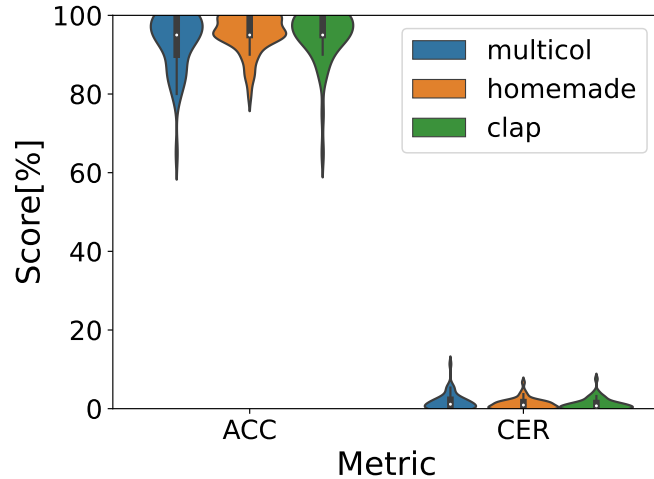


Figure 7.7: User-study performance distribution. We report accuracy (the higher, the better), and CER (the lower, the better).

7.5 ATTACK RESULTS

This section presents the results of our attack in real-life scenarios. We first discuss the attacking scenarios we consider (Section 7.5.1), followed by a presentation of the results of our attack against already deployed ACM (Section 7.5.2) and against ACM following the schema shown in Figure 7.3 (Section 7.5.3).

7.5.1 OVERVIEW

Based on the discussions of ACM deployment done in Section 7.2.2, we aim to verify the following:

1. Do current ACM consider cross-domain samples (e.g., text inside images)? We answer this question in Section 7.5.2 by attacking image moderators with *Clean* samples. We recall that these samples do not contain any transformation and, thus, OCR should successfully extract their text.
2. Are ACM considering cross-domain content vulnerable to CC-CAPA? Section 7.5.3 answers this question by analyzing ACM responses on *Clapcha*, *Multicolor*, and *Homemade* samples.

Tests of the ACM of social networks (e.g., Facebook) are not possible because it would imply the spread of inappropriate and harmful content. Furthermore, we cannot test the attacks to current state-of-the-art solutions

(e.g., hateful memes detection) because, to the best of our knowledge, they all require that the text is successfully extracted through OCR [271, 272, 273]. Moreover, the hateful images presented in our dataset \mathbb{D}_{tox}^{m+g} (see Section 7.4.3) contain only hateful text, while the rest of the background is not harmful. Thus, we opted to test already deployed ACM APIs provided by top IT companies. Note that these services are already adopted from real systems and websites, as can be seen in the APIs presentation pages.

7.5.2 IMAGE MODERATORS

Are current image ACM considering text inside images? To answer such a question, we analyze the scores of only *Clean* samples. An example of *Clean* image is shown in Figure 7.6a. We test the following ACM deployed by top IT companies.

- *Amazon Content Moderation*.¹⁰ The tool aims to classify inappropriate images among different classes, i.e., explicit nudity, suggestive, violent, visually disturbing, rude gestures, drugs, tobacco, alcohol, gambling, and hate symbols.
- *Google Safe Search Detection*.¹¹ The tool returns the likelihood of content containing spoof, medical, violent, or racy content. The likelihood is defined with the following classes: unknown, very unlikely, unlikely, possible, likely, and very likely. We consider content malicious if it is classified as possible, likely, or very likely.
- *Microsoft Content Moderator*.¹² The API identifies if the given image is appropriate for an adult audience (e.g., sexually explicit) or racist.

For each analyzed ACM, we consider a post malicious if it is linked maliciously to at least one of the malicious classes. We measure the attack performance with the *attack success rate* (ASR), defined as the ratio of unsafe content undetected divided by the total number of tests.

We find that all of the services cannot detect offensive text in images, even without obfuscation. *Clean* images reached a success rate of 1 for Amazon and Microsoft, and 0.97 for Google. The 3% images labeled as inappropriate by Google were identified as *spoofed*. This finding suggests that analyzed CV-based ACM do not consider the case of images containing text. We highlight the gravity of such a finding: if an online platform adopts current ACM solutions, attackers could bypass their automatic monitoring systems by *just* putting plain text inside images. Thus, online platforms should manually design defense mechanisms that follow the schema shown in Figure 7.3. We believe that the ACM developers should address this issue since leaving uncovered our proposed scenario (text inside images) weakens their systems' reliability, and expose their users to real threats.

Takeaway 2: *Real world ACM are not considering text within images, which opens severe security threats.*

7.5.3 CROSS-DOMAIN MODERATORS

Section 7.5.2 shows that ACM currently do not consider text inside images for moderation. The natural follow-up question is: “Would adding an OCR module to an ACM effectively ban CC-CAPA?” We thus implement an

¹⁰<https://docs.aws.amazon.com/rekognition/latest/dg/moderation.html>

¹¹cloud.google.com/vision/docs/detecting-safe-search

¹²<https://azure.microsoft.com/services/cognitive-services/content-moderator/>

ACM following the concepts introduced in Section 7.2.2 and preventively assess its robustness. In particular, we defined a pipeline that, given an image, extracts the text using an OCR, and then a textual ACM processes it. In this experiment, we vary the OCR technology while using Microsoft Content Moderator to spot potential harmful extracted sentences. In this stage, since we know a priori that Microsoft considers words $\in \mathbb{D}_{tox}^{m+g}$ as toxic, a misclassification after the text extraction can be due only to an OCR failure. We analyze OCR provided by Amazon¹³, Google¹⁴, Microsoft¹⁵, and the popular free python library Tesseract.¹⁶

We evaluate textual captchas with two metrics: the *attack success rate* (ASR) as defined in Section 7.5.2, and the average *normalized Levenshtein distance* (NLD):

$$NLD(x, x') = \frac{\mathcal{L}(x, x')}{\max(|x|, |x'|)}, \quad (7.3)$$

where x represents the true string in the image, x' the OCR output, \mathcal{L} the Levenshtein distance, and $|x|$ the number of characters in x . The Levenshtein distance measures the number of single-character edits (e.g., addition, modification, deletion) required to make $x = x'$; it is defined between 0, when $x = x'$, and the maximum length between the two strings when they completely differ. The NLD measure defined in Equation 7.3 is thus defined in $[0, 1]$. With the ASR we aim to understand the evasion power of our proposed attack, while with the NLD we aim to understand the number of mistakes that OCR makes.

Figure 7.8 shows the attack performance among the four services. We can first notice that the ASR rate on *Clean* images is very low, meaning that OCR correctly extract the input text. We recall that *Clean* samples do not have any visual transformation (e.g., rotation, complex background), and thus we expect that OCR work properly in such a case. This result suggests that ACM following the schema proposed in Section 7.2.2 are resistant to those attacks that only apply the domain-transfer technique T_1 . Moreover, such a schema presents a valid solution easily adoptable by commercial ACM. Indeed, the results on *Clean* images are much higher compared to the one presented in Section 7.5.2.

On the opposite, the ASR is close to 1.0 for both *Claptcha* and *Multicolor* variants, meaning that offensive textual captchas successfully evaded the ACM in all samples. Our captcha implementation *Homemade* has an average ASR of 0.8, probably due to the fewer number of transformations applied compared to *Claptcha* and *Multicolor*. Similar trends can be found with the NLD measure. The results presented in this section suggest that ACM using the schema proposed in Figure 7.3 are vulnerable to textual captchas with few transformations (e.g., *Homemade* class). Moreover, the more transformations, the higher the attack success rate, reaching the perfect evasion rate for *Claptcha* and *Multicolor*.

Takeaway 3: *Industrial OCR struggle against even the simplest obfuscation techniques.*

¹³<https://aws.amazon.com/it/textract>

¹⁴<https://cloud.google.com/vision/docs/ocr>

¹⁵<https://docs.microsoft.com/en-us/rest/api/cognitiveservices/contentmoderator/imagemoderation/ocrfileinput>

¹⁶<https://pypi.org/project/pytesseract>

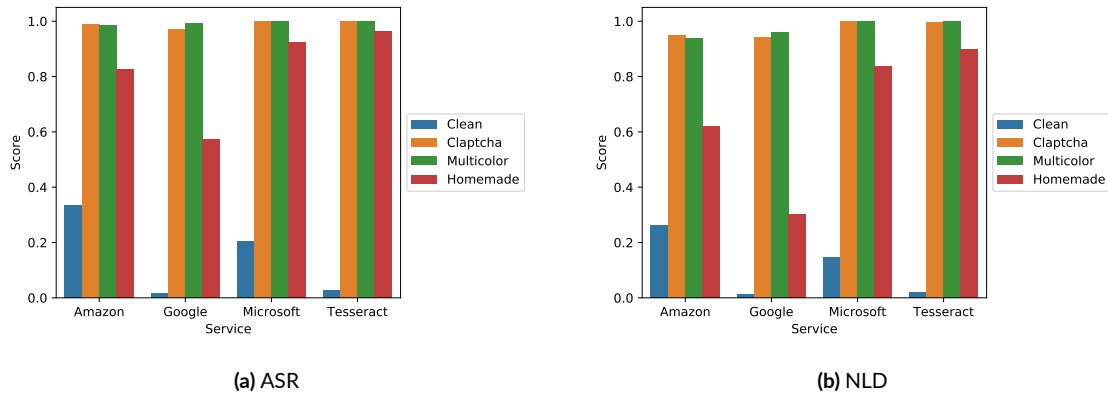


Figure 7.8: Cross-domain evaluation. On the left, the Attack Success Rate (ASR). On the right, the average Normalized Levenshtein Distance (NLD). For both measures, the higher, the more successful the attack.

7.6 CC-CAPA DETECTION STRATEGIES

This section presents CC-CAPA detection strategies. Section 7.6.1 describes possible defense directions. Section 7.6.2 and Section 7.6.3 present, respectively, supervised and unsupervised approaches to tackle the problem. In Section 7.6.4, we discuss general CAPA prevention. Last, we compare our defense strategies to the state-of-the-art in Section 7.6.5.

7.6.1 OVERVIEW

In the previous section, we demonstrated how textual captchas could successfully evade ACM monitoring. Since the generation process of customizable textual captchas is quite naïve, this attack could be massively adopted by many users aiming to spread online messages without being censored. Indeed, social network users are already adopting CC-CAPA (and CAPA in general, see Section 7.3). For simplicity and to demonstrate the proposed attack capabilities, we tested the hate speech evasion task only. Even so, the attack surface is greater than evading hate speech alone, since it can encompass the entire spectrum of text that an online platform could potentially ban (e.g., opinion mining) or analyze (i.e., censorship).

Therefore, it is necessary to discuss potential mitigation to CC-CAPA. We identify three possible directions:

- *Prevention.* Making ACM robust to CC-CAPA is the ideal solution, which involves the deployment of more robust OCR.
- *Detection.* Strategies that identify CC-CAPA samples might help OSN to ban such samples or collect data to train robust OCR.
- *User Reporting.* After CC-CAPA spread in the OSN, users can manually report the presence of toxic content. This data can thus be used to train robust OCR as well.

Although ideal, prevention seems currently unpractical. Indeed CV researchers are currently studying solutions to make OCR robust in complex scenarios [291]. Similarly, user reporting should only compensate for failures of

automatic defensive mechanisms. Indeed, the reported toxic samples have already been spread and (potentially) harmed users. Moreover, we cannot estimate *if* and *when* users will report toxic content.

In this work, we thus focus on the detection scenario. As we previously introduced in Section 7.2.3, captchas are generally a defensive mechanism. So far, the research community has primarily focused on the definition of new captchas or captchas breakers from an attacker’s perspective. The aftermath is that adopting textual captchas as an attack vector creates an uncovered area of cyber security: the captchas identification. Indeed, captcha breakers start from the hypothesis to know a priori if an image is a captcha [2]. OSN can adopt CC-CAPA detectors in three fashions:

1. Detected samples can be directly blocked; this solution might be useful when a platform requires absolute control over its content. Conversely, it might not be ideal in more relaxed scenarios. Indeed, if the platform decides to ban all CC-CAPA samples indiscriminately, users’ (inadvertently) posting benign CC-CAPA samples would feel censored without a reason.
2. Detected samples could be posted, but human operators will revise their goodness.
3. Detected samples could be gathered to create a dataset aiming to build OCR robust to CC-CAPA. We discuss this scenario in Section 7.6.4.

From the literature, we identified two distinct detection approaches:

- *Classification* models (Section 7.6.2). Modeling the problem as a binary task (identify obfuscated images vs. clear ones) might be a simple but effective solution. Indeed, supervised CV classifiers have been demonstrated to be effective in many applications, such as email [292] or image [293] spam detection;
- *Outlier* detectors (Section 7.6.3). We can assume that obfuscated content is the minority of the posts shared in a target platform. Therefore, obfuscated content might be identified as outliers. The identification of malicious activities as anomalies is a well-remarked strategy in cybersecurity and applied in contexts like network intrusion detection systems [294, 295].

7.6.2 SUPERVISED APPROACH: CLASSIFICATION

OVERVIEW. A simple solution is to distinguish CC-CAPA samples from normal OSN posts. In Section 7.6.1 we motivated the need for countermeasures to our proposed attack, and we identified a possible solution: the *textual captcha identification*. We can model such a task as a binary classification problem, where the two classes are *captcha* and *non-captcha*.

DATASET. We now describe the datasets we used to deploy our defense, keeping in mind the following reasons:

1. *The target are OSN.* We must remember that, generally, ACM are deployed on OSN (e.g., Facebook, Twitter, Flickr). It is thus fundamental that the *non-captcha* class captures representative data of the target OSN.
2. *Imbalanced dataset.* Intuitively, we might expect that the majority of the posts in an OSN are not *CAPA* samples. Thus, we expect the dataset to be imbalanced and that the *non-captcha* class contains the majority of the samples.

We built three datasets, starting from three distinct OSN for the *non-captcha* class: Pinterest, Twitter, Yahoo-Flickr. We selected these datasets because images are a substantial portion of their daily content. For the *captcha* class, we used the dataset the authors created in [2], made out of 11 different schemes, each with 700 samples, for a total of 7700 samples. We call this dataset C11. Table 7.3 shows examples of C11 classes along with the applied transformations.

Table 7.3: Captcha schemes used in our experiment coming from [2].

<i>Scheme</i>	<i>Example</i>	<i>Transformations</i>
Alipay		Overlapping, rotation, distortion
Baidu		Occluding lines, overlapping, rotation, distortion, waving, varied font size & color
eBay		Overlapping, distortion, rotation, waving
Google		Overlapping, rotation, distortion, waving, varied font sizes & color
JD		Overlapping, rotation, distortion
Microsoft		Overlapping, solid background, rotation, waving, varied font styles & sizes
Qihu360		Overlapping, rotation, distortion, varied font sizes
Sina		Overlapping, rotation, distortion, waving
Sohu		Overlapping, complex background, occluding lines, rotation, varied font size & color
Weibo		Overlapping, occluding lines, rotation, distortion
Wikipedia		Overlapping, rotation, distortion, waving

Table 7.4 summarizes the statistics of the four sources. We thus created the three datasets: Pinterest + C11, Twitter + C11, and Yahoo-Flickr + C11. Each dataset’s version is split using 70%, 10%, and 20% for the training, validation, and testing partitions. Due to computational limitations, we used just a random subset of Yahoo-Flickr.

MODELS. We utilize two standard techniques for supervised problems in CV: *naïve classifiers* and *fine-tuned classifiers*. In CV, a naïve classifier is usually a Convolutional Neural Network (CNN) made of one or more convolutional layers, followed by one or more linear layers, where the last layer is used for the decision [299]. For

Table 7.4: Datasets' statistics.

<i>Origin</i>	<i>Class</i>	<i>#Samples [k]</i>
C11 [2]	captchas	7.7
Pinterest [296]	non-captchas	70
Twitter [297]	non-captchas	470
Yahoo-Flickr [298]	non-captchas	137

our naïve classifiers, we employed two convolutional layers as CNN backbone, and three linear layers to reach the decision. In particular, we tuned the classifier through grid searching the following parameters on the validation set.

- The CNN backbone: Conv2D with kernel size = {3, 5, 7} and {6, 12, 18} output channels, followed by a second Conv2D with kernel size = {3, 5, 7} and out {6, 12, 18} output channels. Then, the output is flattened and forwarded to the linear layers.
- The linear layers: a first layer with {1000, 10000, 20000} neurons, a second layer with {100, 1000, 2000} neurons, and an output layer with 2 neurons.

Each layer adopts the ReLU as the activation function; moreover, after both Conv2D we apply a MaxPool2D with kernel size = 2. For the fine-tuned models, we use three well-known pre-trained models: Alexnet [300], Resnet18 [301], and VGG [302]. We conducted the experiments in Pytorch. The fine-tuning strategy follows the official Pytorch tutorial [303]. All models are trained using an SGD optimizer (learning rate = 0.001, momentum = 0.9), a cross-entropy loss, and an early stopping mechanism that stops the training if the validation loss is not optimized for five epochs. The models are trained for a maximum of 200 epochs.

RESULTS. We evaluate our models using three standard metrics: F1-score macro, precision, and recall. Table 7.5 summarizes the results. The best naïve classifier had: Conv2D with kernel size = 5 and 6 output channels → Conv2D with kernel size = 5 and out 16 output channels → linear with 10k neurons → linear with 1000 neurons → linear with 2 neurons. In general, all of the classifiers obtain strong classification results close to 100% F1-score in all the scenarios (i.e., Pinterest + C11, Twitter + C11, Yahoo-Flickr + C11). This result implies that companies can easily recognize captchas schemes known at training time with extremely good performances.

Table 7.5: Avg retrieval results of 11 captcha schemes in different OSNs.

<i>Dataset</i>	<i>Pinterest + C11</i>			<i>Twitter + C11</i>			<i>Yahoo-Flickr + C11</i>		
	<i>F1</i>	<i>Prec.</i>	<i>Rec.</i>	<i>F1</i>	<i>Prec.</i>	<i>Rec.</i>	<i>F1</i>	<i>Prec.</i>	<i>Rec.</i>
Naïve	99.8	99.3	99.9	99.5	99.2	99.2	99.9	99.6	100
Alexnet	99.9	99.8	100	99.8	99.5	100	99.9	99.8	100
Resnet18	100	100	100	99.9	99.7	100	99.9	99.7	100
VGG	99.9	99.9	100	99.9	99.6	100	99.9	99.8	100

We further investigated if our supervised technique could spot captcha schemes not seen in training phase (i.e., unknown). Indeed, the definition of new captcha schemes is relatively easy by just varying the number and type of transformations. Moreover, a specific type of transformation can be executed differently; for example, occluding symbols can vary (e.g., lines, segments). As an experiment, we tested our models on \mathbb{D}_{tox}^m (see Section 7.4.3), over the three classes *Claptcha*, *Multicolor*, and *Homemade*, which were unknown at training time. As shown in Table 7.6, our algorithm could detect some *Claptcha* samples, but failed with *Multicolor* and *Homemade*. A possible explanation is that *Claptcha* style is quite similar to some captchas styles presented in our training partition. Thus, if a platform is interested in finding unknown templates, a more generalizable defense solution is needed, which we present in Section 7.6.3.

Table 7.6: Percentage of \mathbb{D}_{tox}^m captchas detected by models trained on data coming from different OSNs.

<i>Dataset</i>	<i>Pinterest</i>			<i>Twitter</i>			<i>Yahoo-Flickr</i>		
	Clap	Multicol	Homemade	Clap	Multicol	Homemade	Clap	Multicol	Homemade
Naïve	11.95	0	0	1.2	0	0	51.79	0	0.2
Alexnet	11.16	0	0	1	0	0	52.19	0	0
Resnet18	0	0	0	0	0	0	0	0	0
VGG	0	0	0	0	0	0	0	0	0

Takeaway 4: Platforms can use supervised techniques to spot samples belonging to a target template, with extremely high accuracy.

7.6.3 UNSUPERVISED APPROACH: OUTLIER DETECTION

OVERVIEW. Supervised techniques guarantee high detection performance on known captcha schemes, while they poorly generalize on unseen styles. Therefore, for unknown styles, we adopt an orthogonal perspective toward our problem. We can assume that CC-CAPA is not (yet) widely exploited on the web platforms, and thus CC-CAPA samples look different from the majority of regular platforms’ posts. Therefore, we adopt an outlier approach, where regular platforms’ posts are inliers and captcha outliers.

DATASET. We use the sources of the same datasets presented in Section 7.6.2 (Pinterest + C11, Twitter + C11, Yahoo-Flickr + C11) but a different training and validation strategy. In particular, the training set contains only samples belonging to the target OSN, while validation and test sets contain both benign and captcha samples. For each OSN, we first take a random subset of 50K samples, and then we split it into training (70%), validation (10%), and testing set (20%). In our investigation, we are willing to understand *how many captcha styles* we should know to build a robust defense. Thus, we vary the number of known captcha styles in the validation set based on the 11 classes available in the C11 dataset. We experiment with different k known styles, $k \in \{2, 4, 6, 8, 10\}$. For each scenario, we repeat the experiment with 5 different styles combinations. The known captcha styles are then randomly split into validation and testing sets, with a 50% of proportion. The unknown captcha styles will belong exclusively to the testing set. Last, we add \mathbb{D}_{tox}^m to the test set (i.e., all the *Claptcha*, *Multicolor*, and *Homemade* captchas).

MODELS. All images are first converted into a 512-dimension embedding representation, using the pre-trained model ResNet-18 [301]. The first component of our defense is a dimensionality reduction module. We opted for the Principal Component Analysis (PCA), on which we vary the number of components: [2, 8, 64, 128]. We then tested the following algorithms: Isolation Forest (IF), Local Outlier Factor (LOF), ECOD [304], and One-Class SVM (OCSVM), using the implementation available in PyOD [305]. For each model, we tune a common hyperparameter, i.e., the contamination level [0.1, 0.05, 0.01]. Moreover, IF are tuned on the number of estimators [16, 32, 64, 128], LOF on the number of neighbors [2, 4, 8, 16], OCSVM on the kernel type [*rbf*, *sigmoid*]. All the models are tuned with a grid-search strategy.

RESULTS. We first visually analyze our data, to better understand possible outcomes. Consider the combination of Pinterest, C11, and \mathbb{D}_{tox}^m datasets. We randomly sampled 2000 items each. From these samples, we first extracted the embedding and obtained a two-dimensional feature space with the combination of a PCA (from 512 to 50 features) and T-SNE (from 50 to 2 features). Figure 7.9 shows the distribution of 2000 Pinterest benign samples among different captcha samples. We can notice that captchas samples have distinct and unique patterns compared to Pinterest ones. However, each captcha style defines its own distinct cluster as well, explaining the poor generalization performance in classification tasks.

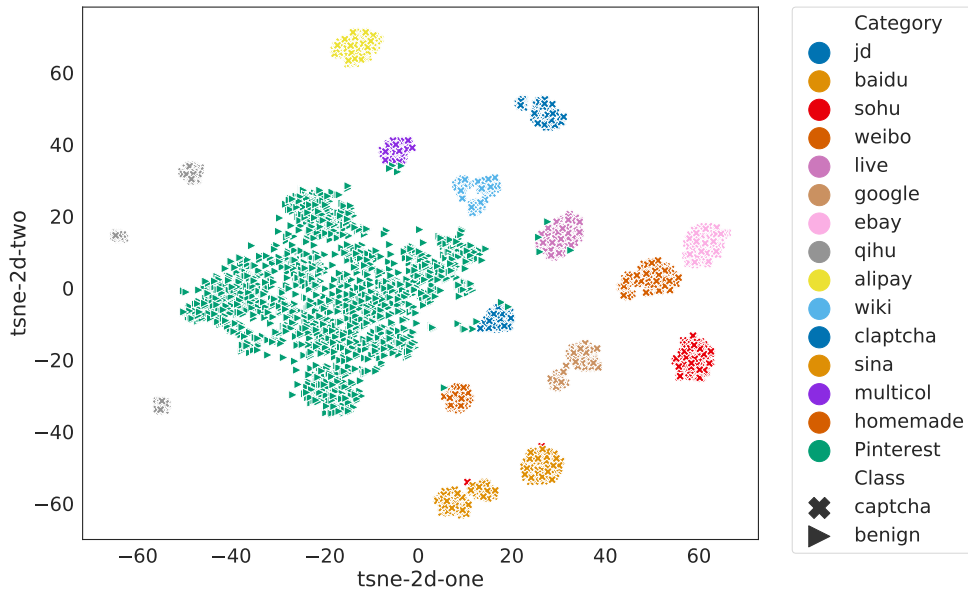


Figure 7.9: T-SNE 2D visualization of 2000 samples benign (Pinterest) and 2000 captchas (C11 and CAPA).

Our next step is to examine the results of three outlier detectors. Figure 7.10 shows the F1-score at testing time at the varying of number of known captcha styles used in the validation set. LOF outperforms Isolation Forest, ECOD, and OCSVM in the three OSN scenarios, reaching, on average, a performance of 80% F1-score. We can also notice that the amount of known styles has a limited impact, finding a performance stabilization starting from 4 styles. Furthermore, we identify consistent trends with both known and unknown captcha styles

recognition. More details are in figures 7.11 and 7.12. Going into more details, LOF labeled as outliers all the captcha classes (both C_{111} and our three classes in \mathbb{D}_{tox}^m), with a minimum of 60% accuracy for *Claptcha*, and up to 96% accuracy for *Homemade*. Thus, through this algorithm, we were able to identify all the captcha schemes through a generalizable solution.

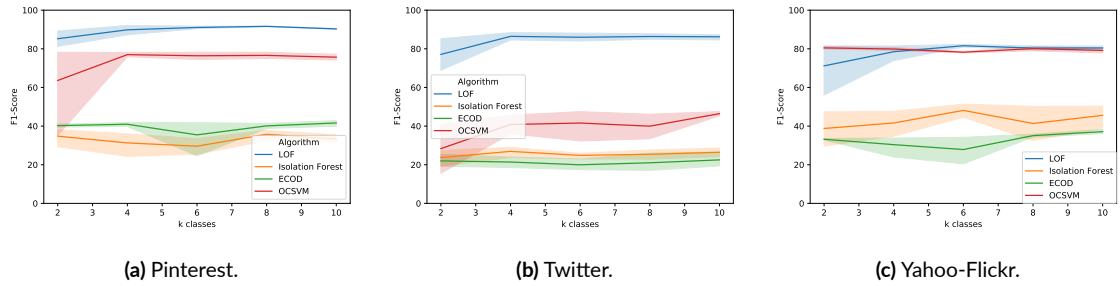


Figure 7.10: F1-score of different Outlier Detection at the varying of the OSN.

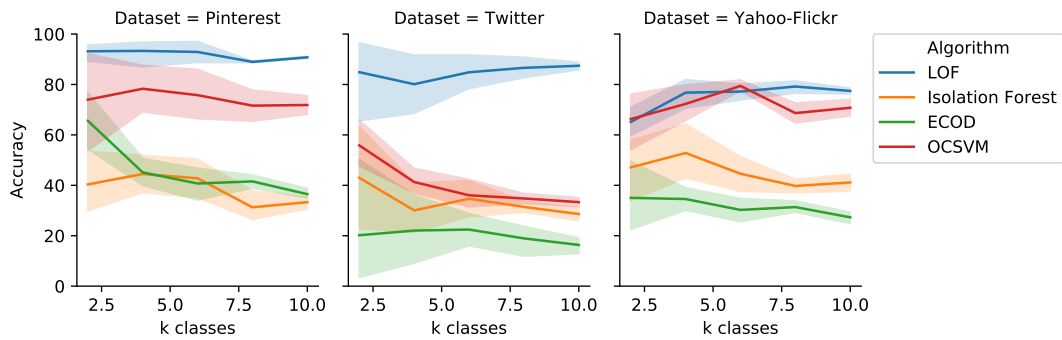


Figure 7.11: Accuracy of different Outlier Detection on known captcha styles at the varying of the OSN.

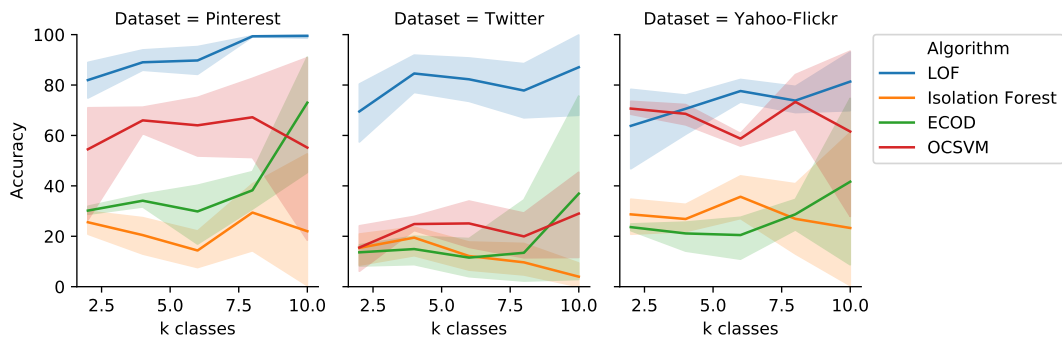


Figure 7.12: Accuracy of different Outlier Detection on unknown captcha styles at the varying of the OSN.

In our experiments, our best algorithm (i.e., LOF) labeled as outliers about 5% of benign posts per dataset, which one could consider as false positives. However, since these posts are considered “unconventional” in the scope of OSN by our algorithm, there is a high chance that such posts could contain some other *CAPA* templates, or more in general, content that could have evaded ACM. Thus, we decided to perform a visual inspection of the three datasets. We manually inspected the 506 Twitter images, 530 Pinterest images, and 479 Yahoo-Flickr images labeled as outliers (i.e., false positives), which confirmed our intuition, i.e., *CAPA* is present in many variants:

- **Pinterest.** Only text without obfuscation: 22%; Text over hard background: 25%; Distorted text: 10%; Occluders: 3%; Natural Scene Text: 7%; Object shape: 1%.
- **Twitter.** Only text without obfuscation: 36%; Text over hard background: 36%; Distorted text: 14%; Occluders: 5%; Natural Scene Text: 11%; Emoji: 3%.
- **Yahoo-Flickr.** Only text without obfuscation: 1%; Text over hard background: 3%; Distorted text: 3%; Occluders: 1%; Natural Scene Text: 7%.

We also found 13 extremely dangerous images on Twitter (8 porn images, 3 gore images, one image inciting racism, and one image advertising drugs). Moreover, we found several posts adopting other *CAPA* templates. In Figure 7.13 we show some examples of false positives we identified through outlier detection techniques, which should require human moderation. In conclusion, through the presented unsupervised approach, we were able to spot all our captcha templates, and even additional *CAPA* templates (see the taxonomy of Figure 7.4), such as all the CAPTCHA Challenges, Scene text, Emoji, and Slang obfuscations.

Takeaway 5: *Unsupervised algorithms offer a solid solutions to spot CC-CAPA. Among the outliers, additional CAPA templates emerged.*



Figure 7.13: Example of false positive found among the outliers that should be moderated.

7.6.4 TOWARD PREVENTING *CAPA*

The presented methods can effectively detect CC-CAPA in the wild, offering a reliable defense strategy. The next step is to shift from *detection* to *prevention* techniques to defend from CC-CAPA, and more in general, from *CAPA*. The main reason we could not focus on prevention techniques, i.e., implementing and training robust OCR, is

that a large dataset currently does not exist. Nevertheless, our supervised and unsupervised approaches can assist in reaching such a goal. We can thus shift the research question from *how to defend against CAPA* to *how to design a large enough dataset with CAPA samples*. Once we answer the latter question, researchers will need to focus on how to design effective OCR. The definition of a CAPA dataset can be divided into two stages:

1. *Identification*, aiming to identify CAPA families that can affect OCR.
2. *Retrieval*, aiming to collect a large number of samples belonging to a specific CAPA family.

Identification. The identification stage can be addressed by *user-reporting*, and *unsupervised techniques*. User reporting a toxic post implies a failure of the ACM, which can be derived from an OCR or toxic classifier failure. If OCR failure occurs, then the OSN identify a potential effective CAPA family. On the opposite, OSN human operators can leverage *unsupervised techniques* (see Section 7.6.3) to identify anomalous posts shared in the platform, seeking potential toxic posts that evaded ACM. Indeed, as shown in Figure 7.9, CAPA families differ from normal OSN posts.

Retrieval. Once human operators identify a new family of CAPA samples, the goal is to collect a large number of similar samples that use such obfuscations, creating a dataset with such a family containing toxic and benign samples. Here, they can use the supervised method we discuss in Section 7.6.2.

7.6.5 COMPARISON WITH STATE OF THE ART

Given the novelty of the attack in OSN, to the best of our knowledge, there are no other defensive detection methodologies to compare our work with. However, in the literature we found a similar topic to ours, i.e., detecting image spam conveyed in emails, sometimes resembling captchas [293]. In the survey of ten years ago that marked the problem of filtering image spam as “solved” [306], three main defense families were presented. The first family involves the usage of OCR extraction combined to text categorization, which is ineffective in our case, as discussed in Section 7.5.3. The second and third families are, respectively, *image classification* and *near-duplicate detection*: both focus on spotting spam images similar to templates known a priori. In particular, they first extract low-level features from a very specific template, and then use machine learning or statistical tests to find images similar to a query image. Through our experiments, we already demonstrated that if the template is “being a captcha”, both methods are ineffective. Indeed, captchas can have very different low-level features based on the adopted obfuscation techniques, and it is impossible to extract pre-determined features for the infinite number of possible obfuscations. Such approaches are better related to our scenario in which we know a captcha schema, and we find images belonging to it. In our experiments, we assessed a near-perfect detection in such a situation. Moreover, our approach adopts Deep Learning (i.e., CNNs) to automatically extract low-level features (compared to manual extraction in prior works), making our methods more scalable and thus superior. For such reasons, we did not conduct further experiments based on prior works.

7.7 CONCLUSIONS AND FUTURE WORKS

Content moderators are essential in our society for the moderation of inappropriate content spread and shared on online platforms. Dangerous content (e.g., hateful words, nudity images) can potentially reach a broad audience,

hurting or harming sensitive people. Online platforms started adopting automatic tools based on deep learning solutions to deal with the massive content volume.

As part of this chapter, we first present Captcha Attack (*CAPA*) and its taxonomy, which is based on the observation of OSN obfuscated posts. We then experimentally demonstrate the ferocity of CC-CAPA, a broad sub-category of *CAPA*, showing that current ACM cannot moderate such samples. We demonstrated how easily an attacker could elude ACM detection by i) changing the domain from text to image and ii) applying captchas schemes. With the first, an attacker can evade those ACM not considering images containing text scenarios. With the latter, an attacker can affect NLP-based tools' performance by exploiting OCR weaknesses. While CC-CAPA, and more in general *CAPA*, is easy to implement and does not require any information about the target model, an ideal countermeasure is far from trivial. Toward this direction, we propose two solid detection approaches that can help to find CC-CAPA (and *CAPA*) samples in the wild.

Our work poses several challenges that might inspire future work. First of all, it is necessary to define a boundary between captchas and non-captchas. Second, for the various categories of the proposed taxonomy, a proper dataset should be collected, to eventually train detectors or sanitizers to help ACM. Last, it would be ideal to build a model that works against all the obfuscation variants described in taxonomy (e.g., emoji, leet speech).

8

Social Honey-pot for Humans: Luring People Through Self-managed Instagram Pages

In recent years, Social Network Analysis (SNA) has emerged as a powerful tool for studying society. The large amount of relational data produced by Online Social Networks (OSN) has greatly accelerated studies in many fields, including modern sociology [307], biology [308], communication studies [309], and political science [310]. SNA success can be attributed to the exponential growth and popularity OSN faced [311], with major OSN like Facebook and Instagram (IG) having billions of users [312, 313]. Researchers developed a variety of tools for SNA [314]; however, elaborating the quintillion bytes of data generated every day [315] is far from trivial [316]. The computational limitations compel scientists to conduct studies on sub-samples of the population, often introducing bias and reducing the quality of the results [317]. Furthermore, the reliability of data is hindered by adversarial activities perpetuated over OSN [318], such as the creation of fake profiles [230], crowdfunding campaigns as seen in Chapter 6, or spamming [319, 320, 321].

Back in the years, cybersecurity researchers proposed an innovative approach to overcome the computational limitation in finding malicious activity in OSN (e.g., spamming), by proposing social honeypots [322, 323, 324]: profiles or pages created ad-hoc to lure adversarial users, analyze their characteristics and behavior, and develop appropriate countermeasures. Thus, their search paradigm in OSN shifted from “look for a needle in the haystack” (i.e., searching for spammers among billions of legit users) to “the finer the bait, the shorter the wait” (i.e., let spammers come to you).

MOTIVATION The high results achieved by such techniques inspired us to generalize the approach, gathering in a *single place any target users* we wish to study. Such a framework’s uses are various, from the academic to

the industrial world. First, *profilation* or *marketing* toward target topics: IG itself provides page owners to know aggregated statistics (e.g., demographic) of their followers and users that generate engagement.¹ Second, *social cybersecurity analytics*: researchers or police might deploy social honeypots on sensitive themes to attract and analyze the behavior of people who engage with them. Examples of themes are fake news and extremism (e.g., terrorism). Although our “general” social honeypot may be used either benignly (e.g., to find misinformers) or maliciously (e.g., to find vulnerable people to scam), in this chapter, we only aim to examine the feasibility of such a tool, and its effectiveness. Moreover, we investigate whether this technique can be fully automated, limiting the significant effort of creating a popular IG page [326]. We focus on IG given its broad audience and popularity. Furthermore, IG is the most used social network for marketing purposes, with nearly 70 percent of brands using IG influencers (even virtual, as shown in Chapter 2) for their marketing campaigns [327].

CONTRIBUTION In this work, we present an automated framework to attract and collect legitimate people in social honeypots. To this aim, we developed several strategies to understand and propose guidelines for building effective social honeypots. Such strategies consider both *how to generate content automatically* (from simple to advanced techniques), and *how to engage with the OSN* (from naive to complex interactions). In detail, we deployed 21 honeypots and maintained them for nine weeks. Our four content generation strategies involve state-of-the-art Deep Learning techniques, and we actively engage with the network following three engagement plans.

The main contributions of this chapter can be summarized as follows:

- We define a novel concept of Social HoneyPot, i.e., a flexible tool to gather *real people* on IG interested in a target topic, in contrast to previous studies focusing on malicious users or bots;
- We propose four automatic content generation strategies and three engagement plans to build self-maintained IG pages;
- We demonstrate the quality of our proposal by analyzing our 21 IG social honeypots after a nine weeks period.

OUTLINE We begin our work discussing related works (Section 8.1). Then, we present our methodology and implementation in Section 8.2 and Section 8.3. In Section 8.4, we evaluate the effectiveness of our honeypots, while Section 8.5 presents social analyses. We discuss the use cases of our approach and its challenges in Section 8.6 and conclude the chapter in Section 8.7.

8.1 RELATED WORKS

HONEYPOT

Honeypots are decoy systems that are designed to lure potential attackers away from critical systems [328]. Keeping attackers in the honeypot long enough allows to collect information about their activities and respond ap-

¹Instagram API provides to the owner aggregated statistics of followers (gender, age, countries) when their page reaches 100 followers [325].

appropriately to the attack. Since legit users have no valid reason to interact with honeypots, any attempt to communicate with them will probably be an attack. Server-side honeypots are mainly implemented to understand network and web attacks [329], to collect malware and malicious requests [330], or to build network intrusion detection systems [331]. Conversely, client-side honeypots serve primarily as a detection tool for compromised (web) servers [332, 333].

SOCIAL HONEYPOT

Today, honeypots are not limited to fare against network attacks. Social honeypots aim to lure users or bots involved in illegal or malicious activities perpetuated on Online Social Networks (OSN). Most of the literature focused on detecting spamming activity, i.e., unsolicited messages sent for purposes such as advertising, phishing, or sharing undesired content [324]. The first social honeypot was deployed by Webb et al. [322] on MySpace. They developed multiple identical honeypots operated in several geographical areas to characterize spammers' behavior, defining five categories of spammers. Such work was extended to Twitter by Lee et al. in 2010 [323], identifying five more spammers' categories, and proposing an automatic tool to distinguish between spammers and legit users. Stringhini et al. [324] proposed a similar work on Facebook, using fake profiles as social honeypots. Similarly to previous works, these profiles were passive, i.e., they just accepted incoming friend requests. Their analysis showed that most spam bots follow identifiable patterns, and only a few of them act stealthily. De Cristofaro et al. [334] investigated Facebook Like Farms using social honeypots, i.e., blank Facebook pages. In their work, they leveraged demographic, temporal, and social characteristics of likers to distinguish between genuine and fake engagement. The first "active" social honeypot was developed on Twitter by Lee et al. [335], tempting, profiling, and filtering content polluters in social media. These social honeypots were designed to not interfere with legitimate users' activities, and learned patterns to discriminate polluters and legit profiles effectively. 60 honeypots online for seven months gathered 36'000 interactions. More active social honeypots were designed by Yang et al. [336]), to provide guidelines for building effective social honeypots for spammers. 96 honeypots online for five months attracted 1512 accounts. Last, pseudo-honeypots were proposed by Zhang et al. [337], which leveraged already popular Twitter users to attract spammers efficiently. They run 1000 honeypots for three weeks, reaching approximately 54'000 spammers.

DIFFERENCES WITH PREVIOUS WORK

To date, social honeypots have been mainly adopted to detect spammers or bot activities. The majority of research focused on Twitter, and only a few works used other social networks like Facebook. There are several reasons behind this trend. First, spamming is one of the most widespread malicious activities on social networks because it can lead to other more dangerous activities. Second, Twitter APIs and policies facilitate data collection, and there are widely adopted Twitter datasets that can be used for further analysis. To the best of our knowledge, there are no works that utilize social honeypots on Instagram, perhaps because it is difficult to distribute, maintain and record honeypots' activities on this social network. Moreover, our goal is to attract *legit users* rather than spammers, which is radically different from what was done insofar. Indeed, many analyses could be easier to conduct by gathering people in one place (e.g., an IG page). For instance, a honeypot could deal with peculiar topics to simplify community detection [9], could advertise a product to grasp consumer reactions [338], understand political

views [339], analyze and contrast misinformation [340], conspiracies [341], and in general, carry out any Social Network Analytics task [342]. Last, owners of IG pages can see the demographic information of their followers (inaccessible otherwise), having extremely helpful (or dangerous) information for further social or marketing analyses [343].

8.2 METHODOLOGY

8.2.1 OVERVIEW & MOTIVATION

The purpose of our social honeypots is to attract people interested in a target topic. The methodology described in this section is intended for Instagram (IG) pages, but it can be extended to any generic social network (e.g., Facebook) with minor adjustments. We define the social honeypot as a combination of three distinct components: (i) the honeypot *topic* that defines the theme of the IG page (Section 8.2.2); (ii) the *generation strategy* for creating posts related to a target topic (Section 8.2.3); (iii) the *engagement plan* that describes how the honeypot will engage the rest of the social network (Section 8.2.4). Figure 8.1 depicts the social honeypot pipeline.

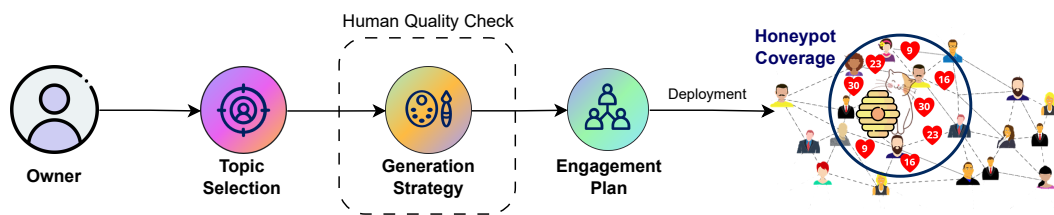


Figure 8.1: Pipeline overview to create a social honeypot. After the owner decides on the topic, generation strategy, and engagement plan, the honeypot automatically generates posts to interact with the social network. After the post is automatically generated, the owner can approve it or request a new one to meet the desired quality.

Our study examines different types of honeypots with a variety of topics, generation strategies, and engagement plans, outlined in the rest of this section. Our experiments aim to answer the following research questions:

- RQ1. Can self-managed social honeypots generate engagement on Instagram?
- RQ2. How do the topic selection, post generation strategy, and engagement plan affect the success of a social honeypot?
- RQ3. How much effort (computation and costs) is required to build an effective social honeypot?

The remainder of the section describes the strategies we adopt in our investigation, along with technical implementation details.

8.2.2 TOPIC SELECTION

Building a honeypot begins with selecting the topic of its posts. Such a choice will impact the type of users we will attract. The topic’s nature might vary, from hobbies and passions like sports and music to sensitive issues like

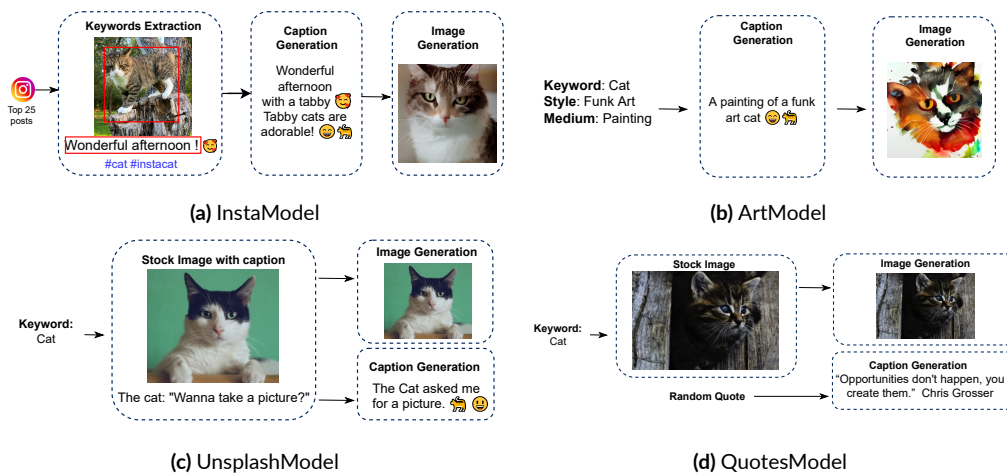


Figure 8.2: Overview of Post Generation strategies.

political views and conspiracies. As an example, if we wish to promote a new product of a particular brand, the topic might be the type of product we intend to promote. Alternatively, if we intend to develop a tool for spam detection, we should choose a topic that is interesting to spammers. This will ensure that they will be attracted to the honeypot’s content. We can even design honeypots with generic topics that can be used for marketing profiling or social studies. In conclusion, the topic should be chosen in accordance with the honeypot’s ultimate purpose.

8.2.3 POST GENERATION STRATEGIES

The generative process aims to create posts pertaining to the honeypot topic. A two-part artifact is produced: the *visual* component of the post (i.e., the image), and its *caption*. We propose four distinct methods to generate posts, each with its own characteristics and algorithms. For ethical reasons, we excluded techniques that might violate the author’s copyright (e.g., re-posting). However, unscrupulous honeypot creators could conveniently use these strategies. In this section, we provide the strategies high-level view to serve as a framework. For technical implementation details (e.g., the actual models we used), please refer to Appendix 8.7.1. Since this stage involves deep generative models that might produce artifacts affecting the post quality, the owner can approve a post or request a new one with negligible effort.

INSTAMODEL

InstaModel is a generative schema that leverages machine learning techniques to generate both images and captions. Figure 8.2.a shows its overview. The schema begins by retrieving one starting post among the 25 most popular IG posts for a popular hashtag related to the honeypot topic.² Next, the pipeline performs, in order,

²Starting from the main topic hashtags (i.e., #cat, #food, #car), we daily create the set of hashtags contained in the top 25 posts, from which we draw the hashtag to retrieve the starting post.

caption generation and image generation steps.

- *Caption Generation.* The algorithm uses an *Object Detector* tool³ to extract the relevant elements of the starting post’s image. In the absence of meaningful information (e.g., is a meme or unrelated to the topic)⁴, we discard that image. When this occurs, the algorithm restarts and uses another sample from the top 25. If the image is kept, the algorithm uses the list of resulting elements (i.e., keywords) to generate a sentence, leveraging a *keyword-to-text* algorithm. Note that we discard from the keywords list those elements with very low probability. The output of the *keyword-to-text* phase (i.e., the new caption) is further refined to align with IG captions, for example, by adding emojis and hashtags, as presented in Section 8.2.4.
- *Image Generation.* The caption generated in the previous step serves as input to produce the post image. To achieve this goal, we use *text-to-image* models, i.e., algorithms that produce more images from a single input. An operator would choose the most appropriate option or a random option in such a case. We remark that *InstaModel* severely adopts generative models. Indeed, we used state-of-the-art computer vision, NLP, and image generation models for object detection, text generation, and image generation, respectively.

ARTMODEL

ArtModel leverages the ability of novel *text-to-image* generative models (e.g., DALL-E) to interpret artistic keywords as inputs. Figure 8.2b shows the overview of the model. Similarly to *InstaModel*, the process starts by generating a caption, and, subsequently, the image.

- *Caption Generation.* Differently from *InstaModel*, the input to generate the caption does not come from other IG posts. Instead, we randomly select the target keyword (e.g., cat), the artistic style of the picture (e.g., Picasso, impressionism), and a medium (e.g., painting, sketch). We create a single sentence by filling pre-defined templates with such three keywords, and add emojis and hashtags as for *InstaModel*.
- *Image Generation.* Similar to *InstaModel*, the caption (without emojis and hashtags) serves as input for a *text-to-image* model, which generates the final image.

UNSPLASHMODEL

This algorithm employs DL models only to generate the caption. In opposition to *InstaModel* and *ArtModel*, *UnsplashModel* starts from the image generation, and then generates the caption (Figure 8.2c).

- *Image Generation.* The image is randomly selected by a stock images website – in this case, Unsplash⁵. The search is based on a randomly selected keyword that reflects the target topic, from a list defined by the owner.
- *Caption Generation.* Unsplash images are usually accompanied by captions free of license. We further refine the caption with a *rephrase* model, and add emojis and hashtags as for the previous models.

³Object detectors are Computer Vision-based tools that identify objects composing a given scene. Each object is accompanied by a probability score.

⁴We discard those images that do not contain at least a topic-related element with a high probability.

⁵<https://unsplash.com/>

QUOTESMODEL

Last, we present *QuotesModel*, a variant of *UnsplashModel*, presented in Figure 8.2d. The objective of this strategy is to determine whether AI-based techniques are necessary to generate attractive IG posts. Therefore, this model does not involve the use of artificial intelligence to create captions and images. In addition, using quotes to caption photos is a diffused strategy [344].

- *Image Generation*. The image generation process is the same as *UnsplashModel*, involving stock images.
- *Caption Generation*. Captions are randomly selected by popular quotes from famous people (e.g., ‘Stay hungry, stay foolish’ – Steve Jobs). Quotes are retrieved from a pool with 1665 quotes [345].

8.2.4 ENGAGEMENT PLANS

Lastly, the engagement plan defines how the social honeypot interacts with the rest of the social network (e.g., other users or pages). We defined three plans, varying in effort required to maintain interactions, and whether paid strategies are involved:

- *PLAN 0*: low interactions and no paid strategies;
- *PLAN 1*: high interactions and no paid strategies;
- *PLAN 2*: high interactions and paid strategies.

PLAN 0

The plan does not involve automatic interactions with the rest of the social network. At most, the owner replies to comments left under the honeypot’s posts. The plan uses the well-known *Call To Actions* (CTA) [346] in the posts. Such a strategy consists in creating captions that stimulate users’ engagement (e.g., liking, commenting, sharing the post). Examples are captions containing simple questions (e.g., ‘How was your day?’), polls and quizzes (e.g., ‘What should I post next?’), or exhorting users to share their opinions (e.g., ‘What do you think about it?’). Following the caption best strategies for IG posts [347], we added 15 random hashtags related to our topic, 8 with broad coverage and 7 with medium-low coverage. More details about the hashtags selections in Appendix 8.7.1. In this plan, paid strategies are not involved.

PLAN 1

The plan is a variant of *PLAN 0* with explicit social networking interactions. We call these actions *spamming*. The spamming consists of automatically leaving likes and comments on the top 25 posts related to the topic (as described in *InstaModel*). Comments resemble legit users (e.g., ‘So pretty!’) and not spammers (e.g., ‘Follow my page!’), and were randomly picked from a list we manually created by observing comments usually left under popular posts. The goal of such activities is to generate engagement with the owner of popular posts, hoping to redirect this stream to the honeypot. When a user follows us, we follow back with a probability of 0.5, increasing the page’s number of followings, resembling a legit page. During our experiments, we also adopted a more

aggressive (and effective) spamming strategy called *Follow & Unfollow* (F&U) [348], consisting in randomly following users, often causing a follow back, and then remove the following after a couple of days. To not be labeled as spammers, we constantly respected the balance $\# \text{ following} < \# \text{ followers}$. In this plan, paid strategies are not involved.

PLAN 2

This plan increments *PLAN 1* with two paid strategies.

BUYING FOLLOWERS When we create a honeypot, we buy N followers. In theory, highly followed pages might encourage users to engage more, and gain visibility from IG algorithm [51]. Therefore, we aim to understand if an initial boost of followers can advantage honeypots. Such followers will be discarded during our analyses. We set $N = 100$, and we buy passive followers only.⁶

CONTENT SPONSORING IG allows posts' sponsoring for a certain amount of time. The target population can be automatically defined by IG, or chosen by the owner w.r.t. age, location, and interests. Since we are interested in studying the population attracted by our content, rather than attracting a specific category of users, we let IG decide our audience, directly exploiting its algorithms to make our honeypots successful.

8.3 IMPLEMENTATION

8.3.1 TOPIC SELECTION

We investigate the honeypots' effectiveness over three distinct topics: *food*, *cat*, and *car*. We selected such topics to account for different audience sizes, measured by coverage levels. Coverage is a metric that counts the total number of posts per hashtag or, in other words, the total number of posts that contain that hashtag in their captions. This information is available on IG by just browsing the hashtag. More in detail, we selected: **Food** (high coverage, #food counts 493 million posts), **Cat** (medium coverage, #cat counts 270 million posts), and **Car** (low coverage, #car counts 93 million posts). We chose these topics, and not more sensitive ones, mainly for ethical reasons. Indeed, we did not want to boost phenomena like misinformation or conspiracies through our posts, nor identify people involved in these themes. However, we designed our methodology to be as general as possible, and adaptable to any topic with little effort.

8.3.2 TESTBED

We deployed 21 honeypots on Instagram, seven for each selected topic (i.e., food, cat, and car), that we maintained for a total of nine weeks. Within each topic, we adopt all post generation strategies and engagement plans. For the

⁶Passive followers only follow the page, but they do not engage further.

post generation strategies, three honeypots use both InstaModel and ArtModel, three honeypots use UnsplashModel and QuotesModel, and one honeypot combines the four. Such division is based on the image generation strategy, i.e., if images are generated with or without Deep Learning algorithms. All posts were manually checked before uploading them on Instagram to prevent the diffusion of harmful or low-quality content. This was especially necessary for AI-generated content, whose low quality might have invalidated a fair comparison with non-AI content.⁷ Similarly, for the engagement plan, two honeypots adopt PLAN 0, two PLAN 1, and three PLAN 2. Table 8.1 summarizes the 21 honeypots settings. Given the nature of our post generation strategies and engagement plans, we set as baselines the honeypots involving *UnsplashModel* + *QuotesModel* as generation strategy and *PLAN 0* as engagement plan (h1, h8, h15). Indeed, these honeypots are the simplest ones, requiring almost no effort from the owner. Setting baselines is useful to appreciate the results of more complex methods, given that there are currently no baselines in the literature.

By following the most common guidelines [351, 352], each honeypot was designed to publish two posts per day, with at least 8 hours apart from each other.

During the nine weeks of experiments, we varied PLAN 1 and PLAN 2. In particular, we started PLAN 1 with spamming only, and PLAN 2 with buying followers. During the last week, both plans adopted more aggressive strategies, specifically, PLAN 1 applied F&U techniques, while PLAN 2 sponsored the two most-popular honeypot posts for one week, paying € 2/day for each post. For our analyses, we collected the following information:

- Total number of followers per day;
- Total number of likes per post;
- Total number of comments per post.

Moreover, IG API provided the gender, age, and geographical locations of the audience when applicable, as explained in Section 8.5.3.

IMPLEMENTATION MODELS In Section 8.2 we presented a general framework to create social honeypots. In our implementations, we employed deep learning state-of-the-art models in several steps. To extract keywords in *InstaModel* we adopted InceptionV3 [353] as object detector, pre-trained on ImageNet [354] with 1000 classes. From the original caption, we extracted nouns and adjectives through NLTK python library⁸. As *keyword-to-text* algorithm, we adopted Keytext [355] based on T5 model [356]; while for *text-to-image* processes we opted for Dall-E Mini [357]. Finally, in *UnsplashModel*, the rephrase task was performed using the Pegasus model [358].

8.4 HONEYPOTS EVALUATION

8.4.1 OVERALL PERFORMANCE

The first research question *RQ1* is whether social honeypots are capable of generating engagement. After nine weeks of execution, our 21 social honeypots gained: 753 followers (avg 35.86 per honeypot), 5387 comments

⁷The effort for the honeypot manager is limited to a quick approval, which could not be necessary with more advanced state-of-the-art models, e.g., DALL-E 2 [349] or ChatGPT [350].

⁸<https://www.nltk.org/>

Table 8.1: Honeypots deployed.

ID	Post Generation Strategy	Engagement Plan
<i>food</i>		
h1 (baseline)	UnsplashModel + QuotesModel	PLAN 0
h2	UnsplashModel + QuotesModel	PLAN 1
h3	UnsplashModel + QuotesModel	PLAN 2
h4	InstaModel + ArtModel	PLAN 0
h5	InstaModel + ArtModel	PLAN 1
h6	InstaModel + ArtModel	PLAN 2
h7	All Models	PLAN 2
<i>cat</i>		
h8 (baseline)	UnsplashModel + QuotesModel	PLAN 0
h9	UnsplashModel + QuotesModel	PLAN 1
h10	UnsplashModel + QuotesModel	PLAN 2
h11	InstaModel + ArtModel	PLAN 0
h12	InstaModel + ArtModel	PLAN 1
h13	InstaModel + ArtModel	PLAN 2
h14	All Models	PLAN 2
<i>car</i>		
h15 (baseline)	UnsplashModel + QuotesModel	PLAN 0
h16	UnsplashModel + QuotesModel	PLAN 1
h17	UnsplashModel + QuotesModel	PLAN 2
h18	InstaModel + ArtModel	PLAN 0
h19	InstaModel + ArtModel	PLAN 1
h20	InstaModel + ArtModel	PLAN 2
h21	All Models	PLAN 2

(avg 2.01 per post), and 15730 likes (avg 5.94 per post). More in detail, Table 8.2 (left side) shows the overall engagement performance at the varying of our three variables, i.e., topic, generation strategy, and engagement plan. The reader might notice that not only our honeypots *can* generate engagement, answering positively to the *RQ1*, but that also topic, generation strategy, and engagement plan have different impacts to the outcomes. For instance, *cat* honeypots tend to have higher followers and likes, while *car* ones generate more comments. Similarly, *non-AI* generation methods tend to have higher likes, as well as *PLAN 1*. We investigate the effect of different combinations later in this section.

8.4.2 HONEYPOT TRENDS ANALYSIS

Social honeypots can generate engagement, but we are further interested in understanding trends of such performance: *is honeypots' engagement growing over time?* A honeypot with a positive trend will likely result in a higher future attraction. On the opposite, a stationary trend implies limited opportunities to improve.

The qualitative analysis reported in Figure 8.3 motivates the trend investigation. The figure presents the average number of Likes per post gained by our honeypots over time, grouped by engagement plan. In general, *PLAN 1* honeypots tend to attract more likes as they grow, followed by *PLAN 2* and *PLAN 0*, in order. In particular, a constantly increasing number of likes is shown by honeypots with *PLAN 1*, especially for food-related pages: starting from an average of ~ 5 likes per post (week 1st) to ~ 12.5 likes per post (week 9th). We evaluate the presence of stationary trends by adopting the *Augmented Dickey-Fuller test* (ADF) [359]. In this statistical test,

Table 8.2: Honeypots overall performance. On the left side, we report the average (and std) engagement generated by the honeypots. On the right, we report the number of honeypots with a non-stationary trend. The results are reported based on the topic, generation strategy, and engagement plan.

	<i>Average Engagement</i>			<i>Engagement Trend</i>		
	#Followers	#Comments	#Likes	#Followers	#Comments	#Likes
<i>topic</i>						
food	38.5±33.7	216.4±18.5	698.4±139.7	6/7	3/7	7/7
cat	47.4±17.5	182.1±23.5	923.1±214.8	6/7	2/7	4/7
car	21.9±9.7	371.0±26.2	625.6±96.6	7/7	3/7	6/7
<i>generation strategy</i>						
AI	37.9±30.9	248.4±94.6	654.2±138.3	7/9	4/9	6/9
non-AI	32.7±21.3	264.2±90.6	842.5±235.2	9/9	3/9	8/9
Mixed	39.3±7.9	257.7±80.0	753.0±125.9	3/3	1/3	3/3
<i>engagement plan</i>						
PLAN 0	11.5±8.4	266.0±105.8	641.3±210.7	4/6	4/6	5/6
PLAN 1	60.0±25.2	254.2±94.3	835.2±210.7	6/6	2/6	4/6
PLAN 2	36.0±14.0	251.8±79.1	763.4±206.1	9/9	2/9	8/9

the null hypothesis H_0 suggests, if rejected, the presence of a non-stationary time series. On the opposite, the alternative hypothesis H_1 suggests, if rejected, the presence of a stationary time series. We conducted the statistical test for each honeypot and the three engagement metrics: #Followers, #Likes, and #Comments. A p -value > 0.05 is used as a threshold to understand if we fail to reject H_0 . Table 8.2 (right side) reports the result of the analysis. The number of Followers and Likes is non-stationary in 19 and 17 cases out of 21, respectively. Conversely, the number of comments per post is stationary in most of the honeypots. This outcome suggests that engagement in terms of likes and followers varies over time (positively or negatively), while the number of comments is generally constant. As shown in Figure 8.3, and given the final number of followers higher than 0 (i.e., at creation time), we can conclude that our honeypots present, in general, a growing engagement trend.

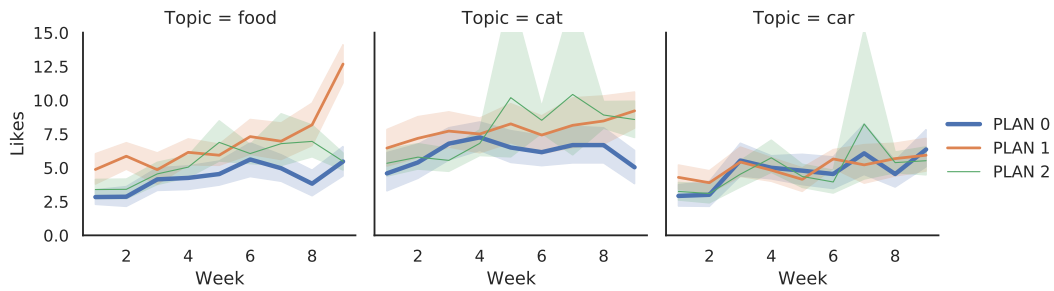


Figure 8.3: Likes trend of our honeypots grouped by engagement plan.

8.4.3 THE IMPACT OF HONEYPOTS CONFIGURATION

We now investigate whether the three variables (i.e., topic, generation strategy, and engagement plan) have a statistical impact on the success of the honeypots, answering RQ_2 and RQ_3 . Given the stationary trend of comments, we focus solely on likes per post and followers per honeypot.

LIKES

Figure 8.4 depicts the distribution of honeypots Likes at the varying of the topic, generation strategy, and engagement plan. In general, there is a difference when the three variables are combined. For example, on average, honeypots belonging to cats, with non-AI generative models, and with PLAN₁ or PLAN₂ have higher values than the rest of the honeypots. Moreover, in general, honeypots adopting PLAN₁ have higher results.

To better understand the different impacts the three variables have on Likes, we conducted a three-way ANOVA. We found that both topic, engagement plan, and generation strategy are significantly (p -value < 0.001) influencing the Likes. Furthermore, we found significance even in the combination of topic and engagement plan (p -value < 0.001), but not in the other combinations. This result confirms the qualitative outcomes we have presented so far. We conclude the analysis by understanding which topic, generation strategy, and engagement plan are more effective. To this aim, we performed Tukey's HSD (honestly significant difference) test with significance level $\alpha = 5\%$. Among the three topics, *cat* is significantly more influential than both *food* and *car* (p -value = 0.001). Regarding the generation strategies, non-AI-based models (i.e., UnsplashModel and InstaModel) outperform AI-based ones. Last, PLAN₁ and PLAN₂ outperform PLAN₀ (p -value = 0.001), while the two plans do not show statistical differences between them.

FOLLOWERS

Tukey's HSD test revealed statistical differences in the number of followers as well. For the analysis, we use the number of followers of each honeypot at the end of the 9th week. We found that *cat* statistically differ from *car* (p -value < 0.01), while there are no significant differences between *cat* and *food*, or *food* and *car*. Regarding the generation strategy, we found no statistical difference among the groups. Finally, all three engagement plans have a significant impact on the number of followers (p -value = 0.001), where PLAN₁ $>$ PLAN₂ $>$ PLAN₀.

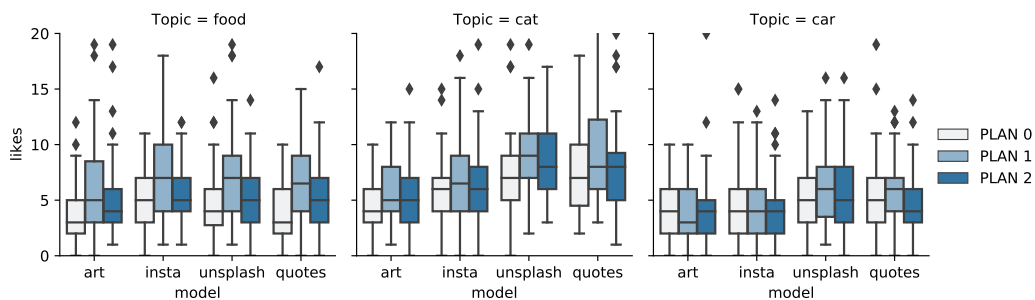


Figure 8.4: Distribution of likes at the varying of topic, model generation strategy, and engagement plan.

AGGRESSIVE ENGAGEMENT PLANS

We recall that honeypots deployed with PLAN 1 and PLAN 2 adopted more aggressive engagement strategies on week 9th: *Follow & Unfollow* for PLAN 1, and *Content Sponsoring* for PLAN 2. Thus, we investigated whether aggressive plans result in more engagement in terms of comments, likes, and followers. The analysis is performed with Tukey’s HSD (honestly significant difference) test with significance level $\alpha = 5\%$. We found no statistical difference in comments in PLAN 1 and PLAN 2. On the opposite, the average number of likes per post shows a statistically significant improvement in PLAN 1 (p -value = 0.01): on average, 7.44 and 9.17 likes per post in weeks 8th and 9th, respectively. No statistical difference is found for PLAN 2; indeed, only the sponsored content benefited (i.e., a few posts).⁹ Last, we analyze the difference between the total amount of followers at the end of weeks 8th and 9th. PLAN 1 honeypots #Followers moved, on average, from 45.7 ± 19.1 of week 8th, to 60.7 ± 26.2 of week 9th, with no statistical difference. PLAN 2 honeypots #Followers moved, on average, from 22.3 ± 11.6 of week 8th, to 30.7 ± 13.9 of week 9th. The difference is statistically supported (p -value < 0.05).

8.4.4 BASELINE COMPARISON

Social honeypots are effective, depending on topics, generation strategies, and engagement plans. Since we are the first, to the best of our knowledge, to examine how to attract *people* using social honeypots (not bots or spammers), there are no state-of-the-art baselines to compare with. Therefore, we compare our methodology with (i) our proposed non-AI generative models with a PLAN 0 engagement strategy (baseline) and (ii) real Instagram pages trends.

BASELINE

This represents the most simplistic method someone might adopt: adding stock images, with random quotes, without caring about the engagement with the rest of the social network. From Section 8.4.3, we statistically showed that the definition of engaging plans is essential to boost engagement in social honeypots. We remark on this concept with Figures 8.5 and 8.6 that show the comparison among the baselines and PLAN 1 social honeypot – which are the most effective ones – in terms of likes and followers over the 9 weeks: in terms of AI and Non-AI strategies, our advanced honeypots outperform in 3 out of 6 cases and 6 out of 6 cases the baselines for likes and followers, respectively. Such results confirm the remarkable performance of our proposed framework. Our strategies might perform worse than the baselines (regarding likes) when the image quality is unsatisfactory. Indeed, as demonstrated in Chapter 3, likes on IG are usually an immediate positive reaction to the post’s image. Since Unsplash images are usually high-quality and attractive, they might have been more appealing than AI-generated images in these cases.

Although comparing our approach with other social honeypots [335, 336, 337] carries some inherent bias (the purpose and social networks are completely different), we still find our approach aligned with (or even superior than) the literature. Lee et al. [335] gained in seven months through 60 honeypots a total of ~ 36000 interactions (e.g., follow, retweet, likes), which is approximately 21.5 interactions per honeypot/week. Our honeypots reached a total of 21870 interactions, which is approximately 115.7 interactions per honeypot/week, i.e., more than five

⁹All sponsored content belongs to weeks before the 9th.

times higher. Yang et al. [336] lured 1512 accounts in five months using 96 honeypots, i.e., 0.788 accounts per honeypot/week. We collected 753 followers, which is 3.98 accounts per honeypot/week, i.e., five times higher. Last, Zhang et al. [337] carefully selected and harnessed 1000 popular Twitter accounts (which they called pseudo-honeypots) for three weeks to analyze spammers. Giving these accounts were already heavily integrated into the social network, they reached over 476000 users, which is around 159 accounts per (pseudo-)honeypot per week. We remind that the purpose of these comparisons is to give an idea of the effectiveness of other social honeypots rather than to provide meaningful conclusions.

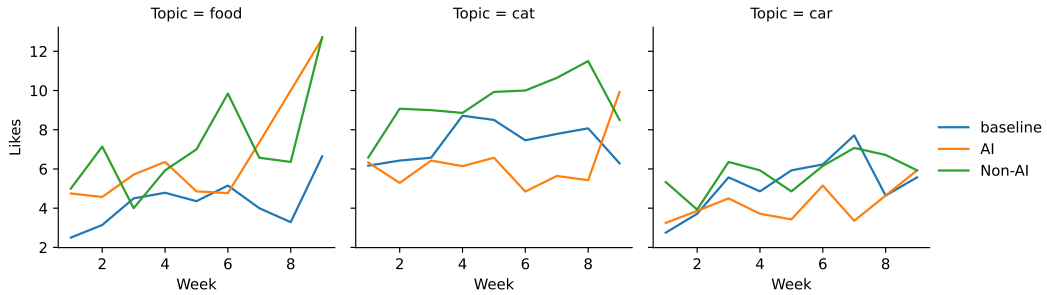


Figure 8.5: Baseline comparison (average likes) with PLAN1 social honeypots.

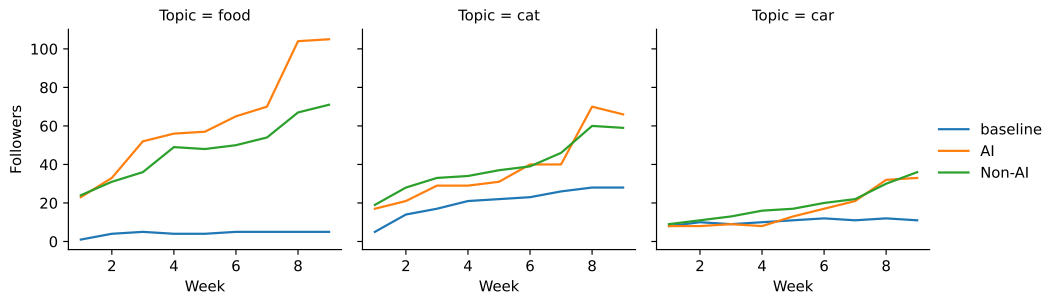


Figure 8.6: Baseline comparison (followers) with PLAN1 social honeypots.

INSTAGRAM PAGES

We now compare our PLAN 1 social honeypots with real IG public accounts. Accordingly, we analyzed the first nine weeks of activities on popular IG pages related to food, cat, and cars. We selected nine popular IG pages for each topic, 3 with $\sim 10K$ followers, 3 with $\sim 100K$ followers, and 3 with more than a million followers. We collected the number of comments and likes for each post published during this period. Due to IG limitations, we could access only information at the time of collection, implying that posts might be a few years old. Monitoring new pages would be meaningless since we do not know a priori whether they will become popular.

We noticed that it is impossible to compare such baselines with our social honeypots because, generally, the considered IG pages contain posts with hundreds of likes and comments even in their first week of activity. For

instance, $+1M$ pages' first posts reached more than 2000 likes. Possible explanations behind this phenomenon are: (i) the considered 18 pages were already popular before their creation (e.g., on a different or older OSN like Facebook); (ii) the considered 18 pages massively sponsored all their content; (iii) we are facing the *earlybird bias*, where older posts contain not just engagement from the first nine weeks, but also engagement from later periods, even years.¹⁰ To further explain this phenomenon, we contacted such IG pages (we extended our survey to 36 pages). Questions focused on the first weeks of activity.¹¹ Unfortunately, up to the submission date, none of the contacted pages replied.

Although there is no evidence in the literature on how long it takes to make an Instagram page famous, most sources consider the initial growth (from 0 to 1000 followers) to be the most challenging part [361, 362], with an overall monthly growth rate of about 2% [363]. Furthermore, success requires lots of dedication to follow best practices consistently [364], which is extraordinarily time-consuming and far from trivial. Being in line with these trends in a fully automated and effortless manner is already an impressive achievement. Our work can serve as a baseline and inspiration for future work.

8.5 SOCIAL ANALYSES

In this section, we perform diverse social analyses based on the engagement we created with our honeypots. We start by analyzing comments in Section 8.5.1, followers in Section 8.5.2, and reached audience in Section 8.5.3.

8.5.1 COMMENTS ANALYSIS

An interesting (and unexpected) result is that, without the premeditated intention of building spammer detectors, most of the comments we received came from spammers. To estimate the total number of spam comments, we first manually identified patterns used by spammers on our honeypots (e.g., expressions like “send pic” or “DM us”). Afterward, using a pattern-matching approach, we found that 95.33% of the comments we received on our social honeypots came indeed from spammers. All spammers' accounts shared similar behavior in commenting: (i) there was always a mention ‘@’ to other accounts, and (ii) they commented almost immediately after the post creation. Such considerations suggest these accounts are bots that target many recent posts, perhaps searching by specific hashtags. Such findings indicate that fresh pages could be a powerful tool to detect spammers with *minimal* effort. We also highlight that spam comments are a well-known issue that affects the majority of IG pages [365] and is not limited to our honeypot pages. Therefore, we argue that creating pages that do not attract spammers is nearly impossible. Nevertheless, IG itself is employing and improving automatic screening mechanisms [237, 238] to limit such behavior. When such mechanisms are enhanced, our honeypots will become more accurate.

¹⁰Earlybird bias appears in other social contexts like online reviews [360].

¹¹For instance, we asked whether the page resulted from an already existing page (on IG or other platforms), or the strategies they adopted to manage the pages (e.g., spam, sponsoring).

8.5.2 FOLLOWERS ANALYSIS

As most of our comments were spam, we investigated whether followers were the same. We manually inspected the followers of our most followed social honeypot for each topic, identifying three categories of followers:

- *Real people*: users that publish general-topic posts, with less than 1000 followers¹², and real profile pictures;
- *Pages and Influencers*: users that publish topic-specific posts (e.g., our honeypots) or with more than 1000 followers;
- *Bots*: users whose characteristics resemble a bot, following well-known guidelines [229], e.g., fake or absent profile picture, random username, highly imbalanced follower/following count, zero or few (< 5) posts.

From Table 8.3, we notice the three honeypots have different audiences. The *food* honeypot obtained the most real followers, *car* reached more bots, and *cat*, was followed mainly by pages. These results confirmed that (i) our honeypots can reach real people, (ii) the audience category depends on the topic, and (iii) spammers' threat is limited to comments. On an interesting note, most pages following our *cat* honeypot were cat-related pages.

Table 8.3: Percentage of real people, pages, and bots for the best social honeypot in each topic.

	Real People	Pages	Bots
Food	48,08%	37,50%	14,42%
Cat	10,61%	72,72%	16,67%
Car	30,30%	21,21%	48,49%

8.5.3 REACHED AUDIENCE

We conclude the experimental results with a detailed analysis of the audience our honeypots reached. In particular, we performed two distinct analyses: (i) *Honeypot reached audience*, and (ii) *Sponsored posts audience*, i.e., IG features available for honeypots with 100 followers and sponsored content, respectively. After nine weeks of computation, one honeypot satisfies the requirement of 100 followers (honeypot ID: h9). About the sponsored content, we obtained information about 9 posts (one per honeypot belonging to PLAN 2).

HONEYPOT AUDIENCE

The honeypot h9 (topic: food, generation strategy: AI, and engagement plan: PLAN 1) gained 103 followers: the majority is distributed over the age range [25 – 34] with 32% (equally distributed among men and women), [35, 44] with 10% of women and 27% of men. Most followers came from India (11.7%), Bangladesh (10.7%), and Japan (9.7%).

¹²After 1000 followers, users are considered nano influencers [366].

SPONSORED POSTS AUDIENCE

For this analysis, we recall that we set our sponsoring strategy leveraging the automatic algorithm provided by IG. Overall, sponsored posts achieved great success in terms of generated engagement. On average, food posts reached 30.6, 116, and 60.6 likes for food, cat, and car posts, respectively. These numbers are strongly above the average likes per post 5.9. IG offers an analytic tool to inspect the reached audience; this feature perfectly fits in the scope of social honeypots, since it allows finding insights about the attracted audience. For each post, the following information is available: quantitative information (i.e., reached people, likes, comments, sharing, saved), and demographic distribution in percentage (gender, age, location). The detailed report is available in Appendix 8.7.3. We observed interesting trends:

- *food* audience: the gender is almost balanced (female audience slightly more attracted), and the predominant age range is 18-34. Top locations: Campania, Lombardia, and Puglia.¹³
- *cat* audience: the gender distribution is toward the female sex, and the predominant age range is 18-34. Top locations: Emilia Romagna, Lombardia, Piemonte.
- *car* audience: the gender is strongly distributed toward the male sex, and the predominant age range is 18-24. Top locations: Lazio, Lombardia.

To conclude, with minimal effort (i.e., € 2/day per post), an owner can get useful information, e.g., to use in marketing strategies..

8.6 TOWARD A REAL IMPLEMENTATION

So far, we have demonstrated our social honeypots can attract real people in a fully automated way. With little effort, they can already be deployed for an array of situations. In this section, we first reason about the use cases of our approach, highlighting both positive and negative outcomes. Then, we present the current challenges and limitations of implementing this work in real scenarios.

8.6.1 USE CASES

Our work aims to show the lights and shadows of social networks such as Instagram. People can easily deploy automated social honeypots that can attract engagement from hundreds or even thousands of users. Upon that, analyses on these (unaware) users can be conducted. As cyber security practitioners, we know that this technology might be exploited not only for benign purposes, but also to harm users [367]. Therefore, this work contributes to the discussion about the responsible use of online social networks, in an era when technologies like artificial intelligence are transforming cyber security. We list in this section possible social honeypot applications.

¹³IG automatic algorithm maximized the audience toward authors country, i.e., Italy, reporting Italian regions.

MARKETING

The first natural adoption of our proposed social honeypots is for marketing purposes. Suppose someone is interested in understanding “who is the average person that loves a specific theme”, where themes might be music, art, puppies, or food. With a deployed social honeypot, the owner can then analyze the reached audience by using the tools offered by IG itself (as we ethically did in this chapter) or by further gathering (potentially private) information on the users’ profile pages [368].

PHISHING AND SCAM

Similarly to marketing, social honeypots can be used by adversaries to conduct phishing and scam campaigns on IG users. For instance, the social honeypot might focus on cryptocurrency trading: once identified potential victims, attackers can target them aiming to obtain sensitive information (e.g., credentials), or to lure them into fraudulent activities such as investment scams, rug pulls, Ponzi schemes, or phishing.

SPAMMER IDENTIFICATION

Social honeypots can also be created to imitate social network users, by posting content and interacting with other users. As we noticed in our experiments, they can attract spammers. Therefore, our proposed framework can be adopted by researchers to spot and study new types of spamming activities in social networks.

MONITORING OF SENSIBLE THEMES

An interesting application of social honeypots is to identify users related to sensible themes and monitor their activities (within the honeypot). Examples of such themes are fake news and extremism [369]. Researchers or authorities might leverage social honeypots to identify users that actively follow and participate in such themes, and then carefully examine their activity. For instance, honeypot owners can monitor how people respond to specific news or interact inside the honeypot.

8.6.2 CHALLENGES AND LIMITATIONS

The first challenge we faced in our work is the massive presence of spammers on IG. Most of them are automated accounts that react to particular hashtags and comments under a post for advertisement or scamming purposes [370, 371]. This factor can inevitably limit our approach when we aim to gather only real people. As a countermeasure, honeypots should include a spam detector (e.g. [370, 372]) to automatically remove spammers. On the contrary, this approach could be useful directly to reduce the spamming phenomenon. Many pages can be created with the purpose of attracting spammers and reporting them to IG for removal.

The second challenge we encountered is the lack of similar works in the literature. Because of this, we have no existing baselines to compare with, and it could be difficult to understand whether our approach is truly successful. However, in nine weeks, we obtained more than 15k likes and gathered ~ 750 followers in total, which is not trivial as discussed in Section 8.4.4. Our most complex methods surpassed the simplest strategies we identify, which can serve as a baseline and source of inspiration for future works.

Among the limitations, we inspected only generic (and ethical) topics. A comprehensive study in this direction would give much more value to our work, especially dealing with delicate topics (e.g., conspiracies, fake news). Moreover, our approach is currently deployable on IG, but would be hard to transfer to other platforms. Even if this can be perceived as a limitation, it would be naive to consider all social media to be the same. Indeed, each of them has its own content, purpose, and audience. Developing social honeypots for multiple platforms can be extremely challenging, which is a good focus for future research. Last, there was no clear connection between the posts of our honeypots. When dealing with specific topics, it might be necessary to integrate more cohesive content.

8.7 CONCLUSIONS

The primary goal of this work was to first understand the feasibility of deploying self-managed Instagram Social Honeypots, and we demonstrated that *it is possible* in Section 8.4.1. Moreover, from the results obtained in our analyses we can derive the following outcomes and guidelines:

1. *Topics* plays an important role in the success of the honeypot.
2. *Generation strategies* does not require complex DL-based models, but simple solutions such as stock images are enough. Similarly, we saw that posts containing random quotes as captions are as effective as captions describing the content;
3. *Engagement plan* is essential. We demonstrated that a naive engagement strategy (PLAN 0) results in a low volume of likes and followers. Moreover, the engagement plan without costly operations (PLAN 1) works as well as plans involving followers acquisition and content sponsoring;
4. *Sponsored content* is a useful resource to preliminary assess the audience related to a specific topic;
5. Social honeypots not only attract *legitimate* users, but also *spammers*. As a result, they can be adopted even for cybersecurity purposes. Future implementation of social honeypots might include automatic tools to distinguish engagement generated by legitimate and illegitimate users.

In conclusion, we believe that our work can represent an important milestone for future researchers to easily deploy and collect social network users' preferences. New research directions might include not only general topics like cats and food, but more sensitive themes like fake news, or hate speech. In the future, we expect generative models to be always more efficient (e.g., DALL-E 2 [349] or ChatGPT [350]), thus increasing the reliability of our approach (or perhaps making it even more dangerous).

ETHICAL CONSIDERATIONS

Our institutions do not require any formal IRB approval to carry out the experiments described herein. Nonetheless, we designed our experiments to harm OSN users as less as possible, adhering to guidelines for building Ethical Social Honeypots [373], based on the Menlo report [141]. Moreover, we dealt with topics (cars, cats, food) that should not hurt any person's sensibility. In our work, we faced two ethical challenges: data collection and the use of deception. Similar to previous works [335, 336, 334], we collected only openly available data (provided by

Instagram), thus no personal information was extracted, and only aggregated statistics were analyzed. Moreover, all information is kept confidential and no-redistributed. Upon completion of this study, all collected data will be deleted. This approach complies with the GDPR. To understand the honeypot's effectiveness, similar to previous works, we could not inform users interacting with them about the study, to limit the Hawthorne effect [250]. However, we will inform the deceived people at the end of the study, as suggested by the Menlo report.

APPENDIX

8.7.1 IMPLEMENTATION DETAILS

8.7.2 MODELS

In this appendix we will describe how InstaModel, ArtModel, UnsplashModel and QuotesModel were implemented. All of them have different characteristics but, at the same time, share some common functionalities that will be explained before of the actual implementation of the four models.

SHARED FUNCTIONALITIES One of the shared functionalities is adding emojis to the generated text. This is done with a python script which scans the generated caption trying to find out if there are words that can be translated with the corresponding emoji. To make this script more effective, it looks also for synonyms of nouns and adjectives found in the text to figure out if any of them can be correlated to a particular emoji. As last operation, the script chooses randomly, from a pool of emojis representing the "joy" sentiment, one emoji for each sentence that will be append at the end of each of them.

CTA are simple texts that may encourage a user to do actions. These CTA are sampled randomly from a manually compiled list and then added at the end of the generated caption.

The last shared feature is the selection of hashtags. As said before, through the Instagram Graph API we are able to get the first 25 posts for a specific hashtag and from them we extracted all the hashtags contained in the caption. Thus we compiled an hashtag list for each of the three topic sorted from the most used to the least used. Instagram allows to insert at most 30 hashtags in each posts but we think that this number is too high with respect to the normal user's behavior. For this reason, we decided to choose 15 hashtags that are chosen with this criteria: 8 hashtags are sampled randomly from the first half of the list in the csv file, giving more weight to the top ones, while the other 7 are sampled randomly from the second half of the list, giving more weight to the bottom part of the list. The intuition is that we are selecting the most popular hashtags together with more specific hashtags.

INSTAMODEL Starting from the caption generation, InstaModel uses the Instagram Graph API to retrieve the top 25 posts for a specific hashtag. In practice, the chosen hashtag will be the topic on which the corresponding honeypot is based. Once we have all the 25 posts, they are checked to save only those that have an English caption before being passed to the object detector block. The object detector is implemented by using the InceptionV3 model for object detection tasks. InceptionV3 detects, in the original image, the object classes with the corresponding accuracy and if the first's class score is not greater than or equal to 0.25, the post will be discarded. Otherwise, the other classes are checked as well and only if their scores are greater than 0.05 will be considered as keywords for the next step. Regarding the original caption, nouns and adjectives are extracted by using nltk python library. Notice that words such as "DM" or "credits" and adjectives such as "double" or similar, are not considered. This is because they usually belong to part of the caption that is not useful for this process.

Keyword2text¹⁴ is the NLP model that transforms a list of keywords in a preliminary sentence. This preliminary sentence is then used by OPT model to generate the complete text. Considering the computational resources

¹⁴<https://huggingface.co/gagan3012/k2t>

available to us, the model used is OPT with 1.3 billion parameters. We suggest to save the text generated by OPT in a file text because it will be used subsequently to generate the corresponding image. Once we have the complete generated text, emojis are added together with a CTA sentence that is standard in any post. The last step for caption generation is to append hashtags: they will be chosen by sampling from the corresponding csv file with the reasoning mentioned above.

The last step of InstaModel is image generation and for this purpose Dall-E Mini ([357]) is used. The prompt will be the text generated after the OPT stage, the one that has been save separately. It is relevant to highlight that the process with Dall-E Mini is not completely automatic and there should be a person that choose the most suitable image for the giving caption.

ARTMODEL ArtModel starts from a prompt generated with a python script and uses Dall-E mini, like InstaModel, to generate the corresponding image. The style and the medium are chosen randomly from two lists. Example of styles can be "cyberpunk", "psychedelic", "realistic" or "abstract" while examples of medium are "painting", "drawing", "sketch" or "graffiti". The topic of the honeypot is used as subject of the artistic picture generated by Dall-E Mini. Once the image is generated, the prompt, added of emojis, CTA and the corresponding hashtags, will be used as Instagram caption.

UNSPASHMODEL UnslashModel does not generate images but uses stock images retrieved from the Unsplash websites. Unsplash has been chosen not only because it gives the opportunity to find images together with the relative captions, but also because it offers API for developers that can be used easily. To avoid reusing the same images more than once, each image's id is saved in a text file which will be checked at each iteration. For the caption generation, the original caption is processed by Pegasus model ([358]) which is an NLP model quite good in the rephrase task. As always, emojis, CTA and hashtags are added to the final result.

QUOTESMODEL QuotesModel makes use of Pixabay¹⁵ stock images website to avoid reusing Unsplash even for this model. Also in this case, we use the topic of the specific honeypot as query tag. As for UnsplashModel, to avoid reusing the same image for different posts, once we have downloaded the image, its id is saved in a text file which will be checked every time needed. For the caption generation, a quote is sampled randomly from a citation dataset [344]. In this case, the model does not add emojis to the text because we think that the quote, by itself, can be a valid Instagram caption. On the contrary, as always, CTA and hashtags are added to the text.

SPAMMING

Honeypots with PLAN 1 or PLAN 2 engagement plans will automatically interact with the posts of other users. The idea is to retrieve the top 25 Instagram posts for the hashtag corresponding to the specific topic of the honeypot and like and comment each of them.

For the implementation we used Selenium which is a tool to automates browsers and it can be easily installed with pip command. Selenium requires a driver to interface with the chosen browser and in our case, since we

¹⁵<https://pixabay.com/>

Table 8.4: Overview of the sponsored content attracted users

<i>Overview</i>									
<i>honeypot</i>	h3	h6	h7	h10	h13	h14	h17	h20	h21
<i>topic</i>	food	food	food	cat	cat	cat	car	car	car
<i>gen. strat.</i>	AI	NON AI	NON AI	AI	NON AI	NON AI	AI	NON AI	NON AI
<i>audience</i>	3126	3412	5337	3245	4597	2863	10698	6824	9633
<i>likes</i>	21	34	37	118	163	67	20	25	127
<i>comments</i>	1	3	7	3	8	1	3	11	3
<i>saved</i>	1	0	21	12	29	7	2	6	44
<i>Gender Coverage [%]</i>									
<i>women</i>	42.2	60.0	87.8	67.2	67.7	59.0	8.6	8.7	5.6
<i>men</i>	57.0	38.7	11.7	31.5	30.7	39.3	89.5	90.7	93.6
<i>Age Coverage [%]</i>									
13 – 17	0.1	0.1	0	0	0	0.1	0.2	0.1	0.1
18 – 24	39.1	37.7	35.9	20.8	33.8	38.6	64.3	45.7	52.5
25 – 34	29.8	12.9	36.0	21.2	25.2	15.2	12.7	31.8	26.8
35 – 44	14.5	11.6	14.3	15.6	13.0	12.4	6.5	10.8	9.4
45 – 54	9.0	18.3	8.2	18.7	14.0	13.7	8.1	5.1	6.1
55 – 64	4.7	12.9	3.8	15.8	9.3	12.4	5.0	3.6	3.0
65+	2.5	6.0	1.3	7.5	4.3	7.2	2.9	2.6	1.8
<i>Geographic Coverage [%]</i>									
<i>Campania</i>	14.7	11.3	9.1	N.A.	N.A.	8.7	7.8	8.7	N.A.
<i>Emilia-Romagna</i>	N.A.	N.A.	N.A.	9.7	8.7	9.2	N.A.	8.6	9.2
<i>Lazio</i>	N.A.	7.9	8.3	9.4	10.5	N.A.	8.2	11.1	9.5
<i>Lombardia</i>	12.4	12.0	13.2	19.6	18.8	17.2	14.0	19.0	20.9
<i>Piemonte</i>	N.A.	N.A.	N.A.	9.0	8.5	7.5	N.A.	N.A.	8.0
<i>Puglia</i>	12.5	10.9	8.9	N.A.	N.A.	N.A.	8.9	N.A.	N.A.
<i>Sicilia</i>	9.0	10.0	9.2	N.A.	N.A.	N.A.	10.4	N.A.	N.A.
<i>Tuscany</i>	N.A.	N.A.	N.A.	7.2	N.A.	N.A.	N.A.	N.A.	N.A.
<i>Veneto</i>	9.0	N.A.	N.A.	N.A.	7.7	8.4	N.A.	8.8	10.1

chose Firefox, we have downloaded the geckodriver. The implementation consists of a python class which has three main methods: `login`, `like_post` and `comment_post`

The `login` method is invoked when the honeypot accesses to Instagram. The `like_post` method searches, in the DOM, for the button corresponding to the like action and then it clicks it. The `comment_post` method searches in the DOM for the corresponding comment button and then clicks it. Afterwards, it searches for the dedicated textarea and write a random sampled comment. Finally, it clicks the button to send the comment.

8.7.3 SPONSORED CONTENT ANALYSES

We report in Table 8.4 the complete overview of audience attracted by our sponsored content. In particular, we report overall statistics in term of quantity (e.g., number of likes), and demographic information like gender, age, and location distribution.

Part III

Security and Privacy Concerns in Modern Social Platforms

INTRODUCTION TO PART III

The concept of social networks has evolved beyond conventional OSNs to encompass modern digital platforms such as video games and the Metaverse. In these virtual environments, individuals actively engage, communicate, and establish connections, fostering online communities and social interactions. However, the extensive adoption of these social platforms leads to the generation of vast quantities of data, often public, which can be susceptible to malicious exploitation. Unfortunately, this domain is mostly unexplored. Yet, techniques such as user identification and profiling may aid in detecting cybercriminals and significantly mitigating these threats. This final part of the doctoral thesis focuses on enhancing users' security and privacy of modern social platforms, namely video games and the Metaverse. The part begins by introducing PvP, a video game identification framework designed to safeguard gamers from malicious activities. Then, the thesis evaluates a novel attribute inference attack on Dota 2, a popular online video game, unveiling a subtle privacy threat impacting millions of gamers worldwide. Last, it presents a user profiling framework tailored for augmented and virtual reality, the foundational technologies underpinning the Metaverse. This framework aims to highlight privacy challenges within this modern platform and enhance both the security and experience of their users.

9

PvP: Profiling Versus Player! A Framework for User Identification in Online Video Games

Over the past few decades, the realm of video games has undergone an astonishing transformation, evolving into a cultural phenomenon of unprecedented magnitude. As of August 2023, 3.2 billion people worldwide, roughly equating to 40% of the global population, actively participate in gaming [374]. The video game industry, once relegated to the fringes of entertainment, is now a thriving behemoth, poised to generate an estimated \$334 billion in revenue in 2023 [375].

However, the increasing popularity of online video games opened up a plethora of new paths for fraudulent activities. In-game purchases have become a ubiquitous feature, accompanied by the trend of one-click payments, where users' payment details are stored within their accounts for swift transactions. Consequently, account takeovers have emerged as a prime target for cybercriminals. Alarming statistics from a recent study [376] reveal that nearly half of all console gamers partake in in-game purchases, with at least one-fifth falling prey to payment fraud. Notable security breaches have marred the gaming landscape in recent years. In 2011, the compromise of the PlayStation Network affected over 77 million accounts [377]. In 2015, Valve, Steam's developer, revealed an average monthly account theft rate surpassing 70,000, leading to implementing a two-factor authorization system [378]. In 2018, prominent titles like League of Legends (LOL) and Fortnite grappled with security breaches. LOL players witnessed massive phishing attacks [379], while Fortnite players reported unauthorized financial transactions linked to their accounts [380]. In 2020, Nintendo succumbed to a significant data breach, resulting in the compromise of more than 300,000 accounts, along with the leaking of users' payment information [381]. Regrettably, the pattern persists, exemplified by the large-scale data breach of Bandai Namco in 2022 [382].

Beyond these high-profile incidents, a common modus operandi among scammers involves adding victims

to their friends' lists to initiate conversations and lure them through malicious links or enticing trade proposals. To bolster their credibility, these malevolent actors often send friendship requests to their victims after a match played together. Despite potential bans and reports, scammers routinely evade penalties by creating new accounts, perpetuating their malicious activities.

Tragically, scams and account takeovers merely scratch the surface of the malevolent activities that pervade the gaming landscape. Cyberbullying, grooming, and harassment are distressingly prevalent issues in contemporary gaming [383, 384]. Again, when these harmful users are identified and their accounts banned, they can create new accounts to perpetuate their activities.

CONTRIBUTION All these malicious behaviors can be mitigated by being able to uniquely identify a player, irrespective of the account they use. Analogous to fingerprints in the real world serving as identifiers and tools for law enforcement, a “video game fingerprint” has the potential to detect and subsequently ban perpetrators across all accounts they employ. We hypothesize that such a digital fingerprint could be discerned from a gamer’s play style. In essence, how an individual navigates, explores, and engages within the virtual world possesses biometric qualities that permit recognition.

In this chapter, we demonstrate that video game players can be identified and distinguished based on their in-game data and play-style attributes. We present a comprehensive methodology for the extraction of game-related features, the aggregation of in-game data, and the application of deep learning techniques for player identification. To this end, we introduce a novel identification framework entitled “Profiling vs Players” (PvP)¹. We tested PvP on data collected from 50 Dota 2 players and 50 CS: GO players, encompassing a total of 10,000 matches. In both games, PvP achieved more than 90% accuracy using two minutes of gameplay data. Noteworthy, Dota 2 and CS: GO represent very different gaming genres (MOBA and FPS, respectively) with distinct in-game characteristics, underscoring PvP’s remarkable adaptability to diverse gaming environments. Our extensive experiments further underscore that player identification can be achieved using rudimentary features commonly found in most video games, such as moving the character or controlling the camera view. By discerning the unique play styles and patterns of players, this framework can aid game developers, esports organizations, and researchers in various domains.

Nevertheless, while PvP holds the potential to mitigate the issues previously mentioned, it could also inadvertently exacerbate harmful behaviors. For instance, victims of cyberbullying who create new accounts to evade tormentors could find themselves pursued through the analysis of their play styles. Thus, we aim to not only demonstrate the feasibility of player recognition but also to raise awareness about potential vulnerabilities and threats associated with this technology.

This chapter extends our previous work “PvP: Profiling versus Player! Exploiting Gaming Data for Player Recognition” [385], which encompasses a case study on the video game Dota 2. This extended version diverges in several key aspects:

- We generalized our methodology, rendering PvP a versatile framework ideally adaptable to a wide array of video games;
- We test PvP on an entirely distinct game, Counter-Strike: Global Offensive (CS:GO). The high identification results achieved in both Dota 2 and CS: GO affirm PvP efficacy in diverse gaming landscapes;

¹The name resembles the widespread game mode Player vs Player.

- We conduct extensive supplementary experiments in both Dota 2 and CS: GO, providing comprehensive insights into PvP’s performance in various gaming contexts.

ORGANIZATION The chapter is organized as follows. Section 9.1 explores related works. Section 9.2 and Section 9.3 present the PvP framework and the games selected for experiments, respectively. Section 9.4 illustrates the data collection, while PvP is tested in Section 9.5. Section 9.6 reports additional experiments in both games, Section 9.7 provides discussion on the findings, and Section 9.8 concludes the chapter.

9.1 RELATED WORK

Video games, deep learning, and privacy have all been examined extensively in the literature, but they have been discussed together only recently. In Section 9.1.1, we explore general video game-related works, including discussions on privacy and machine learning applications. In Sections 9.1.2 and 9.1.3, we focus on Dota 2 and CS: GO-related works, respectively.

9.1.1 VIDEO GAMES RELATED WORKS

IMPACT OF VIDEO GAMES

One of the main interests in the field is understanding the benefits of playing video games and their impact on society. Initially, researchers focused on finding relations between video games, the brain, and the human thinking process. Gong et al. [386] conducted a study on over 27 “expert” players of League of Legends and Dota 2, revealing that playing video games increases the amount of grey matter and promotes better connectivity in a person’s brain. Similarly, Gee [387] discusses how video games can illuminate the nature of human thinking and problem-solving as situated and embodied. He explored why people became more interested in video games to study the human thinking process, analyzing the “projective stance”, an embodied thinking frequent in many video game players. Steffie et al. [388] studied the relation between reciprocity, social capital, social status, and group play, focusing on social behavior in games.

EDUCATION AND PERCEPTIONS

Although video games’ influence on child health is usually perceived to be negative, Kovess-Masfety et al. [389] proved that children who play more video games may be more likely to develop good social skills and build better relationships. Other studies (e.g., [390], [391]) examined the history of games in educational research, trying to understand the true potential of educational video games. Squire [390], in particular, suggests that educators underestimate the potential of educational video games. Moreover, Griffiths [392] showed that playing video games is safe for most people and can help in some situations, such as pain management. Later in the years, Granic [393] demonstrated that children who play strategy-based games usually improve their problem-solving skills, getting better grades in the next school year.

The correlation between violence and video games is also a central theme in video game studies [394]. Often, the results are contradictory. For instance, Gee [395] stated that violent video games pollute the cultural environment of children, stunt their brain development, and provoke aggressive behaviors. In contrast, Markey et al. [396] argued that no proof suggested a relation between violence in video games and real-world violence in the United States. Last, people started considering some video games as pieces of art [397], according to historical, aesthetic, institutional, representational, and expressive theories of art.

MACHINE LEARNING APPROACHES

Conventional Machine Learning (ML) techniques have found applications in video games for various purposes. In 2009, Drachen et al. [398] adopted an unsupervised learning approach to create player models for "Tomb Raider: Underworld." A year later, in the same game, diverse supervised learning algorithms were trained on an extensive dataset of player behavior data to predict player disengagement and game completion times [399]. ML techniques have also been harnessed to categorize player behaviors, such as segmenting Minecraft players based on their time allocation to building, mining, exploration, and combat activities [400]. Lastly, by utilizing ML techniques in conjunction with smart chairs to collect player data, it was demonstrated that one can ascertain a user's professional gamer status [401].

PRIVACY AND PROFILING

Privacy issues in video games are a recent concern. Newman and Jerome [402] evidence companies use sensors to gather players' data through consoles. Players' voices, physical appearance, or geographical location are of main interest. Moreover, players' psychographic information can be obtained from their in-game interactions. In Chapter 10, we will demonstrate how video gamers' personal information, such as age, gender, or personality traits, can be inferred by harnessing their publicly available data, accessible through websites that track their statistics (Tracking Websites). Russell et al. [403] overviewed how modern games align with information privacy norms and notions. Furthermore, it analyzed how users, in particular child gamers, may be affected by data practices and technologies specific to gaming. Many means have been used to identify and profile users in modern technologies, such as movements information published on a social network in [404], or laptop power consumption [405]. However, to the best of our knowledge, a user profiling framework for video games is still missing. In our prior study [385], we achieved over 96% F1-score in uniquely identifying Dota 2 players using Deep Learning. Building on this success, we introduce the PvP (Profiling versus Players) framework, extending our approach to recognize and profile players possibly in *any* video games.

9.1.2 DOTA 2 RELATED WORKS

Different studies have been conducted on Dota 2 over the years. Gao et al. [406] developed several ML classifiers to detect hero roles and positioning from their IDs, achieving 75% and 85% of accuracy, respectively. The work was refined to identify player roles in a team, achieving 96.15% accuracy through a Logistic Regression [407]. More recently, OpenAI adopted deep reinforcement learning to create OpenAI Five, a Dota 2 team composed of five bots trained for over ten months, which was the first artificial team able to defeat world champions in an

esports game, demonstrating that self-play reinforcement learning can overcome a difficult task with superhuman performance [408]. Last, some works tried to detect malicious activities in Dota 2. For instance, Qian et al. [409] proposed an anomaly detection algorithm in player performances to detect cheating. Instead, Ding et al. [410] adopted an unsupervised learning approach to understand if a specific player is a smurf or a booster, reaching 95% accuracy.

9.1.3 CS: GO RELATED WORKS

Counter-Strike: Global Offensive (shortly, CS: GO) is a widely played video game subject of different studies. For example, since CS: GO allows trading skins for real money via Steam, it is possible to find communities devoted to this exchange. Yamamoto and Mc Arthur [411] illustrate how players use this exchange ecosystem to earn money, and identify the most decisive factors determining the skins' value.

Other research studies delve deeper into game analysis and social constructs. For instance, CS: GO has been studied extensively to understand the essence of First Person Shooters (FPS), focusing on their gameplay dynamics and round systems [412]. Another notable example is introducing a context-aware framework for assessing player actions and performance in CS: GO [413]. This framework excels at identifying high-impact actions correlated with a team's likelihood of winning. Further, in this context, Machine learning techniques were applied to CS: GO for winning prediction [414]. Regarding social studies, Rusk and Stahl [415] tried to unravel the social structure within the game, utilizing in-game events such as kills and deaths as markers. Similarly, Sasmoko et al. [416] emphasize understanding the impact of Game Experience (GX) on player emotions during gameplay. Meanwhile, Staahl et al. [417] explored the (co)construction of player identities, examining the tools employed for this purpose and identifying a predominant trend toward a perceived competent player identity influenced by technomasculture norms.

9.2 OUR FRAMEWORK: PVP (PROFILING VS PLAYER)

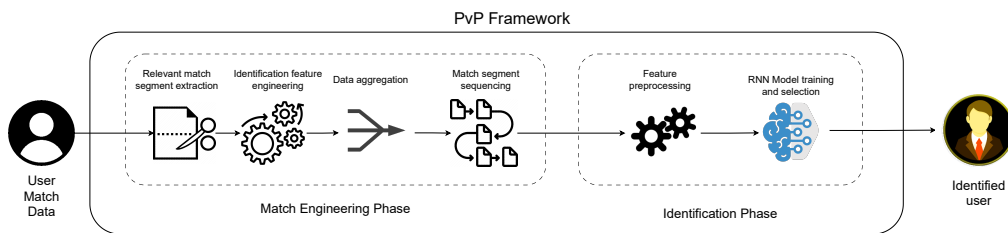


Figure 9.1: Schema of the proposed framework

In this Section, we present our framework Pvp (Profiling versus Player). We start by giving an overview of the framework in Section 9.2.1. Then, we present in detail the data acquisition phase in Section 9.2.2, the game specific phase in Section 9.2.3, and the Deep Learning phase 9.2.4.

9.2.1 OVERVIEW AND GOAL OF THE FRAMEWORK

The framework’s goal is to uniquely identify a player within a known player pool, relying on their individual play style retrievable from their matches. To formalize this, consider a set of players denoted as \mathcal{P} and a corresponding list of matches, \mathcal{M} , played by these individuals. We aim to develop a Classifier C with the capacity to establish a connection between a given match $m \in \mathcal{M}$ (or a part of it) and the specific player $p \in \mathcal{P}$ who played it. Ultimately, C should be able to associate an unseen match, $m \notin \mathcal{M}$, (or a part of it) with the player $p \in \mathcal{P}$ who played it.

The framework foresees several parts. Figure 9.1 presents the overview of PvP. The first step involves gathering players’ data, i.e., their matches. Depending on the framework’s adopter, various strategies can be employed during this phase. The second step entails engineering matches’ data to extract unique features related to players’ play styles (i.e., their “fingerprint”). In this stage, an in-depth understanding of the game can help in identifying these unique characteristics. Last, the identification phase is responsible for training the classifier C to associate a match m to the player p who played it. During this stage, we employ deep learning algorithms, known for their high proficiency in this regard.

APPLICATIONS. PvP offers various potential applications. By creating a unique “fingerprint” for each video gamer, game publishers can effectively curtail or prevent the proliferation of malicious activities. Indeed, activities such as harassment and cyberbullying are unfortunately too common in online video games [384, 383]. Typically, games provide users with the option to report misbehaving players, leading to account bans. However, this approach fails to deter banned players from creating new accounts and continuing their disruptive behavior. With the implementation of a fingerprinting system, malicious players can be traced across every account they create or play on, thus mitigating this issue. Conversely, it is important to acknowledge that attackers may chase their victims in case they create new accounts to escape. Nevertheless, raising awareness is the initial step in enhancing cybersecurity.

Furthermore, PvP can be used to identify smurfs or account boosters [410], enabling punitive actions to be taken against their primary accounts, which they typically value highly. Additionally, the framework can serve as a foundation for authentication mechanisms, offering a transparent background process to prevent account takeovers. Lastly, if multiple individuals share an account (e.g., parents and children), custom experiences can be tailored based on the player’s identity.

9.2.2 DATA ACQUISITION PHASE

The initial step in implementing our framework is data acquisition, and the method depends on the entity utilizing the framework. If the video game’s publisher employs our framework, they can access the data directly from their own servers. Conversely, when a third party seeks to utilize our framework, the data acquisition process becomes slightly more complex. Typically, video games offer APIs and cooperate with tracking websites, offering access to replays or matches with detailed statistics. These data are usually public and retrievable with no effort. In our specific case, we employed tracking websites for data acquisition in both of the video games we used to test the framework.

9.2.3 MATCH ENGINEERING PHASE

This phase entails the transformation of match data into a user’s distinct play style, essentially creating their unique “fingerprint.” Given the substantial differences between various games, the features crucial for distinguishing players can vary significantly. Consequently, having a comprehensive understanding of the specific game in question can greatly aid in identifying these distinctive characteristics. During this stage, the framework’s users should pose a fundamental question: “What in-game actions can each player execute uniquely?”

Here, we present a general approach that is adaptable to virtually any game but should not deter users from enhancing this process with more contextually meaningful features. The outcome of this phase yields a dataset to be utilized in the subsequent “Identification Phase”.

RELEVANT MATCH SEGMENT EXTRACTION

This phase is dedicated to determining the best match segment for the identification process. Given that matches can often extend for a considerable duration², encompassing various in-game situations, it is essential to identify a segment that minimizes classifier noise. Ideally, this segment should capture players’ active engagement while minimizing the influence of random events, often triggered by interactions with other players.

For instance, consider CS: GO, which includes an in-game shopping phase between rounds where players cannot move the camera or perform actions. Such a segment would provide limited insights into players’ play styles and is, therefore, a candidate for exclusion. It is important to note that the relevant time segment may differ from one game to another. When multiple segments are feasible, preference should be given to those occurring earlier in the match, as they can serve as more effective authentication triggers (the sooner, the better).

We clarify that while the framework recommends the extraction of such a segment(s) to assist the identification classifier, it remains flexible, allowing for the use of any match segment, even the entire match, based on user preferences.

IDENTIFICATION FEATURES ENGINEERING

This phase is dedicated to determining the most pertinent features that can effectively distinguish users’ play styles. This is a pivotal aspect of the entire framework, as it lays the foundation for the success of the machine learning-based approach.

We classify these features into two broad categories commonly found in video games: *states* and *actions*. The former encompasses features that are continually recorded and may be shared across various games (e.g., mouse position, avatar movement), while the latter pertains to actions executed at the player’s command, which tend to be more game-specific (e.g., shooting in CS: GO or casting spells in Dota 2).

In our case, we retained nearly all the features presented by the games to potentially filter out less relevant ones at a later stage. The feature selection can also vary depending on the framework user’s objectives. For instance, for basic assessments, utilizing player and camera movement data might suffice, whereas more domain-specific features should be included to enhance performances.

²For example, a typical Dota 2 game lasts around 45 minutes.

DATA AGGREGATION

In many instances, the sequences of gameplay data extracted can be excessively long. Indeed matches of some games can last for hours. Considering that most games maintain a sampling frequency of at least 30 frames per second, a single match can yield hundreds of thousands of states. To alleviate the computational complexity of this task, we adopted a best practice of data aggregation, reducing both the temporal and spatial computational load.

Following the guidance of a previous study [418] and considering our computational resources, we chose to aggregate the data to achieve a sampling interval of 0.5 seconds. This not only eases computational demands but can also enhance pattern identification, making it feasible to uncover patterns that might otherwise remain obscured within the extensive sequence points [418]. However, it is important to note that this step can be considered optional, depending on the available computational resources.

Moreover, the number of states to aggregate to achieve this desired frequency varies from game to game, contingent on both the specific game and the method used for state extraction, as different games operate with different states per second, referred to as *ticks* in our context. For instance, Dota 2 maintains 30 ticks per second, while CS:GO operates at 64 ticks per second.

MATCH SEGMENT SEQUENCING

As the relevant match segment can extend over several minutes, a crucial step involves subdividing this segment into shorter sequences. This division can significantly enhance the effectiveness of the identification process since the algorithm would have more sequences per player to learn from, and Recurrent Neural Networks (RNNs) usually work better with (relatively) short sequences [419]. Furthermore, using shorter sequences would facilitate rapid authentication. The ideal sequence length should be determined through a proper validation strategy. In our experiments, we adopted sequences of two minutes (see Section 9.6.2).

9.2.4 IDENTIFICATION PHASE

This phase is dedicated to the identification process, i.e., given in input a match m , determining the player $p \in \mathcal{P}$ who played it. To achieve this, we rely on Deep Learning (DL). Therefore, it is essential to adhere to all DL best practices, including preprocessing or model validation. Below, we provide a concise overview of the primary stages involved.

FEATURE PREPROCESSING

Before training the classifier, data preprocessing plays a crucial role in refining the dataset and enhancing training performance. The specific preprocessing steps often align with the chosen classification algorithm. In our scenario, we opted for standardization, which involves transforming the data with a mean of 0 and a standard deviation of 1. This standardization step usually leads to a smoother convergence during training. An important consideration here is that scaling should be applied with respect to the values within the training set to guard against potential overfitting.

RNN MODEL TRAINING

Prior to neural network training, it is essential to partition the dataset into distinct training, validation, and test sets. In our case, it's crucial to ensure that data from a single match remains exclusively within one of these sets to prevent overfitting. Our preferred split ratio, commonly employed in our experiments, is 80%-10%-10% for training, validation, and test data, respectively.

Selecting the appropriate deep learning model is equally critical. Given the temporal nature of our data, Recurrent Neural Networks (RNNs) are a natural choice, lauded for their ability to handle time series data [420]. We opted for an LSTM [421] to circumvent the vanishing and exploding gradient challenges often encountered in standard RNNs. Another compelling option is transformer networks, as introduced by Vaswani et al. in [422], where the attention mechanism plays a key role in predictive tasks.

9.3 TESTING THE FRAMEWORK: GAMES SELECTION

To assess the effectiveness of our framework, we chose two popular online video games with millions of monthly players worldwide: Dota 2 and Counter-Strike: Global Offensive (CS: GO). Our rationale for this game selection is detailed in Section 9.3.1. To provide readers with a foundational understanding of these games, Section 9.3.2 and Section 9.3.3 offer background information on Dota 2 and CS: GO, respectively. Lastly, in Section 9.3.4, we delve into the similarities and differences between these two games, shedding light on the challenges inherent in constructing our framework.

9.3.1 ONLINE GAMING PANORAMA

To test our framework, we focused on online multiplayer video games, primarily due to the accessibility of their publicly available data online and the inherent human interactions within them. As stated before and demonstrated by past literature [384, 383], these contexts often expose vulnerabilities to various cybersecurity threats. As possible candidates, we selected popular video games with millions of monthly players. Table 9.1 lists some of the candidates. In the table, we provide information on the following key aspects:

- **Tracking Website Availability:** These are valuable tools for gathering participant information and data;
- **Replay Availability:** This metric gauges the ease of accessing match replays for a given player.
- **Parser Availability:** This indicates the quality and accessibility of replay parsers, crucial for extracting game-related data, and accordingly, a player play style, from match replays.

From the table, we chose CS: GO and Dota 2 for our experiments primarily because of the availability of replays and reliable parsers. Additionally, our expertise in these two games factored into our selection process³. We want to point out that while the presence of replay parsers is advantageous, it is not an absolute requirement. In fact, it is possible to construct a replay parser from scratch if needed. Likewise, match replays can be recorded

³One of our authors currently ranks among the top 1% of Dota 2 players, while another possesses extensive experience in CS: GO.

Table 9.1: Table representing the most played online video games. The monthly players refer to May 2023.

	Monthly Players	Release Year	Free To Play	Tracking Websites	Replay Availability	Parser Availability
Minecraft	180,439,606	2011	No	Yes	Low	Low
LOL	152,973,464	2009	Yes	Yes	Medium	Low
Fortnite	234,546,282	2017	Yes	Yes	High	Medium
PUBG	288,859,492	2016	No	Yes	Medium	High
DOTA 2	14,571,704	2013	Yes	Yes	High	High
CS: GO	5,382,822	2012	Yes	Yes	High	High

Sources: [423][424][425][426][427][428]

through third-party applications. Nonetheless, having readily available parsers and recorded replays significantly facilitates the experiments within our framework. We now briefly explore the two video games and present their characteristics. Given their glaring differences, they serve as ideal test cases to evaluate the adaptability of the PvP framework across diverse video game environments.

9.3.2 DOTA 2

Dota 2 is a Multiplayer Online Battle Arena (MOBA) video game released by Valve Corporation in 2013 and available on Steam. It is currently one of their most played multiplayer game [429]. Two different teams composed of five players, called Radiant and Dire, fight each other to destroy the enemy’s base. Each player controls a single character, called a “hero”, with unique abilities. Through killing enemies, a hero gains experience and gold, that can be used to improve their status.

TRACKING WEBSITES Tracking Websites are web applications that automatically analyze players’ matches to produce individual statistics and general trends about the game. All the matches a gamer played are publicly exposed, along with the URLs to download them from Valve servers. The main tracking websites for Dota 2 are Dotabuff⁴ and Opendota⁵. Players can grant consent within the game client for these data-collecting websites. When a player chooses not to provide consent, their matches become invisible on tracking websites but remain accessible for retrieval.

REPLAY PARSING In Dota 2, a replay is an event stream that involves, for example, the orders players give to heroes and combat logs. In our experiments, we used the parser Clarity⁶, which is written in Java and is the one used by the tracking website Opendota.

9.3.3 COUNTER-STRIKE: GLOBAL OFFENSIVE

Counter-Strike: Global Offensive is a multiplayer, objective-based, First Person Shooter (FPS) released by Valve Corporation in 2012 and available for free on Steam. It is the fourth chapter of the Counter-Strike series.

⁴<https://www.dotabuff.com>

⁵<https://www.opendota.com>

⁶<https://github.com/skadistats/clarity>

The player controls a character that can move around on the map and must complete the given objective. The players are organized into two teams of five players each, called terrorist and counter-terrorist teams, which have opposite objectives.

In our experiments, we focused on the Bomb Defusal game mode, the most common one in the competitive scenario. In each round (up to 30), the terrorists' team must place an explosive in one of two predefined sites (identified as A or B) and make sure it explodes within a time limit, while the counter-terrorists must prevent the bomb from exploding. Figure 9.2 shows an example of a map. In total, there are nine maps, and we considered all of them in our experiments. Every player earns in-game currency based on single actions (such as kills) or team actions (e.g., clearing objectives). At the same time, negative actions (such as killing a teammate) lead to losing the in-game money. Players can use the currency earned in the previous rounds to buy weapons or other utilities for the current round.



Figure 9.2: An example of a game map. The two red areas, labeled A and B, are the two possible areas in which the bomb can be planted, and the two green areas are the spawn areas for the team. In this case, the upper one (nearer to the planting areas for the bombs) is the spawn for counter-terrorists, while the other one is for the terrorists' team.

TRACKING WEBSITES The most famous tracking websites for CS: GO are [csgostats](https://csgostats.gg/)⁷ and [HLTV](https://www.hltv.org/)⁸. HLTV permits downloading match replays directly from the websites. Therefore, we used it during our experiments. It is important to note that both these websites require the player to log into the website using the Steam account, in order for the website to synchronize the matches.

⁷<https://csgostats.gg/>

⁸<https://www.hltv.org/>

REPLAY PARSING In CS: GO, a replay carries pieces of information about the players' positions and actions in the form of logs. We used a parser called `demoinfo`⁹, written in C/C++ and developed by Valve, to access these information.

9.3.4 CONSIDERATIONS ON THE SELECTED GAMES

Dota 2 and CS: GO have many similarities and differences that make them two good representative games to test PVP. In listing them, we also give some key insights into the features that could help us develop a player "fingerprint".

First and foremost, both games are PC-based, utilizing mouse and keyboard input devices. In both Dota 2 and CS: GO, players maneuver avatars within the virtual world using a combination of mouse and keyboard inputs. While the specific in-game actions diverge significantly due to Dota 2's fantasy nature and CS: GO's realistic war-based setting, the frequency and style of these actions can be uniquely attributed to each player. For instance, a CS: GO player's precision in aiming with a sniper rifle or a Dota 2 player's attack frequency can be distinctive traits.

In both titles, players can explore the virtual world by moving their controlled character and manipulating the camera. However, Dota 2 relies on mouse clicks for character movement, while CS: GO utilizes keystrokes. Concerning the camera, Dota 2 provides a top view and allows players to freely navigate it across the map, whereas CS: GO employs a fixed first-person perspective camera that can be adjusted to survey the surroundings. Analogous to the real world, where gait patterns and visual patterns are considered biometrics [430, 431], the virtual world could present similar possibilities.

Furthermore, it is crucial to note that Dota 2 falls under the MOBA (Multiplayer Online Battle Arena) genre, whereas CS: GO is an FPS (First-Person Shooter). These genres inherently differ in gameplay and dynamics. Thus, if our PvP framework performs effectively in both titles, it suggests potential applicability across diverse scenarios. Notably, our experiments consider fundamental features that could be present in any game where a player controls a character, such as character movement and camera systems.

9.4 DATA COLLECTION

In this Section, we present how we collected the data we used in our work. We start with Dota 2 data in Section 9.4.1 and continue with CS: GO in Section 9.4.2.

9.4.1 DOTA 2

DOTA 2 ONLINE SURVEY

To collect data for Dota 2 player identification, we employed an anonymous online survey. Participants willingly consented to the use of their data for research purposes. During the survey, participants provided their Steam

⁹<https://github.com/ValveSoftware/csgo-demoinfo>

ID, a unique identifier that enabled us to locate them on tracking websites. Utilizing these IDs, we accessed their match history, the URLs to download the match replays, and statistics from the Opendota tracking website.

The survey was distributed across various platforms, including Facebook, Reddit, Telegram, and Discord, and typically took participants approximately 4 minutes to complete. As an incentive, we held a prize draw for in-game content.

Additionally, the survey allowed us to investigate the prevalence of issues like scams or harassment within the Dota 2 community. To ensure the validity of responses, participants needed to be visible on tracking websites, and their answers had to exhibit coherence. We implemented best practices developed over the years to validate responses and safeguard the privacy of participants.

SURVEY RESULTS We received a total of 625 responses; however, only 529 were deemed valid. Some responses were disqualified due to inconsistencies in the answers or because they originated from inactive players or individuals not visible on tracking websites. The respondents hailed from 62 countries, with the overwhelming majority (502) being male. Among the respondents, a significant portion (over 47%) identified as students, while others included workers, working students, and unemployed individuals. The age range spanned from 13 to 46, with the majority falling between 16 and 28 years old. Notably, over one-third of the participants declared having multiple accounts, accentuating the boosting and smurfing issues. The dataset we collected is the same presented in Chapter 10, where we demonstrate that our sample of participants is representative of the Dota 2 player population in both size and demographic distribution.

In terms of their encounters with scams and virtual harassment, over 15% reported falling victim to scams, with some experiencing multiple instances. Additionally, approximately 75% of the participants had been contacted by strangers, leading to suspicions of scam attempts at least once. Lastly, more than 40% had experienced harassment during their matches at least once. These results confirm the existence and gravity of malicious behavior within the Dota 2 ecosystem, underscoring the necessity for an identification system like PvP to combat these harmful activities.

BEST PRACTICES ADOPTED We hosted our survey on a website we created, providing a dedicated page to inform participants about the research's objectives, data privacy safeguards, and the intended use of their information. To mitigate the risk of fraudulent responses, we implemented a login system using Steam, based on the OpenID protocol¹⁰. This login process allowed us to retrieve participants' Steam IDs and cross-verify the provided data. However, some users perceived this login requirement as a potential phishing attempt, despite the link being posted on the research group's website. This observation underscores the genuine concerns surrounding online scams.

To assess participants' attention, we included two screening questions. Additionally, we included personal questions related to the game itself (e.g., the hours played per week), which could be readily verified using tracking websites. Furthermore, we considered the time taken to complete the survey; exceptionally short completion times indicated random or insincere responses and were thus disregarded. These measures collectively aided in identifying and filtering out incoherent or spurious responses.

¹⁰<https://steamcommunity.com/dev>

DOTA 2 DATASET CREATION

After acquiring players' IDs through Steam APIs and Opendota, we downloaded their matches from December 2019 to February 2020. Subsequently, we constructed the dataset following a three-step procedure. Initially, we opted to analyze a subset of players to accommodate computational constraints. Next, we identified the most pertinent features for characterizing play styles. Lastly, we delved into parsing the replays to extract these identified features.

PLAYERS SELECTION From our pool of 529 users, we downloaded an extensive set of over 30,000 matches, with an average duration of approximately 45 minutes, totaling around 22,500 hours. However, constrained by our computational resources, we narrowed our analysis to 50 players. These players were selected randomly but with the criterion that they had participated in at least 50 matches as Radiant and 50 matches as Dire, ensuring a balanced dataset. As a result, our final dataset comprised 50 players, each contributing 100 matches, amounting to a total of 5,000 matches.

FEATURES SELECTION To achieve player identification, we focused on general features that could potentially be applicable in other video games, enhancing the transferability of our research across different gaming contexts. Features directly tied to in-game metrics like gold or experience per minute were deemed too game-specific. Instead, we focused on characteristics such as cursor movements, camera movements (pertaining to a player's viewpoint), hero positions, and the available player actions. These attributes are likelier to be found in other video games.

In more detail, the cursor is defined by X and Y coordinates, while the camera is represented by a cell and a vector. Together, these components pinpoint the precise position, each defined along the X , Y , and Z axes. We disregarded the Z axis when considering cells, as it remains constant. A similar approach using cells and vectors is employed to represent hero positions. For actions, we exclusively incorporated the most common ones as features to reduce the problem's dimensionality. Some actions may also include X and Y coordinates for target locations. A comprehensive listing of feature types and values is provided in Table 9.2. The table also reports the aggregated features, which we better explain in Section 9.5.1.

REPLAY PARSING A replay is structured into ticks, with a rate of 30 ticks per second. At each tick, data related to entities (heroes, players, etc.) and players' commands are recorded. We utilized the parser to extract the aforementioned features at every tick, i.e., those associated with the cursor, camera, hero positions, and actions. Consequently, a replay transformed into a sequence of states encompassing cursor, camera, and hero positions, and player actions when occurring.

9.4.2 CS: GO DATA COLLECTION

Similarly to the Dota 2 data collection phase, we followed a three-step procedure to create our dataset.

Table 9.2: Initial set of Dota 2 features considered for the Identification task.

Type	Values	Aggregated Features
Cursor, Camera Cell, Hero Cell, Hero Vector Camera Vector	X,Y X, Y, Z	X_mean, X_std, X_changes Y_mean, Y_std, Y_changes X_mean, X_std, X_changes Y_mean, Y_std, Y_changes Z_mean, Z_std, Z_changes
Action: Move_to_position	occurred, X, Y	n_occurs, X_mean, X_std, Y_mean, Y_std
Action: Move_to_target	occurred	n_occurs
Action: Attack_move	occurred, X, Y	n_occurs, X_mean, X_std, Y_mean, Y_std
Action: Attack_target	occurred	n_occurs
Action: Cast_position	occurred, X, Y	n_occurs, X_mean, X_std, Y_mean, Y_std
Action: Cast_target	occurred	n_occurs
Action: Cast_target_tree	occurred	n_occurs
Action: Cast_no_target	occurred	n_occurs
Action: Hold_position	occurred	n_occurs
Action: Drop_item	occurred, X, Y	n_occurs, X_mean, X_std, Y_mean, Y_std
Action: Ping_ability	occurred	n_occurs
Action: Continue	occurred	n_occurs

PLAYER SELECTION Our initial step involved selecting players for identification. In this instance, we relied on the HLTV Tracking Website ¹¹, which conveniently offers the option to download matches directly from the platform. Similar to our approach in Dota 2, we randomly chose players who had participated in a minimum of 100 matches. Subsequently, we downloaded their matches within the timeframe of the previous year.

To streamline this process, we automated match downloads using the Python library Selenium ¹². We want to clarify that, unlike the Dota 2 case where we obtained explicit consent from the players, our approach with HLTV relied solely on publicly available data. We did not engage in any efforts to uncover their identities or extrapolate any private information.

FEATURES SELECTION We identified a set of features that, based on our knowledge and intuition, could prove valuable for our identification task. In this context, we continued to extract player positions and camera viewpoints. In CS: GO, a player’s position is defined relative to the map’s (0, 0, 0) coordinate, and their speed is measured along the three axes (X, Y, and Z). The camera view is represented by angles that capture the player’s field of vision. These angles, comprising a pair in a 3D polar coordinate system, are particularly noteworthy because they remain consistent across all users, being scaled within the [0, 360) interval and independent of screen size variations.

Concerning player actions, we chose to retain various action types, reserving the decision about which actions to utilize for model training. Notably, the only action that required distinct consideration was ‘crouching.’ CS: GO lacks a specific crouch action; instead, it manifests as a variation in the player’s Y-axis position, from which we can infer the crouching state. Table 9.3 reports a comprehensive list of the selected features.

¹¹<https://www.hltv.org/>

¹²<https://www.selenium.dev/>

Table 9.3: List of parsed features for CS: GO

Type	Features	Type	Features
Camera	X, Y	weapon_zoom_rifle	occurred
Player position	X, Y, Z	player_falldamage	occurred, damage
door_moving	occurred	molotov_detonate	occurred
player_blind	occurred, blind_duration	tagrenade_detonate	occurred
round_mvp	occurred	hegrenade_detonate	occurred
defuser_pickup	occurred	flashbang_detonate	occurred
bomb_pickup	occurred	item_purchase	occurred, item_purchased
defuser_dropped	occurred	ammo_pickup	occurred
bomb_dropped	occurred	silencer_detach	occurred
bomb_abortdefuse	occurred	decoy_detonate	occurred
bomb_begindefuse	occurred	smokegrenade_detonate	occurred
bomb_abortplant	occurred	weapon_fire	occurred, weapon
bomb_beginplant	occurred	weapon_reload	occurred
bomb_defused	occurred	player_jump	occurred
bomb_planted	occurred	player_death	occurred, type (kill, death, or assist)
bullet_impact	occurred	item_pickup	occurred, item_picked
weapon_zoom	occurred	item_equip	occurred, item_equipped

REPLAY PARSING After downloading the replays, we parsed them to extract the aforementioned features. We chose the open-source parser `demoinfo`¹³, developed by Valve. Regrettably, we encountered a challenge due to the absence of comprehensive documentation. Consequently, we invested considerable effort in conducting experiments to decipher its functionality. The parser operates on protobuf messages¹⁴ and functions as a message streamer. It leverages data structures known as “Tables” to distinguish between entities (representing the players’ controlled characters) and actions (executed by the players). To fulfill our feature extraction requirements, we introduced modifications to the parser, enabling us to capture all the identified features.

9.5 FRAMEWORK TESTING ON DOTA 2 AND CS: GO

In the previous section, we showed the game-play features we selected to uniquely identify a player, i.e., to create their “fingerprint”. We now test the rest of our identification framework PvP on Dota 2 in Section 9.5.1 and CS: GO in Section 9.5.2.

9.5.1 DOTA 2

MATCH ENGINEERING PHASE

RELEVANT MATCH SEGMENT EXTRACTION A Dota 2 match typically involves a high level of randomness, which depends on many factors, such as the hero choices, itemization, or players’ interactions. According to the framework’s procedure (Section 9.2.3), we mitigated this problem by extracting a relevant match segment.

¹³<https://github.com/ValveSoftware/csgo-demoinfo>

¹⁴<https://github.com/protocolbuffers/protobuf>

We started by excluding the pre-match phase, during which players select their heroes and purchase starting items. This phase imposes significant limitations on players, such as the inability to move the hero or manipulate the camera. Subsequently, we identified a suitable identification window as the initial 10 minutes of the match. During this phase, players usually remain in the same region of the map while still exhibiting distinct behaviors. Moreover, using this match segment is convenient because it is very rare that the game ends before 10 minutes, and this time-frame remains reasonable for player identification, with shorter identification times being preferable. Following this principle, we reduced each match to its first 10 minutes.

DATA AGGREGATION As anticipated in the framework description, we aggregated data to reduce computational costs and extract high-level patterns. Specifically, we aggregated our data to obtain a datapoint every 0.5 seconds. For spatial attributes, we retained the average and standard deviation values along each axis. Regarding player actions, we preserved the frequency of occurrences along with the mean and standard deviation of relevant axes, where applicable. This aggregation transformed each replay into a sequence of states recorded at 0.5-second intervals, culminating in a total of 61 aggregated features. These features are detailed in Table 9.2.

MATCH SEGMENT SEQUENCING Since the best sequence length is unknown a priori, we decided to set a maximum time for identification of 2 minutes. Therefore, we extracted non-overlapping 2-minute sequences from the reduced 10-minute matches, forming a dataset of 25000 sequences (5 sequences per match * 100 matches * 50 players), each with 240 data points (120 seconds/0.5 seconds). These sequences were individually labeled according to the player responsible for creating them.

TRAINING, VALIDATION, TEST SETS SPLIT The dataset was partitioned into three subsets for training, validation, and testing, with a distribution of 80%, 10%, and 10%, respectively. To prevent overfitting and maintain data integrity, we ensured that each match sequence remained within the same subset during the splitting process.

IDENTIFICATION PHASE

FIRST MODEL Based on our data, the identification challenge can be reformulated as a sequence classification problem, with each class representing a player within our player pool. For the identification algorithm, we opted for a Recurrent Neural Network (RNN), specifically employing a Long Short Term Memory (LSTM) architecture. LSTMs excel at capturing and retaining temporal dependencies, making them highly suitable for our task. Notably, they are adept at learning time-dependent patterns, such as how a player manipulates the mouse or camera, effectively serving as the “fingerprint” we seek.

The employed model comprises two LSTM layers (64 neurons each) utilizing the *tanh* activation function, followed by a fully connected layer (64 neurons) with *RELU* activation. The output layer consists of a *softmax* layer with 50 neurons, each corresponding to one of the players. We employed a standard batch size of 256, conducted training over 100 epochs, and utilized the categorical crossentropy loss function. Our chosen optimizer was Adam, a well-established optimizer in machine learning [432], configured with a learning rate of 0.001, $\beta_1=0.9$, and $\beta_2=0.999$. This model was implemented using Keras with a Tensorflow backend. During training, we standardized our data to achieve faster convergence. As a performance metric, we use accuracy, given our dataset is perfectly balanced.

The model we have outlined is a versatile baseline, suitable for preliminary investigations, and as a foundation for more sophisticated models. Framework adopters have the flexibility to select or design more complex models to achieve superior performance.

FEATURE ELIMINATION The first run of the model did not perform well. This most likely was due to the noise that some of the features introduced. To solve this issue, from the features listed in Table 9.2, we removed the features Hero_Cell and Hero_Vector, and the X and Y coordinates from Attack_move, Cast_position, Drop_item.

MODEL SELECTION

Using the reduced set of features, the model finally started to learn, reaching an accuracy of 90%. We then proceed to optimize hyperparameters through a grid-search approach. We explored the following parameters:

- First LSTM layer neurons: [64, 128, 256];
- Second LSTM layer neurons: [64, 128, 256];
- Dense layer neurons: [64, 128, 256];
- Optimizer: [SGD, SGD Nesterov, RMSprop, Adagrad, Adadelata, Adam];
- Learning rate: [1, 0.1, 0.01, 0.001];
- Dropout after LSTM layers: [None, 0.2];

The model that exhibited the best performance on the validation set (accuracy = 96.48%, loss = 0.179) featured 256 neurons for both LSTM layers and 128 neurons on the dense layer, utilizing Adam as the optimizer with a learning rate of 0.001. The accuracy and loss of the model are depicted in Figure 9.3.

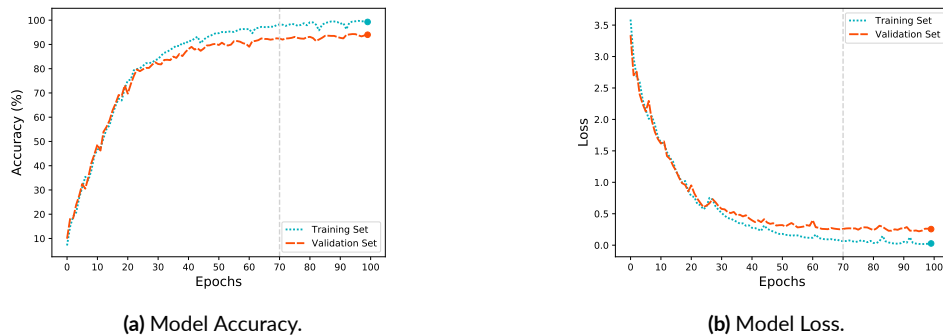


Figure 9.3: Training and validation accuracy and loss in Dota 2 identification model.

Subsequently, we retrained the model using both the training and validation sets and assessed its performance on the test set. This final evaluation yielded an accuracy of 96.32% and a loss of 0.198. This outcome underscores the exceptional generalization capabilities of our model, which demonstrates its ability to achieve highly accurate player identification.

9.5.2 CS: GO

MATCH ENGINEERING PHASE

RELEVANT MATCH SEGMENT EXTRACTION Following the PvP framework’s guidelines, we commenced by identifying a relevant match segment for player identification (Section 9.2.3). In CS: GO, a match begins with a pre-match phase during which players can access the in-game shop. Notably, during this phase, players remain stationary, and camera movement is restricted, even if the player moves the mouse. This presented a potential issue for identification, as all players exhibited limited activity, so we removed the information from the pre-match phase. Then, for similar reasons as in Dota 2, we designated the first 10 minutes of each match as the suitable match segment for player identification. Hence, we reduced each match to its initial 10 minutes.

AGGREGATION As outlined in the framework description, we adopted data aggregation techniques to alleviate computational demands and uncover higher-level patterns. Initially, we encountered occasional data extraction failures by the parser, occurring in less than 1% of the ticks. To address these errors, we performed data interpolation by computing average values between the ticks immediately before and after the erroneous data point. Following interpolation, we proceeded with data aggregation.

Much like in Dota 2, our data aggregation involved reducing the data granularity to establish data points at 0.5-second intervals, equivalent to 32 ticks. For the states, we retained statistical measures, including the mean value, standard deviation, and changes (both positive and negative) in the values. In the case of actions, we retained only the count of occurrences.

SPLITTING THE DATA After data aggregation, we split the data into training, validation, and test set, using an 80%-10%-10% split for training, validation, and test, respectively. To mitigate overfitting and preserve data integrity, we ensured each match sequence remained within its respective subset during the data splitting process.

IDENTIFICATION PHASE

FIRST MODEL We employed the same initial model architecture as in Dota 2, featuring an LSTM-based identification algorithm. This model consisted of two LSTM layers, each with 64 neurons and tanh activation functions, followed by a fully connected layer with 64 neurons using the *ReLU* activation function. The output layer was configured with a *softmax* activation and comprised 50 neurons, corresponding to the number of players. The training was conducted with a batch size of 128 for 100 epochs, implementing early stopping (5 epochs) for model convergence. We utilized the categorical cross-entropy loss function and employed the Adam optimizer with a learning rate of 0.01. The implementation was carried out using the Keras¹⁵ library with a Tensorflow¹⁶ backend. During training, we standardized our data to achieve faster convergence. As a performance metric, we use accuracy, given our dataset is perfectly balanced.

¹⁵<https://keras.io/>

¹⁶<https://www.tensorflow.org/>

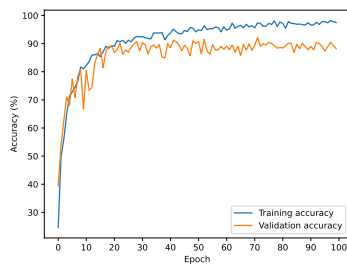
FEATURE ELIMINATION With this model, we achieved approximately 75% accuracy on the validation set. While this result was promising, it fell short of the performance obtained in Dota 2. Likely, certain features led to confusion, as they might have exhibited similarity across players or substantial variability within the same player, particularly concerning map-dependent factors. For example, the player’s mean position along the x, y, and z axes could vary significantly based on the map. Regarding actions, we observed that some parsed actions, such as `door_moving` or `player_jump`, were rarely triggered. Therefore, we excluded the aforementioned actions, and the players’ mean position features from our model to reduce noise introduced by map-dependent variability. Table 9.4 lists the final features utilized in our model.

Table 9.4: Features we kept after the feature selection process

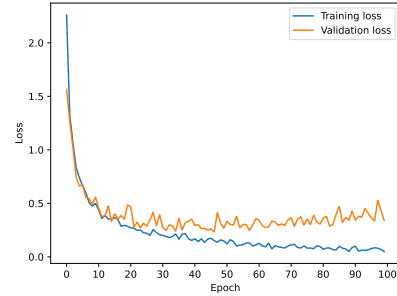
Type	Values	Aggregated Features
Cursor position	X, Y	Mean, standard deviation, changes in positive, changes in negative
Player position	X, Y, Z	Standard deviation, changes in positive, changes in negative
Player velocity	X, Y, Z	Mean, standard deviation, changes in positive, changes in negative
Crouch changes	occurred	n_occurs
Weapon fire	occurred	n_occurs
Weapon reload	occurred	n_occurs
Player jump	occurred	n_occurs
Kills	occurred	n_occurs
Assists	occurred	n_occurs
Item pickup	occurred	n_occurs
Item equip	occurred	n_occurs
Grenade detonate	occurred	n_occurs
Smoke grenade detonate	occurred	n_occurs
Molotov detonate	occurred	n_occurs
Flashbang detonate	occurred	n_occurs
Tagranade detonate	occurred	n_occurs
Hegranade detonate	occurred	n_occurs
Decoy detonate	occurred	n_occurs
Player fall damage	occurred	n_occurs
Ammo pickup	occurred	n_occurs
Silencer detach	occurred	n_occurs

MODEL SELECTION With these new features, we obtained a better-performing neural network, reaching around 80% accuracy in the validation set. To find the best possible model, we applied a grid search approach, trying the same parameters for the Dota 2 case (Section 9.5.1). The best model on validation achieved 90.39% accuracy in the validation, and had 256 neurons in the first *LSTM* layer, 128 in the second one, and 64 in the fully connected layer. The LSTM layers had both a dropout of 0.2. The best optimizer was Adam, with a learning rate of 0.001. Figure 9.4 shows the accuracy and the loss of our model during training and validation.

We then retrained the top-performing model on the combined training and validation sets, resulting in a final accuracy of 91.68%. Once again, our model exhibited impressive player identification accuracy, underscoring the effectiveness of the PvP framework.



(a) Model Accuracy.



(b) Model Loss.

Figure 9.4: Training and validation accuracy and loss in CS: GO identification model.

9.6 FURTHER EXPERIMENTS

For both Dota 2 and CS: GO, we achieved player identification accuracy exceeding 90%, illustrating the versatility of the PvP framework across various video games. Subsequently, we conducted additional experiments to delve deeper into our identification results. In Section 9.6.1, we assessed performance using solely general features commonly found in video games. Section 9.6.2 explores how identification performance varies with sequence length, while Section 9.6.3 investigates whether the timing of sequence selection influences performance. Finally, in Section 9.6.4, we present two game-specific case studies: one focusing on identifying unknown players in Dota 2, and the other examining the impact of the number of players on identification in CS: GO.

9.6.1 GENERAL FEATURES EVALUATION

In both of our case studies, we employed game-specific features, such as spellcasting in Dota 2 and shooting in CS: GO. While these features can be substituted with other game-related attributes in most games, it is valuable to explore how PvP performs using only general features commonly found in various games. For this purpose, we identified cursor movement, camera movement, and character movement as general features, as they are prevalent in nearly all video games where players control characters. Consequently, we retrained our top-performing Dota 2 model using solely `cursor movement`, `camera movement`, and `move_to_position` actions. In the case of CS: GO, we utilized only `cursor position` (camera features), `player position` (excluding the mean value), and `player velocity` (along all three axes).

Table 9.5 presents the results obtained using these general features. CS: GO experienced a slight decrease in accuracy of approximately 6%, while Dota 2 exhibited a mere 1% reduction. This disparity can be attributed to the substantial differences between the two games, with actions playing a more significant role in CS: GO identification. Nevertheless, both results remain notably high, underscoring the versatility of the PvP framework and its potential applicability in a wide range of games, particularly those involving character control.

Table 9.5: Test Accuracy using all features and general features for CS: GO and Dota 2 identification.

	CS: GO	Dota 2
All Features	91.68%	96.32%
General Features	85.83%	95.6%

9.6.2 SEQUENCE LENGTH EVALUATION

For both games, we employed 2-minute sequences for player identification. Now, we explore the viability of using shorter sequences. To investigate this, we divided the 2-minute sequences into smaller sub-sequences of 10, 20, 30, 40, and 60 seconds, while ensuring that these sub-sequences remained within the same set as the original sequences. The accuracy of the models for both games is visualized in Figure 9.5. Additionally, we have included results for 120-second sequences to provide a point of comparison.

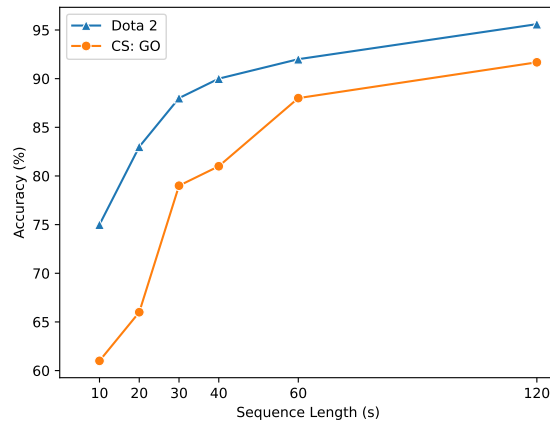


Figure 9.5: Identification accuracy for different sequence lengths.

Results for both games reveal a positive correlation between sequence length and model performance. Accuracy steadily improved, with CS: GO ranging from around 60% for 10-second sequences to 87.4% for 60-second sequences. In contrast, the Dota 2 model exhibited superior performance, starting at approximately 75% accuracy with 10-second sequences and reaching 91% with 60-second sequences. Thus, our findings suggest that identifying a unique play style becomes more reliable in longer sequences. In shorter sequences, players may spend time standing still, and if there is minimal or no movement, the sequences can exhibit significant similarity to one another. While this conclusion holds, it remains desirable to develop alternative models to excel even with shorter sequences.

9.6.3 SEQUENCE PICKING INTERVAL EVALUATION

In this experiment, we aim to discern whether specific time intervals within the 10-minute duration hold greater significance than others, influencing player identification performance. We aim to determine, for instance, whether

the initial two minutes are better to identify a player compared to the final two minutes. To investigate this, we retrained our best model using only the sequences occurring within a two-minute timeframe, i.e., from 0 to 2 minutes, from 2 to 4 minutes, and so on. Given the reduced number of samples in the dataset, we employed a 5-fold cross-validation to assess the model's performance.

The results obtained for both Dota 2 and CS: GO are presented in Figure 9.6. Note that sequences are indexed as follows: sequences with index 1 commence at the beginning of the match, those with index 2 initiate after 2 minutes, and sequences with index 3 begin at the 4-minute mark, continuing in this pattern.

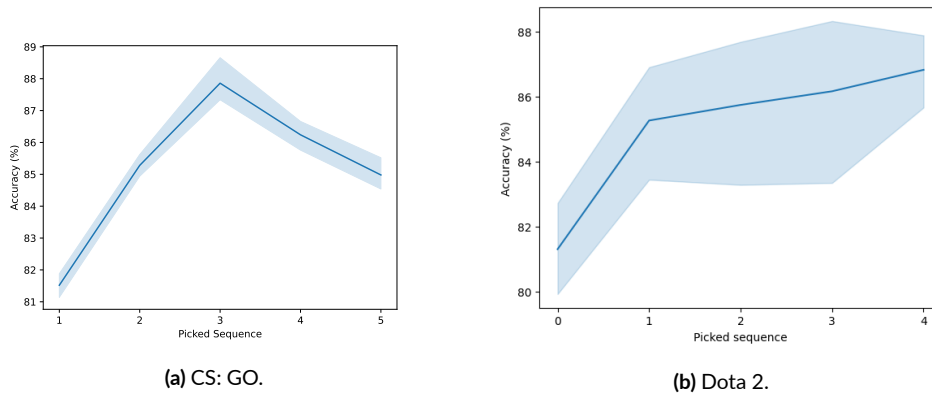


Figure 9.6: Accuracy with sequences picked at different 2-minute time windows.

Observing the figures, it's evident that regardless of the picking time, the accuracy slightly diminishes compared to using all available sequences. Nonetheless, this outcome holds significance as it implies that we can effectively train our classifier with a reduced dataset while still achieving approximately 85% accuracy. Examining both games, we note that the least favorable results manifest within the first two minutes of gameplay. This trend can be attributed to the fact that, at the start of a match, players have not yet had the opportunity to showcase their unique gameplay characteristics, as each player on the team begins from an identical position. As time progresses, we witness a notable improvement in performance, with the peak occurring between the 4-6 minute window for CS: GO and between the 8-10 minute window for Dota 2.

A plausible explanation for Dota 2 could be that, around the 10-minute mark, players begin to roam more across the map, increasing the likelihood of their unique play style emerging. In CS: GO, players may become more active after an initial examination of the situation, thus showcasing their distinct gameplay.

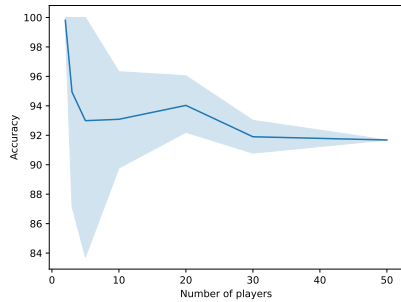
9.6.4 GAME-SPECIFIC ADDITIONAL CASE STUDIES

To test the framework even further, we conducted two additional case studies, one per game.

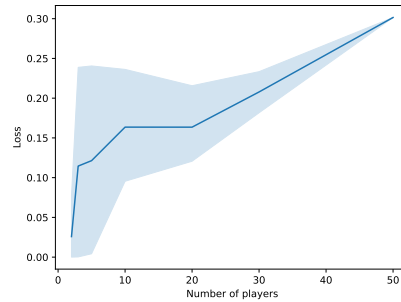
CS: GO NUMBER OF PLAYERS EVALUATION

Intuitively, as more players need to be distinguished, the task becomes more complex. To test this hypothesis, we varied the number of players to be identified, varying it from 2, 3, 5, 10, 20, and 30. For each of these player-count

scenarios, we conducted five training runs, randomly selecting participants each time. The resulting Accuracy and Loss metrics are presented in Figures 9.7a and 9.7b, respectively. Additionally, we provide results using the entire participant pool (50 players) for comparison.



(a) Accuracy with different number of players



(b) Loss with different number of players

As expected, the results reveal that as the number of players increases, the model’s accuracy decreases. This outcome aligns with our predictions, as a larger player pool increases the likelihood of finding two players with similar play styles. Furthermore, it is more probable that players, as they gather information about opponents’ positions, begin to coordinate their movements toward common objectives. Consequently, players may exhibit more similar movement patterns under these circumstances. Interestingly, there is no significant decrease in accuracy between 30 and 50 players, suggesting that our model may perform effectively even with more than 50 players.

DOTA 2 UNKNOWN PLAYERS EVALUATION

In Dota 2, our Neural Network was able to classify a player with very high accuracy. However, we wanted to study how it could face unseen players, i.e., not used in the training phase. Neural Networks do not perform very well with open-set problems, and there is still no very effective solution [433]. To evaluate this case, we adopted two approaches: using a background class and using a threshold for the last predictive layer.

In the first scenario, we introduced an “unknown” background class comprising 45 previously unseen users, each with two matches. Additionally, five other previously unseen users were included in the test set. To maintain class balance, the background class was designed to have the same number of sequences as any other class. While this setup yielded a high accuracy of 93%, it was primarily attributed to the classifier’s strong performance in the original scenario. Indeed, upon closer examination of the confusion matrix (Figure 9.8), we observed that only four sequences from the new users were correctly classified as “unknown” (class 50). The remaining sequences were distributed among other players. This outcome is understandable since the unseen users’ data are still Dota 2 match sequences. Consequently, the network likely attempted to find the most similar player for the unseen sequences, leading to their misclassification.

In the second scenario, we adopted the rationale that if the model had never encountered a particular user, sequences composing one of their matches might be attributed to many players. To test this hypothesis, we provided the network with the five sequences composing a match and examined the network’s output for each sequence.

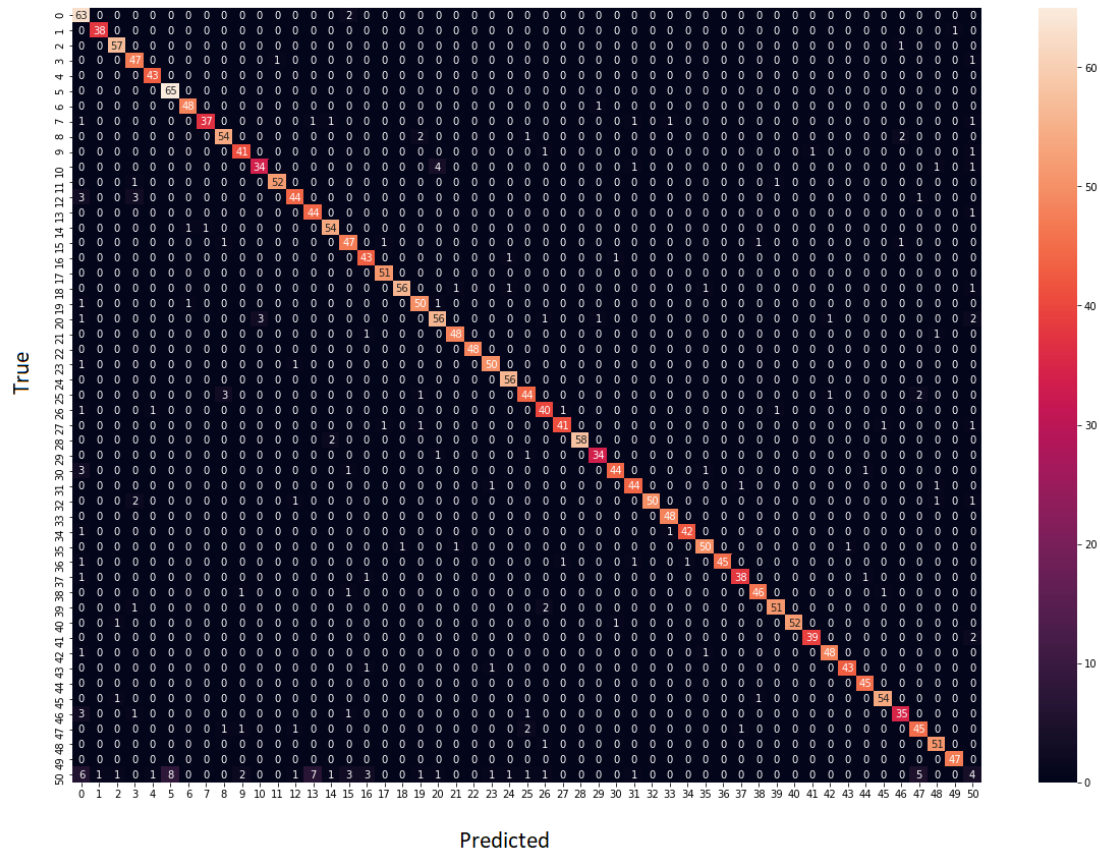


Figure 9.8: Confusion Matrix with Unknown Players involved (class 50).

We decided to assign a player to a match only if at least four out of the five sequences were allocated to the same player. We employed 50 previously unseen players for this experiment, each with two games of five sequences (as in the other experiments).

Out of the 100 games analyzed, only 28 were (mistakenly) attributed to the same known players, resulting in a 72% success rate. Moreover, when considering both games together (i.e., assigning two matches to a player only if both matches were assigned to the same player), only 8 previously unseen players were (mistakenly) assigned to a known player, yielding an 84% success rate. While the system could not distinguish unknown users from known ones with extremely high accuracy, these results indicate the potential to mitigate misclassifications. We intend to explore more effective solutions in our future work.

9.6.5 OVERALL COMPARISON OF THE TWO GAMES USING OUR FRAMEWORK

In our experiments, we demonstrated the effectiveness of our PvP identification framework in two vastly different games: CS: GO and Dota 2. Overall, the experiments in Dota 2 produced superior results, but CS: GO still exhibited very high performance, achieving accuracy rates above 90%.

We employed similar and generalized methodologies for both games to enhance the framework’s versatility. The primary divergence between the case studies lay in the distinct feature engineering approaches necessitated by the inherent differences between the two games. Nevertheless, even when considering only highly general features, we consistently achieved remarkable accuracy levels of around 90%. The framework’s success in these two diverse games leads us to speculate that player identification may be feasible in numerous other video games, particularly those within the MOBA or FPS genres. In both cases, our models effectively learned the unique play styles of gamers, akin to identifying a fingerprint. In the following section, we delve into the implications of our findings.

9.7 DISCUSSIONS

We now discuss the implications our work in Section 9.7.1, the persistence of the identification framework in Section 9.7.1, and the applicability to other games in Section 9.7.2.

9.7.1 REAL WORLD IMPLICATIONS

Being able to identify a player through their play style carries several significant implications.

IMPROVING THE GAMING EXPERIENCE First and foremost, the gaming community can derive significant advantages from our research. Current challenges such as smurfing, boosting, and account selling persist in the realm of video games [434]. To elaborate, “smurfing” refers to the practice where an experienced or highly skilled player creates a new account to be matched against inexperienced opponents, securing easy victories. Unfortunately, this disrupts the gaming experience for players in lower skill brackets. “Boosting” shares similarities with smurfing, wherein a less skilled player pays a highly skilled player to elevate their account’s ranking, resulting in a series of disrupted matches during the boosting process. Account selling can lead to both smurfing and boosting-related issues.

Our system empowers game administrators to impose stringent penalties on smurfs and booster players. By identifying players through their distinctive play styles, we can identify the primary account of smurfs or boosters and apply bans accordingly. Typically, high-rank players invest significant hours and resources into their primary accounts, making a ban on their main account a potent deterrent. Additionally, selling an account typically results in a drastic shift in a player’s play style from one match to another. Our system can easily detect such irregularities. Furthermore, players often create smurf accounts with the intention of selling them at a high price, and our system is equipped to counter this practice effectively.

REDUCING HARMFUL ACTIVITIES Another prominent issue in online video games is the prevalence of scammers [435], as corroborated by our Dota 2 survey results. Typically, scams initiate when a scammer adds a potential victim to their friend list. If the victim accepts the request, they engage in conversations with the risk of falling victim to a scam. Normally, players are cautious about accepting friend requests from new or low-ranking profiles. However, this vigilance tends to wane when the request comes from an average-ranking player after playing a game together. This implies that scammers often need to be active players within the game to increase their chances of success, or they resort to purchasing accounts to carry out their deceitful schemes. As previously demonstrated, our system can help curb account sales, and in cases where a scammer is reported, we can ban them from all associated accounts they play on.

More severe problems, such as harassment, cyberbullying, and stalking, are prevalent in online video games [383, 384]. As indicated in our survey, nearly 50% of players have experienced harassment within Dota 2 at least once. Bullies and stalkers often persist in pursuing their victims, even after their accounts are banned. Leveraging our system, it becomes possible to ban all accounts associated with these individuals by identifying their distinctive play styles. As previously mentioned, these individuals must actively participate in matches or engage with other players to conduct their harmful activities, making our developed system a valuable tool in mitigating this problem.

INCREASING MALICIOUS USERS CAPABILITIES. Regrettably, the system could potentially have adverse consequences by enabling malicious players to effectively pursue their victims and identify them even if they create new accounts. Additionally, there is the risk of exacerbating a more serious concern, such as pedophilia or grooming [436].

While games like Dota 2 or CS: GO may not typically cater to a young audience, other titles like Minecraft or Fortnite have garnered a significant following among children. It is common for both children and their parents to share the same gaming accounts for added safety. However, when parents are absent, children might be left to play alone and develop their unique play styles. If a pedophile were to discern that a child is playing — perhaps by recognizing a very young voice in voice chat — they could obtain match replays and analyze the gameplay data. When the child later plays alongside their parents, their play style would markedly differ, making it less likely for the predator to launch an attack. In essence, the pedophile could potentially identify when a child is playing alone, increasing the risk of their nefarious intentions succeeding.

Lastly, we must consider the substantial privacy concerns that arise with such a system. Many games allow players to log in using their social network accounts. Employing some form of transfer learning, it becomes theoretically possible to track a player across various games until one connected to a social network is identified. At that point, all associated information becomes vulnerable, significantly heightening privacy risks. We remark that, while our intention was never to provide tools to malicious users, acknowledging the existence of these vulnerabilities serves as the initial step toward developing effective countermeasures.

One effective countermeasure for potential victims is to restrict third-party websites, like tracking platforms, from collecting their data, substantially raising the bar for attackers seeking to identify them. Additionally, individuals can proactively vary their play styles by experimenting with different characters or by adopting novel strategies. Similarly, they may change their gaming equipment frequently, reducing potential biases that might aid in player identification.

A NEW AUTHENTICATION SYSTEM. An intriguing implication, supported by the recognition of play style as a form of biometric data, is the potential use of the system as an authentication method. This authentication approach could prove invaluable in preventing account theft, often facilitated through scams, or for a wide range of actions that typically require authentication. For instance, in the event of a forgotten password, users could be prompted to engage in a two-minute gameplay session to reset it. This authentication method could extend to safeguarding in-game purchases, permitting transactions only after a player has participated in at least one game during the current session. In cases where an account is stolen, there is a high likelihood that the thieves may exploit the victim’s linked credit card to make in-game purchases. Our system could introduce an additional layer of security to protect users in such scenarios. Furthermore, this system could pave the way for the creation of “protected” platforms where access is restricted to previously enrolled users. Instead of relying on potentially vulnerable passwords, authentication would hinge on the recognition of a user’s unique play style.

SYSTEM PERSISTENCE An important question arising from our work concerns the persistence of a player’s play style over time. While further research on this topic is warranted, we believe that both player improvement and hardware changes pose minimal threats to the concept of play style as a biometric identifier. In the case of player improvement, this phenomenon tends to occur gradually and continuously over time. Our model can be designed with this in mind, continuously updating itself with new match data. Indeed, the model can incorporate continual learning techniques, ensuring that a player’s gradual improvements are seamlessly integrated into the learning mechanism.

Concerning hardware devices, console controllers are typically standardized, making replacements straightforward. Mice and keyboards, on the other hand, are known for their durability, often boasting longevity of up to 60 million clicks¹⁷. As a result, hardware changes are rare. Moreover, we argue that the movement of a player’s avatar and the number of actions performed are not dependent on the specific input device, further supporting the stability of play style identification.

9.7.2 APPLICABILITY TO OTHER GAMES

Starting from our good results in two *very different* games, it becomes apparent that the concept of play style as a biometric identifier can potentially apply to multiple video games. While our experiments were confined to two specific games, our belief in this hypothesis has solidified. This high confidence is attributed to the prevalent feature found in the vast majority of video games, namely, the ability for players to control an avatar, manipulate their visual perspective, and perform in-game actions. These fundamental features transcend various game genres and mechanics.

Looking at the popular games listed in Table 9.1, League of Legends (LoL) bears similarities to Dota 2, while Fortnite and PUBG share similarities with each other and belong to the same shooter genre as CS: GO. Minecraft, in contrast, stands out as a unique experience. Despite these differences, all these games share the common thread of allowing players to control an avatar and manipulate the camera, albeit through different mechanics such as first-person or third-person perspectives. This common characteristic can manifest through features that closely align with those extracted using our framework. As suggested by our experiments in Section 9.6.1, these general features

¹⁷<https://steelseries.com/gaming-mice>, Accessed on: 30 May 2023.

can serve as a valuable baseline for identification, and their effectiveness can be further enhanced by incorporating specific game-related features to optimize performance.

9.8 CONCLUSIONS AND FUTURE WORK

In this chapter, we introduced PvP, an innovative framework designed for user identification within online video games. PvP leverages the distinct play styles of gamers as digital fingerprints to uniquely identify them across various accounts. By harnessing in-game data and employing deep learning techniques, PvP achieved over 90% accuracy on games from diverse gaming genres, such as Dota 2 and CS: GO. While this achievement holds the potential to reduce malicious activities, PvP can also be a new tool in the hands of cyber-criminals. Therefore, our aim is also to raise awareness of threats potentially affecting millions of gamers.

This work paves the way for numerous future research avenues. First, PvP should be tested on a larger pool of players, to assess its performance degradation. Additionally, the development of new deep learning algorithms, such as employing twin neural networks for player verification, could further enhance its generalizability. Similarly, improving identification accuracy in the case of unknown players or using shorter sequences is auspicious. Moreover, exploring PvP's applicability in gaming genres beyond MOBA and FPS offers another intriguing avenue for investigation. Lastly, examining whether a user's play style can transfer across different games, thereby enabling identification with minimal training across a wider range of video games, presents an exciting area for future exploration.

ETHICAL CONSIDERATIONS

In this study, we exclusively utilized publicly available data. All our experiments strictly adhere to the Menlo report guidelines [141]. We maintain the utmost respect for the privacy and security of Dota 2 and CS: GO users, ensuring that our research poses no risks, such as drawing undue attention or attempting to unveil their real-world identities. While our investigation acknowledges the potential dual-use nature of PvP as a tool in the hands of cybercriminals, our primary objective, as cybersecurity researchers, remains firmly focused on promoting awareness of potential threats within the gaming community and developing robust countermeasures to safeguard gamers.

10

Attribute inference attacks in online multiplayer video games: A case study on Dota 2

More than 3 billion people played Video Games (VG) in 2021, whose industry is constantly expanding, attracting new players every day [437]. A recent study highlighted that over 71% of participants increased their playtime, and that VG improved their well-being [438]. Within the broad VG landscape, one category stands out: *online multiplayer VG*. These VG allow players to interact with each other in a ‘controlled’ environment (i.e., the game) that is separated from their private life [439]. Specifically, players can interact in two distinct settings: cooperative or competitive. This chapter focuses on the latter, motivated by the rise of the Electronic Sports (E-Sports) panorama, which generated over \$1B of revenues in 2021 [437].

In E-Sports, players compete in VG matches [440]. Notable examples of E-Sports VG are Fortnite, ApexLegends, CS:GO, and DOTA2. All such VG are addictive (on average, DOTA2 players have over 1600 hours of playtime), and have an heterogeneous playerbase [441]. Some individuals “play for fun”, e.g., to spend their free-time with friends, or to entertain their audience on streaming platforms [442]. Others, however, “play to win”, and their primary aim is improving so that they can participate in (and, perhaps, win) one of the many competitions held regularly. Such competitions have rich prize-pools (up to \$40M [443]) which attract thousands of contestants. Indeed, winning matches is difficult due to the highly competitive environment (which is ultimately a zero-sum game [444]), and ‘mastering’ an E-Sport VG requires constant dedication [445].

Several resources, typically referred to as Tracking Websites (TW), were born to track players’ activities on a specific VG. Indeed, an intuitive way to improve is learning from past mistakes, and TW greatly facilitate such process by providing their users (i.e., the players) with detailed statistics of their in-game performance. We provide a screenshot of a TW focused on DOTA2 in Fig. 10.1, showing an overview of the in-game activities of the player

“Dendi”. Such statistics include, e.g., data of past matches, the days in which the player is more active, their friends; additional information is available by navigating the webpage. Despite the undeniable advantages provided by TW (over 70M of DOTA2 players use TW [446]), we observe that *all data retrieved and elaborated by TW is publicly available*: anyone can observe, collect, and use such data. We thus ask ourselves: “what if players’ in-game data are used *against* them to violate their privacy?” If this were true, then the E-sport setting would be prone to Attribute Inference Attacks (AIA). Such attacks, enabled by the capabilities of Machine Learning (ML), aim to infer private information about a given target (i.e., a player) by using their publicly available data [18].

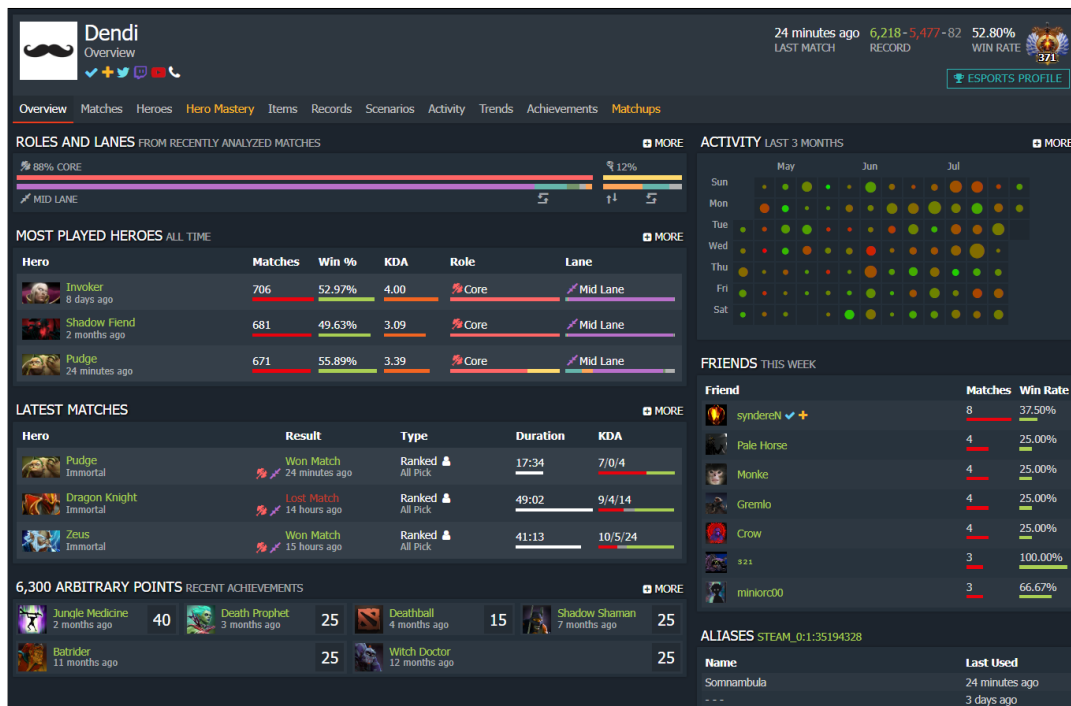


Figure 10.1: A TW, showing the statistics of the professional Dota2 player “Dendi” [1]. All such information is constantly updated and publicly accessible: <https://dotabuff.com/players/70388657>.

Although TW report only in-game statistics, we cannot exclude the existence of a link between such data and personal attributes (e.g., gender, age, personality) or even sensitive ones (e.g., health [447, 448])—the latter being outside our scope. Prior research (e.g., [449, 450]) revealed that a correlation exists between the in- and off-game traits of a given player. Surprisingly, however, no study has been carried out within the specific context of DOTA2. Such a lack is concerning: the in-game data provided by DOTA2 is semantically different from that of other VG. Hence, to this day, it is still *uncertain whether AIA are a threat to DOTA2 players*. Consequently, it is also unknown (i) *how* AIA can be carried out and (ii) what is the *impact* of an AIA in the DOTA2 context. Inspired by Biggio and Roli [451], we proactively assess the likelihood and the effects of this subtle privacy issue.

CONTRIBUTION. This chapter investigates the threat of AIA against DOTA2 players. We begin (Section 10.1) by contextualizing the E-Sports ecosystem (with a focus on DOTA2) and summarizing the fundamental concepts

of AIA (building on related work). Then, we provide four major contributions—which go *beyond the research domain*.

- **A threat model of AIA against DOTA2 players** (Section 10.2). We describe *how* to (legitimately) launch an AIA to infer private information on players while knowing only their DOTA2 handle. We also explain *why* attackers would do so.
- **We prove the existence of correlations between DOTA2 players’ in-game data and their personal attributes** (Section 10.3). By conducting an (informed) survey, we collect in-game and personal data of 484 DOTA2 players, corresponding to over 26k matches. We then perform a correlation analysis, showing the existence of statistically significant ($p < 0.01$) and strong (Spearman’s $\rho > 0.3$) relationships between in-game (public) and off-game (private) attributes.
- **We proactively evaluate the impact of AIA in DOTA2** (Section 10.4). We use the data gathered from our survey to (ethically) enact an AIA, and measure its success rate. We develop multiple ML models, by assuming attackers with varying domain expertise on DOTA2. We show that even simple AIA can be successful (almost 70% F1-score on gender), and that more sophisticated AIA can further increase such impact (over 75% accuracy on predicting the occupation).
- **We assess AIA that can be staged in practice** (Section 10.5). We assume the viewpoint of an attacker with *specific* goals, and elucidate the real-threat of AIA in DOTA2 by demonstrating a realistic application of our findings, showing AIA with near-perfect success rate (almost 100% precision).

Finally, we discuss our results, describe some countermeasures, and explain how our AIA can be extended to other E-Sports VG (Section 10.6). We then conclude the chapter and provide ethical considerations (Section 10.7).

TRANSPARENCY. We release a repository containing exhaustive details on our study, as well as the source code we developed for our analyses—available at: <https://github.com/hihey54/Dota2AIA>. Finally, note that the author of the thesis is a top-1% DOTA2 player, who brought his knowledge to the analyses.

DISCLAIMER. This chapter tackles a delicate privacy issue that potentially touches millions of video-gamers. All our evaluations are conducted ethically [452], but attackers are not bound to such ethics. At the time of writing, the problem is still open.

10.1 BACKGROUND AND RELATED WORK

This chapter tackles two emerging domains: competitive video games, and attribute inference attacks—which we now summarize.

10.1.1 THE COMPETITIVE VIDEO-GAME ECOSYSTEM

Competitive *video-games* (VG), and E-Sports in particular, are receiving a lot of attention [437], leading to a constant increase of players all aiming to “reach the top” [453]. To improve their performance, players can analyze

their in-game statistics [454]. Such statistics are typically provided by the VG itself, but are limited to a single *match*. Even if most VG allow players to inspect their history, analyses can only be performed on a match-by-match basis. Such limitation was overcome by *Tracking Websites* (TW), which collect and aggregate information pertaining to all matches of a given player(s), providing a comprehensive overview of their activity (cf. Fig. 10.1). An illustration of such ecosystem is in Fig. 10.2, which we now describe from the viewpoint of our VG of choice—DOTA2.

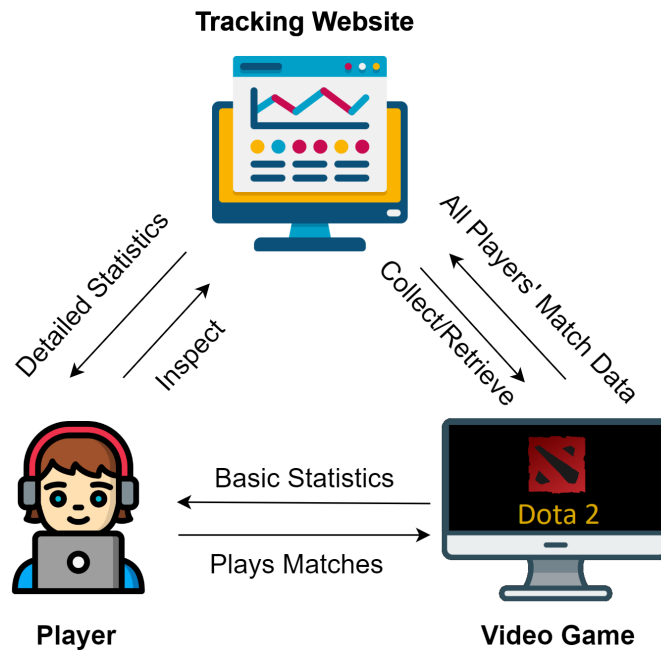


Figure 10.2: The E-Sport ecosystem. *Players* engage in *matches* of a *video-game*, which publicly releases data on such matches. These data are collected by *tracking websites*, whose elaborations are made public.

- **Video-Game.** DOTA2 is a Multiplayer Online Battle Arena (MOBA) VG. Released in 2013 and available for free, it is one of the most popular VG, counting up to 6M daily players and over 15M monthly players [4]. In a match, two teams of five players fight in real time with a common objective: destroy the enemy team’s base before they do it to yours.
- **Players.** Each player in a team has a crucial role in ensuring their team’s victory, and such roles are difficult to master. Indeed, DOTA2 is extremely competitive: in 2021, its biggest tournament had the largest prize pool in the entire history of VG, amounting to \$40M [443]. Such prizes are enticing for players, who continuously strive to get better: every DOTA2 player has more than 1600 hours [7] of playtime (on average). It is not surprising, hence, that DOTA2 players will resort to any (legitimate) tool to maximize their efficiency.
- **Tracking Websites.** A massive amount of DOTA2 players leverage the services provided by TW [6]. Reportedly, some TW tracked the activities of more than 79M players, aggregating the results of ~3B matches [446]. In our context, TW constantly interact with specific DOTA2 APIs to retrieve all historical

data pertaining to a player’s matches. Before using a TW, a player must explicitly allow DOTA2 to share their match details with external sources; however, considering the benefits provided by TW, only few players do not give their consent. Every player (and corresponding DOTA2 activity) tracked by a TW is publicly visible on the platform.

Such context begs the question: “**why are TW publicly releasing players’ data?**” The answer is: “because players themselves want such data to be public.” Indeed, such availability allows players to:

- browse other players’ statistics, so as to learn how the game is played by top-players;
- increase their visibility to professional organizations, which can hire them if they show good performance;
- share their activity with friends, teammates, or even unknown players that paired up with them;
- climb TW-specific rankings (e.g., players who get most wins with a given character).

Simply put, players benefit from their in-game data being publicly released by TW—thereby exposing players to the threat of AIA.

10.1.2 ATTRIBUTE INFERENCE ATTACKS

We summarize the fundamentals of Attribute Inference Attacks (AIA), and then highlight the research gap motivating this chapter.

AIA IN A NUTSHELL

The underlying goal of AIA is inferring *private* information on a given target by exploiting *publicly available* data on such target. For example, an attacker can use the (public) ratings posted on a video streaming platform by a given user to infer their (private) gender [455]. Such inference can be done leveraging the predictive capabilities of Machine Learning (ML): By training a ML model on a representative dataset, and then providing such ML model with some user’s public data, the ML model will output the personal attributes of such user. We remark that AIA are semantically different than membership inference attacks (e.g., [456, 457]), whose goal is inferring information on the ML model’s training set.

AIA are becoming problematic due to the lack of education of most internet users, who publicly share their data while overlooking (or ignoring) the corresponding risks (e.g., [458, 459, 460]). For instance, most data published on social networks can be easily retrieved via OSINT [461] and then used to setup an AIA. Indeed, most prior research considers the ecosystem of social networks, due to the ease of retrieving information linking public data with private attributes: Goelbeck et al. [462] infer personality traits of social media users. Jurgens et al [463] consider Twitter, and predict the location of the users based on their tweets. More recently, Gong et al. [464] focus on Google+ users, whereas Zhang et al. [465] consider, e.g., YouTube, and predict users’ gender (above 70% F1-score) based on their historical activity. Similarly, [466] focus on Facebook, showing that the gender can be predicted (~80% accuracy) by analyzing the usage of emojis. (The authors of [467] also consider Facebook, and infer sensitive data which is outside our scope). Other examples are [468, 455, 469, 470]. All such works show that AIA can be enacted in the real world, representing a subtle privacy risk.

MOTIVATION: AIA AND VIDEO GAMES

Surprisingly, no efforts consider AIA exploiting (public) VG data to infer players’ (private) attributes—to the best of our knowledge. As shown in Section 10.1.1, the competitive VG ecosystem (and especially the one of DOTA2) is particularly prone to the risk of AIA. A trace of such exposure is provided by the few works analyzing the correlation between the players’ in-game behaviour and their personal life—albeit for VG of different genres. For instance, Oggins et al. [471] highlighted that MMORPG players have a similar in- and off-game behaviour. Martinovic et al. [450] reasoned on how such similarity can be used by producers of MMORPG. For instance, some players’ physical traits can be inferred from their in-game avatar—which tends to be alike [472]. In this context, Likarish et al. [473] analyzed the in-game avatars to predict the age of the corresponding player; whereas Symbolski et al. [474] predicted the gender. Besides physical characteristics, some researches also studied personality indicators. Spronck et al. [475] found correlations between personality traits of 36 players and their playing-style. The only paper we are aware of that considers a competitive VG is [449], showing correlations between Battlefield 3 players’ in-game data and some of their personality traits.

Most related studies on VG (i) did not consider MOBA—which are our focus; and (ii) adopted the perspective of the producers of the VG—i.e., they assumed the availability of in-game data that was not publicly available [476, 477]. The latter is crucial: a *real* attacker is unlikely [478] to have access to a company’s databases—especially in domains with a high market share, such as (competitive) VG. Granted: such studies showed that correlations exist between players’ in- and off-game characteristics, *but in different VG*. No paper, however, investigated: (i) whether a correlation exists also in DOTA2; and, if it exists, (ii) ‘how’ and ‘to what extent’ it can be exploited in the DOTA2 ecosystem by real attackers—who are not omnipotent. The only work that considers a similar setting as ours is [6], but it focused on recognizing the play-style of DOTA2 players across different accounts—which is an objective orthogonal to ours. To the best of our knowledge, we are the first to investigate AIA in VG.

10.2 DOTA2 ATTRIBUTE INFERENCE ATTACKS

Our primary contribution is the first threat model for feasible AIA against DOTA2 players. We describe ‘how’ AIA can be staged in the DOTA2 ecosystem (Section 10.2.1); and ‘why’ attackers would do so (Section 10.2.2).

10.2.1 PROPOSED THREAT MODEL

Our AIA is mostly tailored for players who actively engage in *competitive* DOTA2 matches. (Some DOTA2 players do not “play to win”, and hence are less likely to use TW.) For simplicity, we assume that a player only owns a single ‘handle’ (e.g., “Dendi” in Fig. 10.1 is the handle of the player “Danil Ishutin”), which is used to retrieve data from any public source (e.g., tracking websites).

FORMAL DEFINITION. We describe the viewpoint of our considered attacker according to the following four criteria [451]:

- *Goal:* The attacker wants to infer the personal attributes of a set of players whose real identity is completely private.

- *Knowledge*: The attacker knows the handles of a set of players, and is well-aware of the DOTA2 ecosystem.
- *Capability*: The attacker can only access and retrieve data that is either publicly available, or that users are willing to share (e.g., social networks, public surveys).
- *Strategy*: The attacker first (legitimately) gathers information associating players’ in-game data with their respective personal attributes. Then, the attacker trains a ML model to perform AIA against players whose personal information is completely private, i.e., by only using their (known) handle.

We implicitly assume that the targeted players enabled in-game data sharing with external sources (e.g., TW). We stress that the attacker shall *not* perform any data breach to obtain the desired private information—an attacker will never launch an AIA otherwise.

PRACTICAL SCENARIO. We present in Fig. 10.3 an illustration of our threat model, which is divided in three stages: *prepare, infer, exploit*.

1. *Prepare*. First (left of Fig. 10.3), the attacker must collect a representative dataset associating DOTA2 players’ in-game data (e.g., daily matches played, win/loss ratio) with the corresponding ground truth (e.g., the players’ gender).
2. *Infer*. Then (middle of Fig. 10.3), the attacker uses the harvested dataset to train a ML model, which is the tool to carry out the AIA. The inference is done by providing public in-game information on a target player (obtainable, e.g., by querying a TW with the handle of a player) as input to the ML model.
3. *Exploit*. Finally (right of Fig. 10.3), the attacker benefits by either stalking a victim (targeted AIA), or by profiting from the inferred attributes (an indiscriminate AIA).

10.2.2 FEASIBILITY OF AIA IN DOTA2

Any attack is theoretically possible, and several papers (e.g., [479]) advocate to always consider worst-case scenarios. Nonetheless, we argue that our proposed AIA are not only possible “in theory”, but also likely to occur “in practice” due to their high feasibility [478]. Indeed, real attackers have a cost-benefit mindset [480]. In our case, AIA will be launched only if an attacker finds them easy to setup (in terms of cost and risk), and if they lead to tangible benefits.

In particular, we focus the attention on three aspects—each pertaining to a given stage of our exemplary use-case, namely: acquiring the dataset to train the ML model (i.e., the *capabilities* of the attacker); improving the performance of such ML model (i.e., the *knowledge* of the attacker); and how a successful AIA can be exploited (i.e., the *goal* of the attacker).¹

- **Data Harvesting.** Obtaining *public* in-game data of multiple players (i.e., the “features”) is straightforward in DOTA2: it is simply necessary to go to a TW² and retrieve all information related to a set of players. In contrast, obtaining the corresponding personal attributes (i.e., the “labels”) may appear harder, as

¹We observe that our threat model is significantly different from the one in [481].

²We observe that abundant information is also available directly from DOTA2, hence TW are not strictly required (we will discuss this in Section 10.6).

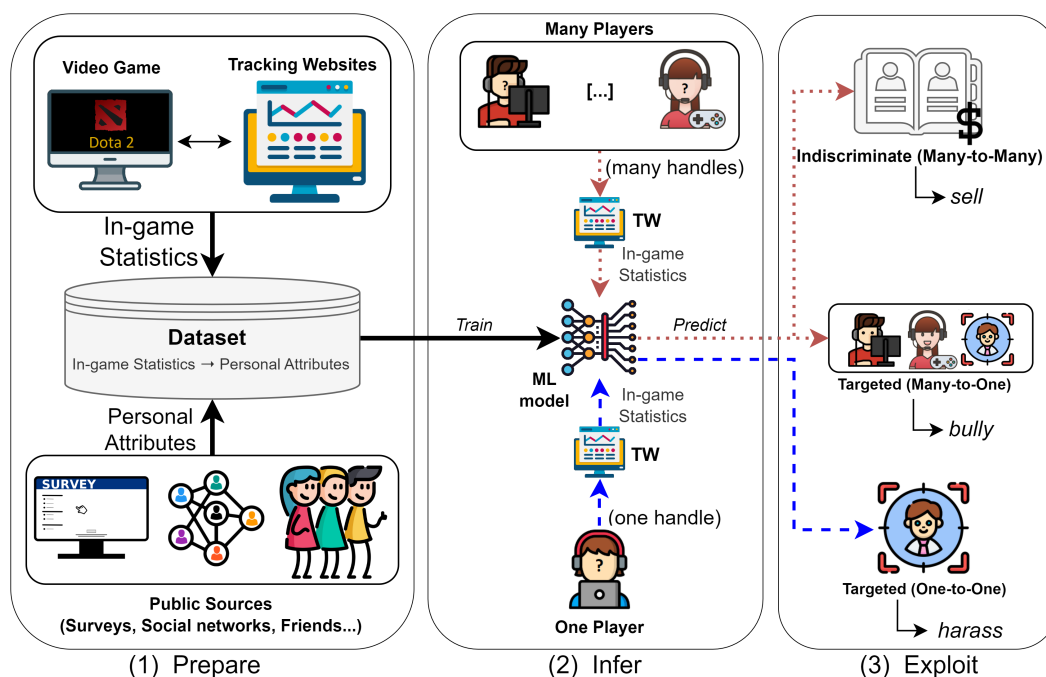


Figure 10.3: Overview of our proposed AIA against Dota2 players. Public information is used to infer personal (private) attributes.

such information is typically kept *private*. Unfortunately, this is not the case in the DOTA2 ecosystem.³ For instance, the real identity of many players (e.g., professionals or streamers) is well-known. Moreover, it is possible to search for a given handle on popular search-engines and inspect the results. For example, a given player may use the same handle also on social media; some people even announce their handle on public forums to facilitate establishment of partnerships. Alternatively it is also possible to conduct surveys in which interviewees must input their handle, as well as some inconspicuous private information (e.g., gender, age). For instance, two large surveys were carried out in 2016 and 2021, receiving 30k and 8k responses respectively, by simply posting announcements on popular boards [3].

- **Refining the ML model performance.** Even if an attacker can acquire a suitable training dataset, it is unlikely that such dataset can yield a proficient ML model from the start—hence, *naive* attackers will hardly be successful in their AIA. *Expert* attackers, however, can use their superior knowledge on the DOTA2 scene to improve the success rate of their AIA. In our evaluation (Section 10.4) we will show some pre- and post-processing techniques that boost the predictive performance of the ML model. Given that attackers interested in our AIA are well-aware of how DOTA2 works, this characteristic further aggravates the threat of AIA.
- **Exploiting AIA.** We identify three ways in which an attacker can benefit from AIA in DOTA2. (We will consider all of these ways in our evaluation.) First, they can launch an *indiscriminate* ‘many-to-many’

³Zhang et al. [465] also state that ground-truth harvesting is easy in today’s landscape.

AIA, i.e., by using many handles (belonging to many players) to infer the respective personal attributes; such attributes can then be sold⁴ to any potential buyer—e.g., dark web, or even to ad-companies which want to send customized ads [483]). Second, they can launch a *targeted* ‘one-to-one’ AIA by inferring the attributes of just one player—e.g., after losing a match, an attacker can launch an AIA against a player of the opposing team and harass them [6]). Third, they can launch a *targeted* ‘many-to-one’ AIA by inferring the attributes of a set of (many) players, and then finding a (single) player within such set that meets some criteria—e.g., finding an underage player and then bully them [484, 485, 486].

Finally, we observe that the results of the two surveys [3] showed similar trends despite the 5 year timespan. Such stability may suggest that even **data collected many years prior can still be used to enact successful AIA**. Considering the high likelihood of such a threat, we embrace Biggio and Roli’s [451] recommendation: we must proactively assess the impact of AIA in DOTA2.

TAKEAWAY: Attackers can – cheaply and legitimately – use many methods to setup an AIA, which can be exploited in various ways to violate DOTA2 players’ privacy.

10.3 PRELIMINARY ASSESSMENT

A prerequisite for a successful AIA is the existence of relationships between the players’ in-game data, and their corresponding personal attributes [487]. We recall (Section 10.1.2) that past research found some correlations—but *in different VG* (e.g., Battlefield3 [449]).

Hence, as our second contribution, we now investigate whether there is some evidence hinting that AIA “can be successful in DOTA2”. To this purpose, we perform an extensive survey on real DOTA2 players (Section 10.3.1 and Section 10.3.2), and analyze the correlation coefficient between their in-game statistics and personal attributes (Section 10.3.3).

10.3.1 COLLECTION OF PERSONAL ATTRIBUTES (SURVEY)

We conduct a survey to collect the handles of DOTA2 players, together with their personal attributes.

METHOD. The handle consists in the Steam ID of each player. For the personal attributes, we consider: gender, age, occupation, purchase_habits, as well as the “Big Five” personality traits [488]). Such attributes are those typically envisioned by past research (e.g., [462, 465, 466, 469]); the only exception is purchase_habits, which is an ‘original’ attribute that we propose due to the given DOTA2 context, in which players typically purchase “cosmetics” to embellish their characters. Nevertheless, all such personal attributes represent information that is *not available* from any resource linked with DOTA2: hence inferring such information without the explicit consent of the corresponding player represents a privacy violation.⁵ Our survey entailed 10 questions used to determine the

⁴This is a popular strategy adopted by some real companies [482].

⁵Even purchase_habits is not public: a player may have many “cosmetics”, which can have been *gifted*; moreover, a single purchase may include *more* than a single “cosmetic”, which can also be obtained via “bundles”.

personality traits [489]; 4 questions which explicitly referred to the remaining four attributes considered in this chapter; as well as one question for the country. We also included 10 questions, which served both as ‘attention checks’, but also for verifying the authenticity of the answers (e.g., we asked “what is your favorite DOTA2 hero?” and we verified on a TW whether the answer was genuine).⁶ Overall, the survey began in Oct. 2019 and ended in Dec. 2019. In this timeframe, we hosted our survey on a website, whose link was distributed on many online social media platforms such as Facebook, Reddit, Discord, and Telegram. Upon landing on the survey’s website, participants had to login with their Steam account (via OpenID), thereby ensuring that all personal attributes were correctly linked to the actual player.

Table 10.1: Personal attributes considered in our study. Our population is of 484 Dota2 players. The distribution resembles the one in [3].

<i>Private Attribute</i>	<i>Description</i>	<i>Classes Distribution</i>
gender	Gender at birth	<i>Female:</i> (4.96%), <i>Male:</i> (95.04%)
age	Current age	<i>13–18:</i> (13.43%), <i>19–24:</i> (53.72%), <i>25–38:</i> (32.85%)
occupation	Whether a player is employed or not	<i>No:</i> (57.44%), <i>Yes:</i> (42.56%)
purchase_habits	Frequency of in-game purchases	<i>Never:</i> (10.54%), <i>Rarely:</i> (61.16%), <i>Regularly:</i> (28.30%)
openness	Inventive/curious (high) vs. consistent/cautious (low)	<i>Low:</i> (19.22%), <i>Medium:</i> (24.38%), <i>High:</i> (56.40%)
conscientiousness	Efficient/organized (high) vs. extravagant/careless (low)	<i>Low:</i> (39.46%), <i>Medium:</i> (23.97%), <i>High:</i> (36.57%)
extraversion	Outgoing/energetic (high) vs. solitary/reserved (low)	<i>Low:</i> (47.31%), <i>Medium:</i> (21.07%), <i>High:</i> (31.62%)
agreeableness	Friendly/compassionate (high) vs. critical/rational (low)	<i>Low:</i> (20.87%), <i>Medium:</i> (19.42%), <i>High:</i> (59.71%)
neuroticism	Sensitive/nervous (high) vs. resilient/confident (low)	<i>Low:</i> (53.51%), <i>Medium:</i> (19.21%), <i>High:</i> (27.27%)

ANALYSIS. We received 625 answers from 62 different countries. We filtered out: 18 invalid answers; 43 participants who were not visible on any TW; and 78 inactive players (i.e., less than 5 games in the last month). Thus, our sample size consists in 484 players. Despite being far smaller than the overall amount of DOTA2 players, such number still allows to draw statistically significant result. Indeed, we are above the minimum sample size of 384 required by setting a confidence level of 95%, a margin of error of 5%, population proportion of 50%, and a population size of 7 million [490]. We report in Table 10.1 the considered personal attributes, as well as their class-distribution in our population. We grouped age in three bins (similarly to [468]): very young/underage, young adults, and over 25 (our ‘oldest’ respondent was 38); the frequencies for purchase_habits are never, less than once a month (rarely), and monthly or more often (regularly); for occupation, we consider a student as unemployed. Since our survey quantified each personality trait as an integer [0–100], we group such values into three categories (similarly to [491]) differentiating low, middle, or high scores.

VALIDATION. By observing Table 10.1, we can see that some classes may present a high imbalance, such as gender or age. However, our class-distribution is strikingly similar to those of the surveys carried out in previous years [3]: specifically, we focus on the largest survey from 2016, whose sample size was of 29,351. Let us make some exemplary comparisons, so as to *validate* all our subsequent analyses: if our population significantly differs from the ‘real’ one, then we cannot claim that the threat is ‘real’. According to [3], *male* players are 96%, which match our results of 95%. The same can be said for age: according to [3], minors represent 15% of the population (ours

⁶Our repository includes the full questionnaire. Some questions found therein asked for other (non-sensitive) information that do not pertain to this chapter.

is 13.4%), whereas young adults are 66% (ours is 54%), with over 25 being 20% (ours is 33%). (Small differences are due to slightly different thresholds for the bins). For occupation, the unemployed are 67% in [3] (ours is 57%).

Summary: from our survey, we derive that: our population (i) is representative of the DOTA2 community, and (ii) is large enough to derive statistically significant conclusions. Moreover, our survey also shows that (iii) the DOTA2 community is willing to participate in online surveys—representing one of the means an attacker can use to harvest players’ private information for a (real) AIA.

We will use \mathcal{A} to indicate the dataset containing the (personal) *attributes* of our 484 players—collected via our ethical survey.⁷

10.3.2 COLLECTION OF IN-GAME STATISTICS (TW)

Once we obtained the handles of the participants, we retrieved their in-game statistics via public Tracking Websites.

METHOD. Our TW of choice is OpenDota because it provides free APIs⁸ usable to retrieve in-game statistics. We used two APIs:

- `player`, which, given a handle, returns some summary statistics (e.g., win/loss ratio) of the corresponding player, as well as the list of matches⁹ played by such a player;
- `matches`, which, given the identifier of a match (obtained from the `player` API), returns all information on that specific match (e.g., kills, deaths, assists).

We report in Table 10.2 the information returned by our invoked API. Some fields are provided as lists, which include additional entries. For example, `matches_chat` includes all messages exchanged by the two opposing teams during a DOTA2 match. For a detailed explanation of all fields, we refer the reader to the official documentation.

Overall, after querying the `players` API for each of the 484 players, we found out that our population participated in 26241 matches during the considered timeframe. Therefore, we invoked the `matches` API on all these entries.

PREPROCESSING. By applying original feature engineering techniques on the data retrieved from OpenDota, we distill additional knowledge to assist in our analysis. Such techniques involve both ‘traditional statistics’, but also our own ‘domain expertise’ on DOTA2.

- *Traditional Statistics.* The most straightforward operation involves computing some aggregated metrics on the details of each match played by a given player (e.g., average match length). We also perform some more refined operations. For instance, the `players` API does not directly provide the playtime trend of a given player, but such information can be computed by using the results from `matches`: by inspecting

⁷We never attempt at inferring additional (private) information of our respondents.

⁸OpenDota API: <https://docs.opendota.com/>

⁹For simplicity, we only considered the matches played in the previous 30 days since making each API call (i.e., from December 2019 to January 2020).

Table 10.2: Data returned by the player and matches OpenDota APIs.

<i>Type</i>	<i>Field</i>	<i>Type</i>	<i>Field</i>	<i>Type</i>	<i>Field</i>
num	player_rank_tier	num	match_human_players	list	match_radiant_team
bool	player_plus	num	match_lobby_type,	list	match_dire_team
list	player_matches	list	match_objectives	num	match_skill
num	match_match_id	list	match_picks_bans	list	match_players
num	match_barracks_status_dire	num	match_positive_votes	num	match_patch
num	match_barracks_status_radiant	list	match_radiant_gold_adv	num	match_region
list	match_chat	num	match_radiant_score	list	match_all_word_counts
list	match_cosmetics	bool	match_radiant_win	list	match_my_word_counts
num	match_dire_score	list	match_radiant_xp_adv	num	match_throw
list	match_draft_timings	num	match_start_time	num	match_comeback
num	match_duration	list	match_teamfights	num	match_loss
num	match_first_blood_time	num	match_tower_status_dire	num	match_win
num	match_game_mode	num	match_tower_status_radiant		

the dates of the matches played, we can identify, e.g., which day of the week a given player is most likely to play DOTA2.

- *Domain Expertise.* By applying knowledge on the DOTA2 context, we further increase the amount of information usable for our analysis. As an example, we inspect all chat messages to determine whether players use words that are typical of DOTA2 slang (e.g., “cd”, “b”, “rat”, “smurf”, “gank”). We provide in our repository (see Appendix A) an additional description of how we computed the features related to `match_chat`.

Overall, we compute over 300 features—all of which are novel in the context of AIA.¹⁰ Such features identify three datasets: \mathcal{P} , focused on the players, containing 484 samples, each described by 187 features; \mathcal{M} , focused on the matches, containing 26241 samples, each described by 137 features; and $\overline{\mathcal{M}}$, containing 11117 samples and 160 features, which is a ‘distilled’ version of \mathcal{M} . In particular, $\overline{\mathcal{M}}$ differs from \mathcal{M} in two ways: First, we address the problem of the highly imbalanced distribution of \mathcal{M} in terms matches-per-player (some players in \mathcal{A} have only 5 matches in \mathcal{M} , while others have hundreds); we thus reduce the potential bias by randomly sampling¹¹ at most 30 matches for each player. Second, we augment the features in \mathcal{M} with those derived with our domain knowledge; the intention is determining how much of an impact our intuitions (resembling those of an attacker) have on all our experiments.

10.3.3 CORRELATION BETWEEN DOTA2 IN-GAME STATISTICS AND PERSONAL ATTRIBUTES

We can now objectively determine whether a relationship exists between DOTA2 players’ in-game statistics and their personal attributes. This step is crucial to provide a theoretical foundation supporting the effectiveness of

¹⁰A complete description of all our considered features is provided in our repository.

¹¹To mitigate the effects of randomness, we create 20 versions of $\overline{\mathcal{M}}$ and will use all of these for our experiments, averaging the results.

AIA in this context.

METHOD. We perform a correlation analysis between our three dataset containing in-game statistics, and the dataset containing corresponding personal attributes. Inspired by [462], we compute the correlation between each feature of $(\mathcal{P}|\mathcal{M}|\overline{\mathcal{M}})$, with each feature of \mathcal{A} . To conduct a rigorous analysis, for each pair of features we compute: (i) the *statistical significance* of the correlation—measured with a p -value; and (ii) the corresponding *strength of the relationship*—whose measure varies depending on the chosen correlation metric. We consider two metrics [492]: *Cramer’s V* for categorical variables; *Spearman’s ρ* for numerical variables. We remark that low p denotes strong significance (we set $p < 0.01$ as default threshold), whereas strong relationships are denoted by high absolute values of the corresponding metric (ranging between 0 and 1).

RESULTS. We report in Fig 10.4 the correlation between \mathcal{P} and \mathcal{A} as measured by the ρ metric. For each numerical variable in \mathcal{A} , we report the top-3 variables¹² of \mathcal{P} (as measured by ρ), all of which obtain $p < 0.01$. We can see that age is correlated with kills, probably because younger players have an aggressive playstyle. A strong correlation exists between purchase_habits and (i) cosmetics_prices, i.e., the money spent by a player in skins; and (ii) special messages (i.e., hero_msg and counter_thank_msg) that can be unlocked with a paid subscription. Moreover, extroversion is highly correlated to chat usage (i.e., rank_chat and ratio_chat_msg); whereas agreeableness to wins in unranked games (i.e., normal_win). Interestingly, neuroticism is correlated with denies (a unique mechanic of DOTA2), openness to the type of selected heroes, and conscientiousness is low for players that play on Thursdays. Although not shown in Fig. 10.4 (because they are categorical features), we also mention high correlation between the gender of the player and the gender of the most played heroes (which is common in cooperative VG [474]); whereas the occupation is strongly correlated to paid subscriptions.

TAKEAWAY. A correlation exists between DOTA2 players’ in-game data and their personal attributes. Our finding demonstrates the risk of AIA in DOTA2.

Additional analyses (as well as the variants of Fig. 10.4 for \mathcal{M} and $\overline{\mathcal{M}}$) and heatmaps are provided in our repository (see Appendix B).

10.4 PROACTIVE EVALUATION OF AIA IN DOTA2

Our preliminary assessment provides evidence that AIA against DOTA2 can be successful. Hence, as our third contribution, we set out to proactively evaluate the impact of such AIA. To this purpose, we use the data derived from our survey (described in Section 10.3.1 and Section 10.3.2) to perform various ethical and controlled AIA.

Specifically, we find instructive to study three diverse AIA, each requiring different amounts of preparation. First, we consider the most *simple* way to carry out an AIA, i.e., by using only the aggregated data of each player (Section 10.4.1). Second, we evaluate the success rate of AIA that use information derived from just *one match* (Section 10.4.2). Third, we analyze *sophisticated* AIA in which the attacker leverages all their expertise to maximize their impact (Section 10.4.3). Finally, we perform a reflective exercise by discussing the general context of

¹²We remark that $\rho > 0.1$ is a valid signal indicator for orthogonal tasks [493].

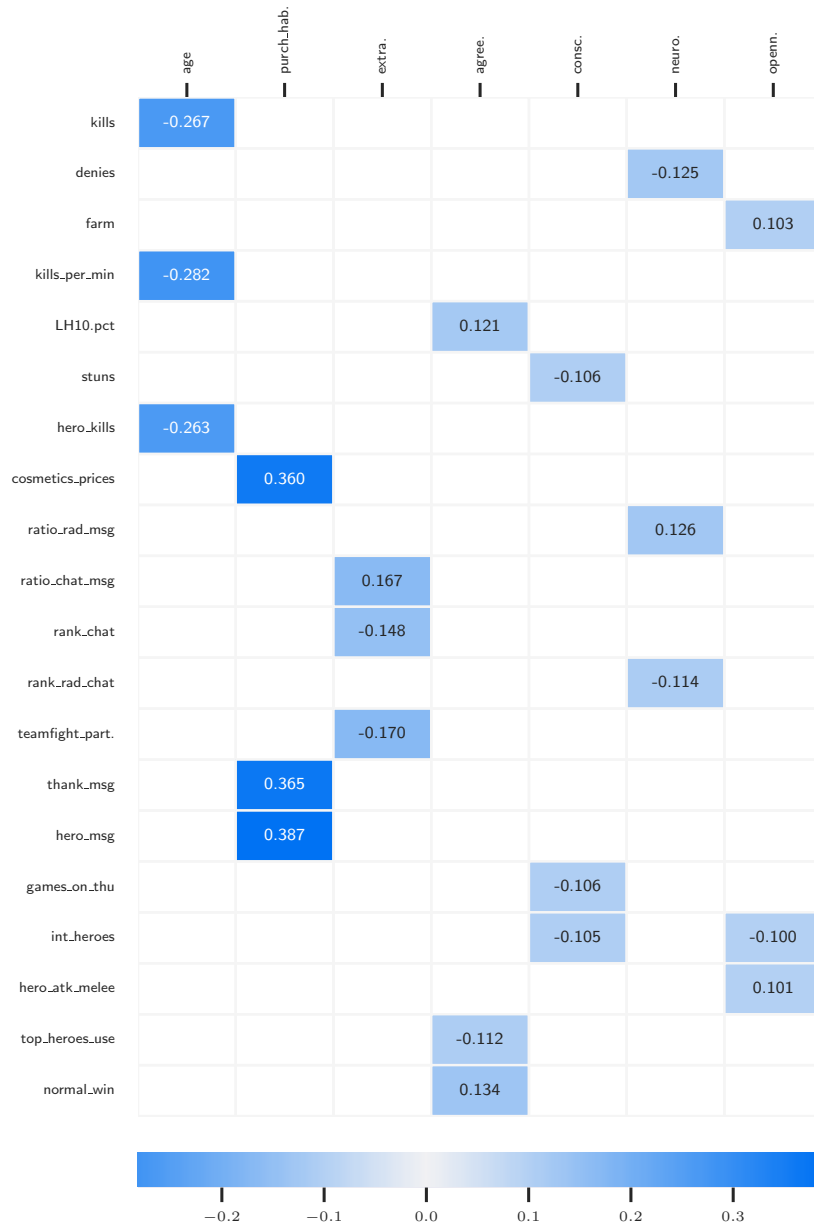


Figure 10.4: Top-3 Spearman's ρ between \mathcal{P} and \mathcal{A} (at $p < 0.01$). Higher absolute values denote stronger correlation, while the sign indicates the direction of the correlation.

AIA in light of the results achieved in research (Section 10.4.4). For a statistical validation of all our findings, see Appendix C in our repository.

COMMON SETUP. We always adhere to our threat model (Section 10.2.1). The attacker knows the handle of one or more players, and uses such handle to retrieve in-game data from TW, which are then provided as input to an ML model for inference. Moreover, we also assume that the attacker gathered the private attributes for training the ML model via a survey (i.e., the one described in Section 10.3.1). Indeed, as evidenced by [3], thousands of DOTA2 players willingly participate in game-related surveys. For ethical reasons, we do not violate our respondents' privacy by performing OSINT, or crawl their social media profiles (which are both viable means that an attacker can – legitimately – use to improve their AIA).

10.4.1 SIMPLE AIA (AGGREGATED PLAYER DATA)

The underlying principle of these AIA is that they only use the information contained in \mathcal{P} , i.e., which aggregates the statistics of all matches played by any given player. Such information is simple to compute, but is lossy. For instance, the `average_match_length` includes the duration of all matches, and inevitably leads to oversimplifications. However, due to their simplicity, such AIA are feasible to stage and it is important to assess their impact.

TESTBED. For these experiments, we merge \mathcal{P} with \mathcal{A} , generating a single dataset containing 484 samples, each described by 187 features (from \mathcal{P}) and associated to 9 attribute labels (from \mathcal{A}). To develop the ML model for the AIA, we consider four ML algorithms: Logistic Regression (*LR*), Decision Trees (*DT*), Random Forest (*RF*), and Neural Networks (*NN*). We validate our results through a nested stratified 10-fold cross-validation, during which we also apply feature selection and hyperparameter optimization for each considered ML model. Finally, to address the imbalance of some target attributes (e.g., age), we apply well-known under- and over-sampling techniques [494, 495] (as also recommended in [479]).

IMPACT. We report in Table 10.3 the results of the simple AIA. Rows denote the target attributes, whereas columns denote the considered ML algorithms; the rightmost column refers to a ‘Dummy’ stratified classifier (simulating a random guess) which we use as baseline for comparison. Cells report the predictive macro F1-score (and standard deviation) across all our trials.

From Table 10.3, we observe that at least one of our models always outperforms the baseline. The *NN* achieves remarkable performance (almost 70% F1-score) to predict gender, whereas occupation is correctly predicted with almost 60% F1-score. In contrast, some attributes are very difficult to predict, such as `purchase_habits` for which the performance hardly goes 3% above the baseline. We can conclude that such simple AIA can be effective in some cases, but real attackers can easily improve the success rate by considering additional information—as we will show in Section 10.4.3.

Table 10.3: Impact of the *simple* AIA (based on \mathcal{P}) as measured by the F1-score. Rows report the attributes and columns our ML models (boldface denotes the best model for a given attribute).

	<i>LR</i>	<i>DT</i>	<i>RF</i>	<i>NN</i>	<i>Dummy</i>
gender	64.97 \pm 10.9	59.71 \pm 12.7	50.91 \pm 5.33	67.24 \pm 13.4	51.62 \pm 10.9
age	40.47 \pm 6.30	39.38 \pm 8.76	44.08 \pm 6.17	28.06 \pm 7.59	32.21 \pm 5.70
occup.	53.23 \pm 7.22	47.44 \pm 8.34	56.08 \pm 7.88	59.89 \pm 7.15	43.76 \pm 9.56
purch.	32.05 \pm 10.1	31.74 \pm 4.53	34.40 \pm 8.20	32.17 \pm 7.19	31.20 \pm 6.26
open.	28.94 \pm 5.94	40.76 \pm 6.80	32.6 \pm 7.77	30.89 \pm 7.60	29.59 \pm 2.04
consc.	26.52 \pm 5.65	33.87 \pm 8.78	34.27 \pm 5.60	23.83 \pm 8.18	33.23 \pm 8.94
extrav.	30.15 \pm 7.53	36.16 \pm 5.14	36.49 \pm 5.56	28.59 \pm 5.95	32.27 \pm 7.01
agreeab.	29.46 \pm 6.29	34.11 \pm 8.58	33.68 \pm 6.25	24.54 \pm 9.43	33.39 \pm 7.35
neurot.	32.38 \pm 6.56	40.76 \pm 6.80	32.6 \pm 7.74	31.6 \pm 8.30	30.07 \pm 4.46

10.4.2 ONE-MATCH AIA (ABLATION STUDY)

We now assess the effects of AIA carried out by using the statistics of just *a single match*. This scenario can be considered as either a best-case or a worst-case depending on the viewpoint. Indeed, we can expect that using only one match to predict the personal attributes may yield poor results—which is a best-case for the defender. However, if such an AIA is successful, it would turn into a worst-case because the attacker can infer the private attributes with limited information (e.g., less queries to the TW API).

Moreover, we consider two attackers: an ‘expert’ attacker that uses their *domain expertise* to distill additional knowledge from the single match; and a ‘naive’ attacker that does not do so. Hence, the results of the ‘naive’ attacker can serve as an *ablation study*, allowing to gauge the effects of domain expertise in AIA.

TESTBED. To simulate the ‘naive’ attacker, we merge \mathcal{M} with \mathcal{A} . Hence, for each of the 26241 matches in \mathcal{M} (described by 137 features), we append the 9 attributes of \mathcal{A} . For the ‘expert’ attacker, we merge $\overline{\mathcal{M}}$ (having 11117 matches, each with 160 features) with \mathcal{A} , because $\overline{\mathcal{M}}$ is augmented with DOTA2 domain knowledge. We consider the same ML algorithms as in the simple AIA (i.e., *RF*, *LR*, *NN*, *DT*). We then train and test ML models by adopting a split of 80:20 (such split is done on the basis of the unique players in \mathcal{M} (or $\overline{\mathcal{M}}$) to avoid overfitting); we reserve 10% of the training set for validation purposes. Finally, we repeat all our experiments 20 times to account for the random sampling of $\overline{\mathcal{M}}$.

IMPACT. We report the results in Table 10.4; for simplicity, we only consider the models using *RF*, because they consistently outperformed all the others. The three columns show the F1-score obtained by the ‘naive’ (left) and ‘expert’ (middle) attackers, as well as that of a ‘Dummy’ classifier (right) that simulates a coin-toss.

From Table 10.4, we can see that the ‘naive’ attacker cannot successfully predict 8 out of 9 attributes, because the F1-score is always comparable (or even inferior) than the Dummy classifier. The only exception is the age attribute, for which the F1-score is 10% superior (albeit still hardly usable). We also note that such results are inferior to those of the simple AIA (cf. Table 10.3). From a defender’s viewpoint, these results may appear encouraging.

Table 10.4: Impact of the *one-match* AIA (F1-score). Columns refer to the ‘naive’ attacker (using \mathcal{M}), ‘expert’ attacker (using $\overline{\mathcal{M}}$), and the Dummy (random guess). The expert attacker is always superior.

	Naive attacker (ablation study)	Expert attacker (domain knowledge)	Dummy (baseline)
gender	49.03 \pm 0.18	58.47 \pm 5.21	49.75 \pm 0.55
age	43.72 \pm 2.66	56.82 \pm 3.01	33.28 \pm 0.46
occup.	49.42 \pm 4.56	68.42 \pm 1.90	49.87 \pm 0.89
purch.	35.61 \pm 5.06	49.71 \pm 3.85	33.37 \pm 0.53
open.	32.26 \pm 3.75	43.73 \pm 2.96	33.48 \pm 0.41
consc.	29.49 \pm 3.63	46.11 \pm 3.20	32.88 \pm 0.62
extrav.	32.33 \pm 2.47	46.82 \pm 1.96	33.25 \pm 0.56
agreeab.	33.62 \pm 2.28	45.36 \pm 3.37	34.09 \pm 0.46
neurot.	27.39 \pm 4.78	46.60 \pm 2.72	33.65 \pm 0.58

Unfortunately, the ‘expert’ attacker is much more successful, with 10–20% improvements over the Dummy classifier. Notably, occupation reaches \sim 70% F1-score (up from 49%), whereas gender almost 60% (up from 49%). Such results prove that using domain knowledge of DOTA2 substantially improves the success of AIA. What is surprising is that such AIA require the statistics of *a single match* (i.e., just one API query).

10.4.3 SOPHISTICATED AIA

We now assess AIA launched by a sophisticated attacker who, alongside using their domain expertise during pre-processing, exploits post-processing methods to further improve the AIA success rate.

INTUITION. We build from the one-match results of the the ‘expert’ attacker (Section 10.4.2). Then, we leverage the fact that a given DOTA2 player (i.e., the one targeted by the attacker) typically plays many matches. It is reasonable to assume that said player exhibits a *stable behaviour* across all such matches. Indeed, taken individually, a single match may not capture the true behaviour of a given player, thereby leading an ML model to make a wrong prediction; however, by considering the predictions of the *same* ML model to *many* matches (from the same player), the stable behaviour (i.e., the desired attribute) of the targeted player is more likely to emerge. For example, a player that has ‘high’ openness may not show such trait in every single match; but such trait may emerge by (independently) analyzing more matches, and aggregating the results.

TESTBED. We use the ML models trained with $\overline{\mathcal{M}}$ using the *RF* algorithm. Then, we provide as input to such models an increasing amount of matches from the same targeted player: specifically, we consider from 1 up to 30 matches (if available), which are randomly sampled (from the test portion of $\overline{\mathcal{M}}$). Then, for each attribute in \mathcal{A} , we take the predictions (provided as probabilities) of the ML model for all such matches, and we average all such predictions, choosing the one with the higher value.¹³ To reduce bias, we repeat all such experiments 20 times for

¹³E.g.: we want to predict the occupation (which is binary) of a player by analyzing 4 matches. The ML model analyzes 4 matches and outputs 4 probabilities, e.g., {0.1, 0.2, 0.8, 0.2} (i.e., values below/above 0.5 denote em-

each different variant of $\overline{\mathcal{M}}$; and, we repeat the draw of the chosen matches 1000 times.

IMPACT. The results of our sophisticated AIA are shown in Fig. 10.5, showing accuracy (y-axis) as a function of the matches analyzed by the ML model (x-axis). Lines correspond to the target attributes; shaded areas show the standard deviation. We do not report gender because the highly unbalanced population would inflate the results.

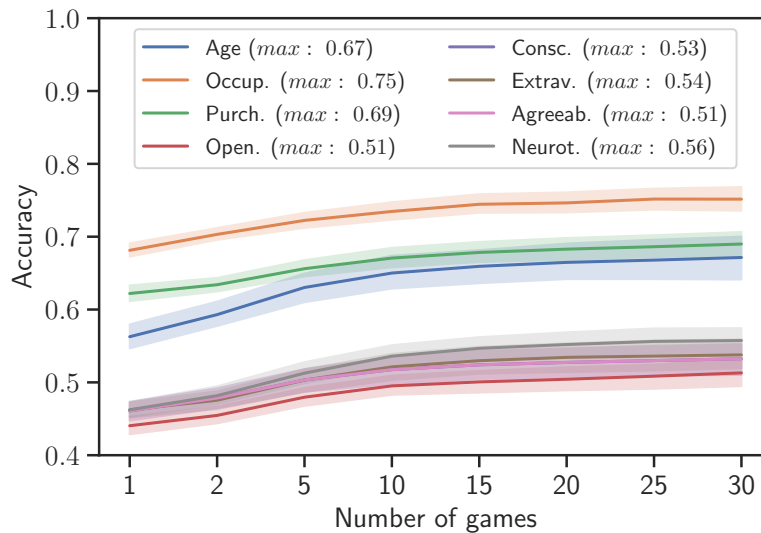


Figure 10.5: Impact of Sophisticated AIA. We post-process the predictions of the ML model over multiple matches of the same targeted player.

From Fig. 10.5 we can see that the accuracy increases as more matches are analyzed. For example, predicting the occupation goes from 68% up to 75% after 15 matches. Similarly, age goes from 58% up to 65%. What makes these results concerning is that retrieving the information on such extra matches requires little effort by the attacker, because (i) it is free and (ii) it can be automatized.

10.4.4 REFLECTION: AIA IN RESEARCH AND IN PRACTICE

As a reflective exercise, we report in Table 10.5 the results (according to a given ‘Metric’) obtained by some prior works attempting to predict the same attributes considered in this chapter (we exclude purchase_habits because it is novel). We stress that Table 10.5 is not meant to be a way to compare our AIA with previous ones, since we are the first to consider the DOTA2 setting (Section 10.1.2). Moreover, past works envision (i) different classes having (ii) different distributions for each attribute—making any comparison unfair.

From Table 10.5, we can see that – from a general viewpoint – obtaining high performance (e.g., overall accuracy) is difficult for some attributes. However, the real threat of AIA lies in the fact that they can be customized: although precisely inferring, e.g., the age of *all* individuals among a population may be unfeasible, it is different

ployment/unemployment). We assign the class after averaging the probabilities, thereby ‘filtering’ the noise (i.e., the 0.8).

Table 10.5: Results of prior work on AIA. Cells denote the value of a given ‘Metric’ for each of the attributes considered in this work.

Prior Work	Metric	gend.	age	occup.	open.	consc.	extrav.	agreeab.	neurot.
Goelbeck [462]	MAE	—	—	—	0.09	0.10	0.14	0.11	0.13
Weinsberg [455]	AUC	0.84	—	—	—	—	—	—	—
Al [496]	Acc.	0.80	0.80	—	—	—	—	—	—
Chen [468]	AUC	0.82	0.61	—	—	—	—	—	—
Fang [497]	Acc.	0.80	0.73	0.25	—	—	—	—	—
Bunian [491]	Acc.	—	—	—	0.58	0.60	0.58	0.58	0.58
Yo [469]	Acc.	0.70	0.80	0.70	—	—	—	—	—
Mei [498]	MAE	—	0.09	—	—	—	—	—	—
Pijani [466]	F1	0.83	—	—	—	—	—	—	—
Zhang [465]	F1	0.74	0.38	0.13	—	—	—	—	—
Eidizadehakhcheloo [470]	AUC	0.95	0.98	—	—	—	—	—	—

when the objective is more specific. For instance, an attacker may want to identify just a specific group of people (e.g., children—see Section 10.2.2), and they can tweak their ML models for this exact purpose.

A POSITIVE MESSAGE. This chapter tackles an open issue, and our ultimate goal is to cast light on a real^a problem—and not to aggravate such problem. Hence, for the sake of responsible research, we will now showcase only a few ‘practical’ AIA, having near-perfect success rate.

^aThe problem is real, and we demonstrated it. Our survey resembles DOTA2 population (Section 10.3.1), the statistical analysis proves the existence of correlations (Section 10.3.3) and our evaluation shows improvements over the baselines (Section 10.4).

10.5 PRACTICAL AIA (THE TRUE THREAT)

Insofar, the objective of our AIA was always to infer *each* class by independently considering every attribute. According to our threat model (Section 10.2.2), such AIA conformed to the targeted ‘one-to-one’ category: given *any* player, infer (*all* of) their attributes. The results (in Section 10.4), despite being arguably serious, may not induce real attackers to launch most of such AIA (aside from, perhaps, those on occupation): some players exhibit traits that are difficult to infer.

However, attackers can also launch two other categories of AIA, which can yield ‘devastating’ results while being surprisingly simple to carry out. As our fourth and last contribution, we now elucidate the effects of some indiscriminate ‘many-to-many’ AIA (Section 10.5.1), and of some targeted ‘many-to-one’ AIA (Section 10.5.2).

10.5.1 INDISCRIMINATE ‘MANY-TO-MANY’ AIA

Let us assume an attacker whose goal is to sell the inferred attributes to the black market. Such an attacker may want to advertise their data as being “most likely correct”; put differently, the attacker wants to ensure that the

inferred information is “unlikely to be completely incorrect”, thereby accepting some margin of error.

METHOD. We use exactly the same setup as in the ‘sophisticated’ AIA (Section 10.4.3), where the inference is done after analyzing multiple matches. However, for these AIA, we assume an attacker who is satisfied as long as the prediction is not completely wrong. For instance, assume that a player has ‘high’ openness (cf. Table 10.1): we consider the AIA to be successful if the probability associated to ‘high’ is either at the first or second place among all the possible classes (three in this case). A similar scenario describes an AIA in which the attacker wants to find, e.g., a player who is “likely to be open” (i.e., has ‘high’ openness either at the first or second place).

IMPACT. We report the results of these AIA (after using 30 matches) in the central column of Table 10.6, in which rows denote the attributes (we exclude those that only have two classes, as it would be unfair to include them); the leftmost column denotes the accuracy obtained by the sophisticated AIA (cf. Fig. 10.5), whereas the rightmost column denotes the improvement (as a flat difference). From Table 10.6, we can see a big jump in predictive accuracy with respect to Fig. 10.5. For instance, inferring age reaches 89% accuracy, whereas purchase_habits goes from 65% to 96% accuracy. Remarkably, this method is the only one that provides usable results for agreeableness and openness, both with ~80% accuracy. Despite bearing some intrinsic margin of errors (because the predicted class is not guaranteed to be the exact one), an attacker can still benefit from such imprecision, making these AIA a tangible threat.

Table 10.6: Indiscriminate ‘many-to-many’ AIA (mid column). Compared to the baseline (cf. Fig. 10.5), the accuracy substantially increases.

	Sophisticated AIA (30 matches)	Indiscriminate AIA (30 matches)	Improvement
age	67.15 \pm 6.87	89.15 \pm 4.66	+22.00%
purch.	68.99 \pm 3.81	96.13 \pm 2.86	+27.14%
open.	51.30 \pm 3.87	77.86 \pm 3.39	+26.56%
consc.	53.24 \pm 4.88	80.19 \pm 4.12	+26.95%
extrav.	53.78 \pm 3.90	81.51 \pm 4.40	+27.73%
agreeab.	50.71 \pm 4.65	76.84 \pm 5.59	+26.13%
neurot.	55.74 \pm 3.88	80.64 \pm 4.02	+24.90%

10.5.2 TARGETED ‘MANY-TO-ONE’ AIA

We now assume an attacker who wants to find players that present specific traits among a large population, e.g., finding very young players. In these cases, the attacker would train their ML models to maximize the *precision* on a given class, so as to minimize the amount of false positives. Although a similar strategy inevitably leads to a reduced *recall*, this is not an issue in reality: the attacker is not interested in, e.g., “finding *all* young players” (which is an unfeasible objective), but rather “finding a subset of those players that are *guaranteed* to be young”.

Such scenario is even more problematic than the previous ones, especially given that a low recall is not an issue when the population counts millions of players.

TARGETS. We consider an attacker that is interested in identifying four “vulnerable” groups of players¹⁴. Specifically: “very young” ($\text{age}=13-18$), “purchasers” ($\text{purchase_habits}=\text{Rarely} \vee \text{Regularly}$), and “introverts” ($\text{extraversion}=\text{Low}$.) Moreover, we also consider an attacker that attempts an ‘intersectional’ AIA, wherein the targeted group conforms to two specific classes of two *distinct* attributes. In this case, the attacker wants to pinpoint “purchasers & workers” ($\text{occupation}=\text{Yes}$, and $\text{purchase_habits}=\text{Rarely} \vee \text{Regularly}$), which could be ideal to identify players to which advertise new products—because such players tend to make purchases, and are likely to have the economical resources for doing so (as they have a job).

TESTBED. We adopt a similar setup of the sophisticated AIA (Section 10.4.3), i.e., we use $\overline{\mathcal{M}}$ as dataset, and evaluate the performance of our ML models as they analyze increasingly more matches of the same player, and then averaging the output probabilities. The crucial difference, however, lies in the problem formulation, which now reflects a *binary classification* setting: the objective is predicting the targeted class, and anything outside of such class is irrelevant. To this purpose, we first merge all players that do not belong to the targeted class (i.e., the “positive”) into a single class (i.e., the “negative”). Then, for each target, we train a (binary) classifier by using the precision as optimization metric (whereas in the sophisticated AIA, we used the macro F1-score). We find the best models and hyper-parameters using a validation set having players never seen at training time, simulating that the attacker can use only data that has gathered. The good results achieved on the validation set (combined with our correlation findings described in Section 10.3.3) suggest that the attack is feasible, and would incentivize the attackers to launch it in reality. Last, we evaluate the best models on the test set, having players not included in either the training or validation sets. For each targeted attribute, we repeat all these procedures five times to reduce bias and account for randomness.

IMPACT. We report in Fig. 10.6 the *precision* in identifying the targets as a function of the matches analyzed by the ML models.

It immediately stands out that we obtained much ‘dangerous’ results than in any of the previously considered scenarios. For instance, by analyzing 10 matches, our ML models can detect “very young” with almost perfect precision. Obviously, this comes at the cost of a low recall, which was about 47% after 30 matches.¹⁵ Moreover, our models ably detect “purchasers” after a single match, achieving a stable 90% precision—surprisingly exhibiting also a recall of 98% after 30 matches (not shown in Fig. 10.6), suggesting that purchasing indicators are well defined, and the mistakes happened probably when users are gifted expensive items. The models devoted to “introverts” achieve 76% precision (and 73% recall) after with 30 matches, indicating that players belonging to this group have many characteristics in common. Finally, for the ‘intersectional’ AIA focusing on “purchasers & workers”, the models obtain 86% precision (and 47% recall) after 30 matches, suggesting that roughly half of such players exhibit distinctive traits.

¹⁴There are over 8000 possible combinations of all our classes, and investigating all of them is clearly unfeasible and outside our scope.

¹⁵Roughly speaking, we detected half of the “very young”, but with no mistakes—i.e., the ML model found ~ 5 guaranteed “very young” out of ~ 81 players in the test-set.

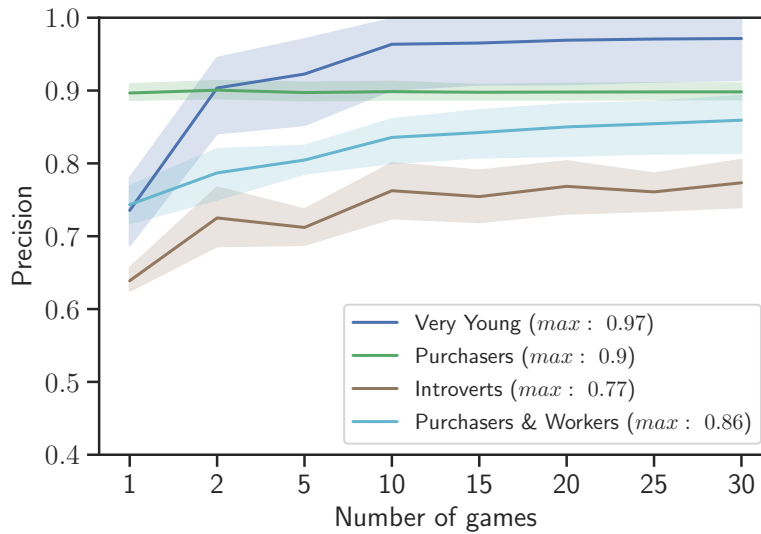


Figure 10.6: Targeted ‘many-to-one’ AIA. We train our ML models by maximizing the *precision* on a single targeted class. Such AIA are very effective after analyzing ~ 10 matches for each player in the test-set.

TAKEAWAY. Attackers with *specific* goals can easily setup AIA that are highly successful, thereby confirming the exposure of DOTA2 players to such privacy threat.

10.6 DISCUSSION

Our proactive evaluation showed that AIA can be highly successful in DOTA2. A legitimate observation is that our experiments consider a small subset of all DOTA2 players. However, our population still allows to derive statistically significant results (see Section 10.3.1). Another observation is that we (ethically) simulated an AIA by collecting personal attributes through a survey (instead of, e.g., scraping social networks [464]). However, as explained in Section 10.2.2, DOTA2 players are willing to participate in similar surveys (even when promoted by random users [3]). Hence, our (ethical) AIA represents a feasible scenario for an attacker, and our results are statistically significant. Finally, there exist infinite ways in which an attacker can use the collected data to carry out AIA; yet, those considered in this chapter confirm our point, i.e., that AIA are a threat to the DOTA2 player base.

We now discuss some possible mitigations (Section 10.6.1), and explain how our threat model can be applied to other E-Sports (Section 10.6.2).

10.6.1 COUNTERMEASURES TO AIA IN DOTA2

Our AIA are rooted in the fact that players’ in-game statistics are publicly obtainable from TW. The most obvious countermeasure would be denying public access to all such statistics *from the VG itself*. Unfortunately, players are the ones (implicitly) asking for such public availability (see Section 10.1.1). Alternatively, DOTA2 developers can

Table 10.7: Overview of E-Sports VG. Numbers are taken from various sources [4, 5, 6, 7, 8].

	Release Year	Genre	Monthly Players	Concurrent Players Avg	Playtime Avg (Hours)	Age Range (PEGI rec.)	Tournament Revenue	Exemplary TW	Replay System	Max Players per Lobby
<i>League of Legends</i>	2009	MOBA	127 M	700 K	832 H	11–50 (12+)	\$93 M	lolprofile.net	Yes	10
<i>CS:GO</i>	2012	FPS	34 M	560 K	611 H	13–40 (18+)	\$134 M	csgostats.gg	Yes	18
<i>Rocket League</i>	2016	Sport	90 M	25 K	315 H	6–35 (3+)	\$18 M	rltracker.pro	Yes	8
<i>Fortnite</i>	2017	Battle Royale	270 M	4 M	1800 H	6–54 (12+)	\$121 M	fortnitetracker.com	Yes	100
<i>PUBG</i>	2018	Battle Royale	510 M	200 K	356 H	12–55 (16+)	\$45 M	pubg.op.gg	Yes	100
<i>Apex Legends</i>	2019	Battle Royale	118 M	195 K	91 H	8–37 (16+)	\$10 M	apex.tracker.gg	No	60
DOTA2	2013	MOBA	15 M	450 K	1700 H	12–50 (12+)	\$283 M	opendota.com	Yes	10

use our analyses to make the features with stronger correlation to some attributes impossible to compute with public data; however, attackers are free to derive also other features—potentially with stronger correlations with (also) other attributes.

It is hence difficult to find a ‘general’ mitigation that preserves the functionalities of TW while ensuring players’ privacy. Yet, in an attempt to reduce the feasibility of an AIA, we propose two countermeasures. (1) *TW could allow players to select ‘what content’ is public.* For instance, a player can have only their last few matches to be visible by anyone. This solution has two drawbacks. First, if the statistics of *other* players in the same match are visible, an attacker could still launch an AIA—albeit at a higher cost, because they need to retrieve the information from the other players (of which they need to know the handle). (2) *TW could allow user to choose ‘who’ can see their profiles.* For instance, two players could browse each other’s statistics only if they are friends *within the VG*—which is a different environment than the TW (e.g., Fig. 10.1 shows the friends within the TW). Such a countermeasure requires, however, a deep cooperation between TW and the VG. Alternatively, visibility can be granted *upon request*.

Unfortunately, both countermeasures impair the use of TW to learn from other players, because their matches would be hidden. The only exception are professional players, whose profiles can be public since they are less likely to be targeted AIA in the first place.

Summary: Countermeasures against AIA present tradeoffs. Our work will hopefully inspire the search for a cost-effective solution.

10.6.2 EXTENSION TO OTHER E-SPORTS

Our threat model can cover also other VG beyond DOTA2. Indeed, we observe that our AIA necessitates access to in-game statistics, which are mainly retrievable through TW. However, **the existence of TW is not a strict requirement.** In fact, TW elaborate statistics and replays by directly interacting with the VG itself—because it is the VG that makes such data publicly available. Therefore, an attacker could harvest this information and elaborate it autonomously. Obviously, the amount of effort required in this scenario is much higher than relying on a TW, but an AIA would still be feasible (especially for targeted ‘one-to-one’ AIA).

Let us summarize the panorama of other E-Sports VG, for which we provide an overview in Table 10.7. All these VG present at least one TW akin to those of DOTA2 TW. Moreover, for all these VG, the in-game details of a player are public *by default* (except for DOTA2 and CS:GO), and they often have a replay system that could relax the requirement of a TW. We remark, however, that the other requirement for a successful AIA is the existence of a relationship between players’ in-game statistics and personal attributes. Although there is no proof (yet) of the

existence of such a relationship in other contexts, we believe in its existence. In fact, many DOTA2 features can be found in the other VG. Examples are the kill/death/assist ratio, paid subscription plans and cosmetics, chat usage, or information about the play-time. Finally, we highlight that players' of some VG (e.g., Fortnite) are children, increasing the risk of AIA [484, 486].

10.7 CONCLUSION

We address the problem of Attribute Inference Attack (AIA) in competitive video-games (VG), with a focus on DOTA2. We observe that DOTA2 players are naturally exposed to AIA due to the abundant in-game statistics that are publicly available. Based on this observation, we propose a threat model of AIA in DOTA2, and (ethically) evaluate its impact. Our results demonstrate that with little preparation and domain expertise, attackers can predict the personal attributes of DOTA2 players with high success (e.g., near-perfect precision). Countermeasures to such AIA are unfeasible due to tradeoffs that would disrupt the entire DOTA2 ecosystem.

By elucidating this subtle threat, which can affect also players of other VG, this work will hopefully inspire the development of effective mitigations (either by the VG producers, or by the TW administrators), therefore fostering an increased privacy of video gamers (who should be made aware of such risk).

ETHICAL CONSIDERATIONS

Our institutions do not require any formal IRB approval to carry out the experiments described herein. Nonetheless, our survey and corresponding evaluation are all performed by adhering to the guidelines of the Menlo report [141]. All interviewees were informed that their responses would be used for research purposes. Our questionnaire does not ask for sensitive data, or for private details such as name or address. We never released our dataset publicly (not even in anonymised form). All participants are also aware of the email address to contact should they be willing to have their entry removed from our dataset. Since our user-base is located in Europe, we also strictly complied with the GDPR, and all underage participants were located in countries which allowed their participation in research surveys without explicit parental consent [499]. For our AIA, we always infer the attributes that the participants of our survey willingly provided to us, hence there is no privacy violation. We do not attempt to infer personal attributes of players who did not participate in our survey (i.e., we do not collect in-game data of randomly chosen DOTA2 players, and use such data to infer their private information). The attributes we infer are non-sensitive.

11

You Can't Hide Behind Your Headset: User Profiling in Augmented and Virtual Reality

In recent years, the pandemic has increased the need for remote connections, and we have witnessed to mass adoption of virtual technologies, particularly for teamwork. Different platforms have opened up new perspectives for virtual interactions with others, and fostered the already ascending development of the Metaverse. The Metaverse has been recently defined as a "post-reality universe, a perceptual and persistent multiuser environment merging physical reality with digital virtuality" [500]. While being designed around the human, which constitutes the physical reality of this interplay, digital virtuality relies on immersive technologies that allow spatial and interactive features, namely Augmented Reality (AR) and Virtual Reality (VR), collectively known as Extended Reality (XR). Eventually, these devices became the core of the fourth wave of computing innovation [501].

Currently, there is an ongoing discussion on the potential protocols that will govern the Metaverse, with a particular focus on the controversial interplay between openness and privacy [500]. The latest virtual devices allow tracking many behavioral data, such as the headset's and controllers' position and rotation, or eye movements. All these data can induce leak of personal information, and even the user's identity (e.g., [502, 503, 504]). While remaining private, this information would help to restrict the use of the headset to specific individuals. For example, it would be possible to allow authentication only to those with appropriate permissions, thus increasing the security of such technologies.

To date, many studies demonstrated the feasibility of user profiling tasks in XR such as authentication [505, 506], users identification [504, 507, 503, 502], and gender inference [507]. Nevertheless, the variety of XR devices and interactions have led researchers to build specific profiling mechanisms for each of their experiments, which were conducted on a single technology and single (or few) actions. Indeed, creating an ad-hoc system for every situation requires significant effort [504, 503]. Moreover, if features are bound to a specific action (e.g., hands distance in a grab action), they will hardly generalize in different scenarios, with the risk of introducing bias. For

instance, Miller et al. [502] used the raw Y-axis of the Head Sensor (i.e., roughly the person’s height) as a principal descriptor for user identity. However, as pointed out by the authors and recent literature [508], such a feature is not *persistent*. Last, the comprehension of which factors impact profiling in XR technologies is currently limited. Indeed, the literature suggests that profiling performances might depend on the technology (i.e., AR, VR) [509], user actions [503], cognitive workload [510], experimental bias [508], and XR sensors [503], but they were never examined altogether.

CONTRIBUTIONS. In this work, we propose a comprehensive study of XR user profiling by leveraging behavioral data obtained through the use of VR and AR headsets. As a first contribution, we introduce a general profiling framework applicable to different virtual devices (e.g., VR, AR), applied fields (e.g., everyday use cases, work scenarios), and types of user behaviors (e.g., walking, searching, pointing). We test our framework on data from our previous works [510, 511], showing the generability of the approach. Since such previous studies revealed gender differences under diverse workload conditions, we additionally investigate the workload impact on profiling. Ultimately, our framework leverages task-independent and free-of-bias features, aiming to become a baseline for XR profiling.

As a second contribution, we implement our framework to study users’ profiling at different privacy levels (i.e., identification, personal information), introducing - to the best of our knowledge - the profiling of gender and age in virtual contexts through modern and widespread XR devices (AR Microsoft HoloLens, VR HTC VIVE Pro Eye).¹

As a third contribution, we explore the impact of device sensors on users’ profiling. Precisely, we assess the relevance of the headset’s position and rotation, the controllers’ position and rotation, and the eye tracker information available in the VR device. Last, we fill a gap in the literature on users’ profiling in AR scenarios, which is largely understudied compared to VR. Overall, we summarize our contributions as follows:

- we propose a general profiling framework for XR technologies, which can serve as a generic baseline for future XR profiling studies;
- we examine users’ profiling with respect to identification and private information (age and gender) in virtual scenarios, which is novel in the AR context;
- we introduce and explore the role of task workload in user profiling, which is a new concept in the area;
- we conduct extensive studies to assess sensors’ importance in the profiling tasks.

ORGANIZATION. In Section 11.1, we provide background and review literature on users’ profiling. Section 11.2 presents the general profiling framework we adopted in our experiments. The dataset and experimental settings are shown in Section 11.3 and Section 11.4, respectively. We report our results in Section 11.5, and discussion in Section 11.6. We conclude in Section 11.7.

¹Our VR device is a commercially widespread device that comes with an embedded eye-tracker, and as such, can potentially consolidate the findings of previous works based on hand-crafted devices.

11.1 BACKGROUND & RELATED WORK

This section describes the importance of security and privacy in XR technologies. Section 11.1.1 summarizes the application of virtual technologies in different fields, highlighting benefits deriving from user profiling. Section 11.1.2 introduces privacy in XR technologies, while Section 11.1.3 analyzes the state of the art in XR user profiling.

11.1.1 XR USE-CASES AND BENEFITS OF PROFILING

INDUSTRY AND REMOTE WORK. As Industry 4.0 progressed, virtual devices have proven their benefits in many sectors: in the design cycle of products and manufacturing systems [512], for programming machines [513], in the teleoperation industry [514, 515] and also for training novices [516, 517]. In any of these applications, virtual technologies provide the operator with a faithful virtual equivalent of the physical environment. Automatically identifying workers wearing headsets could improve workplace security. For example, authentication could be enabled only for those with appropriate permissions (e.g., site manager). Further, since older workers may prefer a different virtual environment design [518], user profiling could help customize virtual features according to age.

EDUCATION. Online education through virtual environments is one of the key pillars of Metaverse [500]. Several studies have examined how immersive virtual technologies are successfully integrated into education, as well as how they positively influence learning. Subject-specific benefits include improving skills, living more realistic experiences, and enhancing motivation and interest in learning [519]. Additionally, [520] assessed VR applications for higher education are becoming increasingly popular in engineering, medicine, and computer science education, and are mature enough to teach declarative, procedural, and practical skills. With XR being widely adopted in education, a profiling or identification algorithm comes in handy. For example, teaching methods and content can be tailored based on each student's needs or age.

GAMING AND ENTERTAINMENT. While VR games have been popular since the 1990s (e.g., Virtual Reality Gear [521]), AR has been gaining popularity since 2016 with Pokémon Go, Snapchat, Apple's ARKit, and Google.com's ARCore [522]. The sector is expected to grow exponentially, as it encompasses entertainment markets beyond gaming and arcades: the film and music industry, live show sectors and sports are just a few examples [523]. Following the pandemic's devastating effects in these markets, immersive virtual platforms can help support the cinema, music, and live-show industries [521, 524]. Last, the recent proliferation of virtual influencers (Chapter 2) demonstrates the importance of virtual technologies in both entertainment and marketing. Clearly, user profiling could be used for marketing strategies in this sector (e.g., delivering customized advertising). Further, particularly in gaming platforms, user identification might help detect banned individuals and prevent their access to virtual games.

MEDICINE. Both doctors and patients have found virtual technologies to be trustworthy. For instance, VR-simulated surgeries can be beneficial for medical education and training [525], while AR can support surgeries

by overlaying salient clinical records or visual aids over the patient’s body [526]. Virtual control systems for remote robotic surgery operations are also rising [527]. For patients, VR can help improve cognitive abilities after a traumatic brain injury [528] or increase engagement in Parkinson’s motor training [529]. Through identification, detecting whether a user is a surgeon or a student can restrict their rights during an XR surgical procedure. Similarly, profiling patients could allow training customization and automatic recordings of clinical improvements.

AR AS A SMART WEARABLE TECHNOLOGY. The latest AR smart glasses are fully wearable devices with computational functions, providing various functionalities by freeing the user’s hands [530]. For instance, Vuzix² developed AR smart glasses for navigation in unknown areas, while Zhao et al. developed an AR assistive navigation device [531]. Recently, Facebook has partnered with Ray-Ban and launched their Ray-Ban stories, which have raised important questions about ethical and privacy issues [532]. In the foreseeable future, the next generation of smart glasses will likely allow projecting e-mails and notifications from social networks on the user’s field of view. Reliable automatic identification of the user during everyday activities would allow private messages to be viewed only by the owner.

11.1.1.2 PRIVACY IN XR TECHNOLOGIES

Table 11.1: State of the art overview. Legend: ○ = AR, ● = VR, ● = AR & VR.

Reference	#Participants	Technology		Privacy Level			Sensors				Analysis		
		AR	VR	Age Authentication	Gender Identification	Head Position	Head Rotation	Eyes	Controller Position	Controller Rotation	Multiple Actions	Workload Impact	Sensors Importance
Roger et al. [504]	20	Google Glass			●	●	●	●					●
Li et al. [505]	95	Google Glass		●		●							
Mustafa et al. [506]	23		Google Cardboard VR	●		●	●						
Steil et al. [507]	20		Oculus DK2 + Pupil-Lab		●	●		●					
Pfeuffer et al. [503]	22		HTC Vive		●	●	●	●	●	●	●	●	●
Miller et al. [502]	511		HTC Vive		●	●	●	●	●	●	●	●	●
Liebers et al. [508]	16		Oculus Quest HMD		●	●	●	●	●	●	●	●	●
<i>Our</i>	34 (AR) and 35 (VR)	Microsoft HoloLens	HTC VIVE Pro	●	●	●	●	●	●	●	●	●	●

The increasing popularity of big data [533] coupled with the rapid adoption of various “smart” devices has resulted in parallel increases in privacy concerns. In today’s society, most people consider data collection incessant and believe that the risks outweigh any benefits [534]. To prevent (or at least reduce) the exposure of personal data, current and emerging technologies should support privacy by default [535], following recent legislation such as GDPR [536]. Fortunately, researchers are actively focusing on studying and adding a security and privacy level to XR and, more in general, emerging technologies. For instance, Adams et al. [537] deeply investigated VR security and privacy perceptions from users and developers, outlining a “code of ethics” for developers. Abraham et al. [538] interviewed XR experts from industry and academia to investigate issues relating to security, privacy, and influencing behavior, providing guidelines for future XR devices supporting security and privacy by default.

²<https://www.vuzix.com/>

Recent works [539, 540] deeply discussed security and privacy issues arising in the metaverse, allowing a better understanding and a consequent improvement of the technology concerning its users. Similarly, Nair et al. [541], proposed a system to browse metaverse in incognito, protecting their privacy from companies, surveillance agencies, or data brokers. Researchers have also focused on incorporating privacy-preserving measures on daily usage systems, such as authentication [542], and more recently, de-authentication techniques[30].

Besides protecting users' data from unwanted usage or sharing, past literature shows how attackers can use *public* data in unconventional ways to profile users or to infer *private* users' data (e.g., gender, age, personality traits). Examples include video games data [6], Social Networks interactions [467, 463], or online ratings [465]. The results of such studies highlight the high risks connected with public data availability, highlighting the need for further research to enhance user privacy.

11.1.3 USERS PROFILING IN AR AND VR APPLICATIONS

Few works discussed user profiling in AR and VR technologies, which are synthesized in Table 11.1. First, we classified previous works based on the *technology* (AR vs VR), given the diverse level of immersion they provide [509]. Second, we distinguished the *privacy level* they operate, i.e., whether they tackle private data profiling (age and gender), authentication, or identification. We remark that identifying a person (i.e., recognizing a given user among a group of known people) is substantially different than inferring their personal attributes (i.e., age, gender).³ Third, we considered the *sensors* they adopted for the profiling. Several works [504, 507, 503] built their algorithms on eye trackers, motivated by the connection found between eye movements and personal information [543, 507, 544, 545, 546]. However, researchers have proposed a variety of methods [547, 548, 549] to hide personal identifiers from eye movements, and XR devices integrate a greater number of sensors (e.g., gyroscope, accelerometer) which require additional studies. As we will demonstrate in our experiments (Section 11.5), eye movements are not strictly necessary for user profiling. Last, we report whether they tested their algorithms on *multiple actions* (i.e., generability) and evaluated the *sensors' importance*, factors that might affect the profiling performances [503]. As a novel point, we introduce the role of *cognitive workload* in profiling, since it affects how users interact with XR technologies [510].

The reader can notice that existing works demonstrated that user profiling in XR technologies is feasible, but to what extent, as well as the required conditions, is currently unclear. We briefly present the limitation of current literature, and how we address such gaps.

SINGLE TECHNOLOGY. Previous works focused solely on one technology, AR [504, 505] or VR [506, 507, 503, 502, 508], developing customized and task-dependent algorithms. Given AR and VR both aims to provide an immersive environment and embed similar sensors, future XR studies would highly benefit from a cross-technology profiling framework.

LIMITED PRIVACY UNDERSTANDING. Researchers tackled mainly a single privacy level profiling, ignoring other privacy issues associated with XR devices. For instance, to the best of our knowledge, there are no at-

³For instance, we might identify a person within a population by their surname, which is uncorrelated to their age or gender.

tempts in the literature to infer users' private data (e.g., age, gender) from modern XR devices. Indeed, many works [543, 537, 538] theorized that private data inference in XR was possible based on eye trackers studies [550], but none of these theories were empirically proven. The only evidence of gender profiling comes from Steil et al. [507], who purposely equipped the VR headset Oculus DK2 (2016) with an eye-tracker (Pupil⁴).

RESTRICTED SET OF SENSORS. The most impacting results were gained primarily by leveraging eye-movement features [504, 503]. Others leveraged different behavioral features such as head position and rotation [504, 505, 506, 503, 502], often being prone to experimental bias [508] (see Section 11.2.3). Therefore, it is still unclear how different features contribute to the accuracy of a profiling task, nor if the feature choice should be task-dependent.

LACK OF GENERABILITY. Only two works [503, 502] tested their algorithms on multiple actions, questioning their generability. In AR, no works tested generability. We also noticed that no works analyzed the actions' cognitive workload impact, which was demonstrated to be crucial in XR interactions [510].

It follows that testing a general framework, which (1) leverages the same algorithms for profiling users in all XR technologies, (2) systematically considers multiple features, (3) extends to different levels of profiling tasks (identification, private data inference), and (4) works for multiple actions, might be helpful in view of higher generability and broader comprehension of XR user profiling.

11.2 METHODOLOGY

This section describes our methodology to execute user profiling within virtual technologies. Section 11.2.1 motivates the reasons for our investigation. The overview of our proposed framework is presented in Section 11.2.2, while the details are provided in Section 11.2.3.

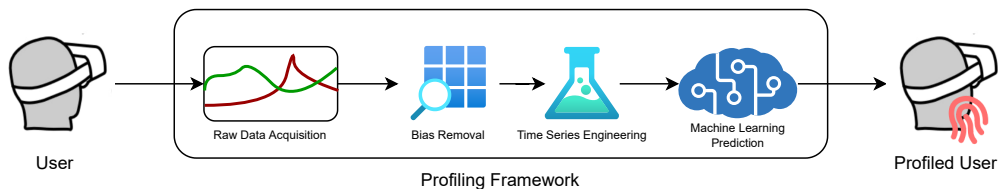


Figure 11.1: Overview of the proposed framework for user profiling in Augmented and Virtual Reality.

11.2.1 SCOPE OF THE WORK

This study examines whether users can be profiled by leveraging their interaction with AR and VR devices. In particular, we consider two privacy levels of user profiling:

1. *User identification*, where we aim to identify a given user within a known population;

⁴<https://pupil-labs.com/products/vr-ar/>

2. *Private information inference*, where we aim to infer users' gender and age.

Thus, we propose a general framework to accomplish both tasks, extendable to infer additional users' information. Further, our framework requires to:

- work across different XR devices and actions;
- reduce the experimental bias by leveraging features uncorrelated from the task.

By satisfying these requirements, our framework can become a simple yet effective baseline to test user profiling over general XR devices and applications. The use of a generic-purpose framework can indeed simplify future research and comparison between multiple applications and devices.

11.2.2 INFERENCE FRAMEWORK OVERVIEW

Our goal is to define a generic pipeline that can be adapted and applied to any virtual technology (e.g., AR, VR) context to profile a user, in terms of identification or private information. As shown in Figure 11.1, the pipeline consists of four steps, starting from the *user* from whom we record the behaviors, to their actual profiling:

1. *Raw Data Acquisition*. In this phase, users' behavioral data are acquired. XR technologies' devices continuously generate data from users' interactions with the virtual environment (i.e., time series). From these data, we can describe users' behavior. The amount and type of information depend on the virtual technology and its devices. For instance, data might come from users' input (e.g., pressing joystick's buttons) and users' movements.
2. *Bias Removal*. This phase aims to remove potential biases from time series that might lead to train erroneous machine learning models.
3. *Time Series Engineering*. This phase aims to extract insightful information from the time series.
4. *Machine Learning Prediction*. This phase aims to infer users' private information from the data elaborated in the previous phase by leveraging machine learning algorithms.

11.2.3 FRAMEWORK DETAILED DESCRIPTION

RAW DATA ACQUISITION

Users interact with AR and VR applications through devices such as headsets and joysticks. These devices embed several functional sensors to offer users an immersive experience. For example, users move and explore the virtual environment through sensors like accelerometers and gyroscopes embedded in the headset. Thus, by combining information retrievable by each sensor s^i of the equipment, we can trace users activity a at a given time t :

$$\vec{a}_t = [s_t^0, s_t^1, \dots, s_t^n], \quad (11.1)$$

where the subscript denotes the timestamp and the superscript the sensor involved. We call this process *acquisition phase*. The acquisition phase can be repeated over time, resulting in a temporal user-behavioral description. Thus,

by acquiring data in $\Delta t = t - t_0$, we obtain a behavioral time series, described as follows:

$$\vec{\mathbf{B}}_{\Delta t} = [\vec{a}_{t_0}, \vec{a}_{t_1}, \dots, \vec{a}_{t-1}, \vec{a}_t]. \quad (11.2)$$

$\vec{\mathbf{B}}_{\Delta t}$ represents an atomic sample of a user action (or task) of duration Δt that we will use in the next phases to infer their private information.

BIAS REMOVAL

The acquisition phase might lead to an enormous quantity of raw data. Such data might describe not only users' behavior, but also environmental information strongly correlated to experimental sessions. For example, using the raw headset height to identify users might be erroneous since such information might not be persistent over time (e.g., different shoes, different body position) [502]. The problem of *spurious correlations* in cybersecurity applications is well known [479]. Thus, care must be taken to understand if sensors might lead to erroneous and inconsistent machine learning performance. The process of bias removal depends on the sensors' nature and requires an ad-hoc analysis. We explain in detail our implementation in Section 11.4.2. The de-biasing phase results in a new vector of de-biased actions:

$$\vec{\mathbf{B}}_{\Delta t} = [\vec{d}_{t_0}, \vec{d}_{t_1}, \dots, \vec{d}_{t-1}, \vec{d}_t], \quad (11.3)$$

where d_{t_i} is the de-biased version of the feature a_{t_i} .

TIME SERIES ENGINEERING

Raw temporal data should be properly elaborated to extract meaningful information. Moreover, given the vast amount of data, such sequences should be aggregated (i.e., compressed) to limit the computational cost of their analyses. The aggregation strategy can consider the whole sequence of specific features, or just a subpart of it. For example, given a sensor $s_{\Delta t}^i$ and its de-biased values over the time $d_{\Delta t}^i = [d_{t_0}^i, d_{t_1}^i, \dots, d_{t-1}^i, d_t^i]$, the aggregation of a whole sequence results in a unique number x^i , while the partial aggregation (e.g., a transformation every q times step) in a vector of numbers $[x_0^i, x_1^i, \dots, x_m^i]$, where $m = t/q$. Note that the subscript does not denote the temporal axis anymore. Popular features derived from the aggregation phase are the mean, standard deviation, min, max [502]. At the end of the process, we obtain, for each participant action or task, an aggregated datapoint $x = [x^0, x^1, \dots, x^n]$ that will be used by the machine learning models.

MACHINE LEARNING

The last phase of the pipeline involves machine learning approaches like Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). Training a well-performing model requires validation strategies that consider the nature of the inference. For instance, if the aim is to identify a user within a known population, the training, validation, and testing splits should contain samples of the whole population. However, to avoid trial (or session) bias, the three splits should consider samples from different collection trials. Conversely, when inferring information like age and gender, the three splits should contain different sets of users, since we want to infer the

characteristics of people not seen at training (and validation) time. Regarding the type of machine learning algorithm, we suggest the use of *inherently interpretable* models (e.g, LR, DT) to better understand models' decisions during inference. Moreover, interpretable models allow a transparent debugging phase to identify the presence of spurious features [551]. Finally, given the unbalanced nature of the problem (i.e, not all the classes are distributed equally), we suggest using performance metrics like F1-score with macro average.

11.3 DATASET OVERVIEW

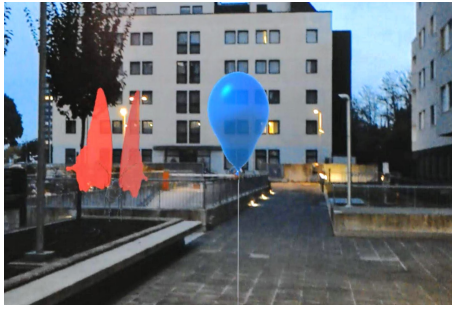
Our previous studies assessed behavioral and workload aspects in individuals using AR while walking outdoor [510], and in users wearing VR for guiding an industrial robotic arm [511, 552]. In the present work, we leverage behavioral and eye-tracking data of both the AR and VR scenarios with the purpose of profiling users. For each technology, we considered tasks and actions to test the generability of our profiling approach and study the conditions or actions which might be more (or less) successful.

11.3.1 AR EXPERIMENT

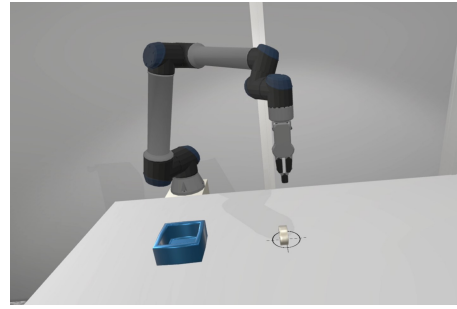
OVERVIEW. The AR experiment investigated multitasking effects in participants using AR while walking outdoor [510]. Participants wore the Microsoft HoloLens 1st generation smart glasses, and interacted with the augmented targets both via an Xbox One controller and physical collision with the virtual objects (e.g., walking through an augmented target). They performed: i) a *visual task*, in which they discriminated between different augmented targets presented in their peripheral view, ii) a *navigation task*, in which they reached a series of augmented landmarks via physical walking outdoor, and iii) the combination of these tasks, i.e., a *dual-task*. For more details about the tasks, please refer to [510]. Figure 11.2a shows an example of the virtual environment. Each participant performed 80 trials of the visual task, 50 trials of the navigation task, and 50 trials of the dual-task. While the original dataset was composed of 45 participants, we excluded 11 participants whose headset position data were not correctly recorded, and finally run our analyses on 34 participants (10 females age mean = 24.28, SD = 2.22 - 24 males age mean = 24, SD = 2.62). We continuously recorded through the device (60 Hz) the following measures: position (in meters) of the AR headset in the three axes (x, y, z), and rotation of the AR headset in Euler angles.

TASKS. From the experimental design, we identified the following tasks:

- *Mental Task (MT)*. The mental task corresponds to what [510] call Visual Discrimination Task. Specifically, participants were discriminating between different colored and lateralized augmented objects while standing still.
- *Navigation Task - Low workload (NT-Low)*. Participants were looking for augmented targets in their surroundings and then walked through them.
- *Navigation Task - High workload (NT-High)*. Participants were executing the navigation task concurrently with the mental task. The concurrent execution of two tasks is known as “dual-task paradigm” and is usually deployed in cognitive science research to create higher mental demand on the participant.



(a) Augmented Reality



(b) Virtual Reality






Figure 11.2: Virtual environments adopted in the experiments.

ACTIONS. Each task is composed of smaller operations that we named actions. The dataset contains the following actions:

- *Button interaction.* Participants were standing still while discriminating between the lateralized colored targets. Specifically, they were instructed to press specific buttons on the joystick according to the hemi-field where the virtual object was displayed.
- *Search.* Participants were engaged in the visual inspection of the surroundings to find a virtual landmark; this action was performed while participants were standing still and just rotated their heads to inspect the surrounding.
- *Walk.* Participants were physically walking to the identified virtual landmark.

Both the search and walk actions were performed as single-task and concurrently with the secondary mental task (namely, the visual discrimination task). Based on the results obtained in our previous work [510], participants perceived a lower workload in the single-task compared to the dual-task. Therefore, we here refer to the dual-task as the high workload condition, while the single-task is considered as a low workload condition. Furthermore, the button interaction action was categorized as a high workload condition since it entailed high and sustained attentional processes for correctly discriminating the stimuli appearing lateralized to the participant's field of view. Table 11.2 represents the actions isolated in the AR environment.

Table 11.2: Augmented Reality actions organized per type of action and workload level.

<i>Workload</i> \ <i>Action</i>	<i>Button Interaction</i>	<i>Search</i>	<i>Walk</i>
<i>Low</i>	—		
<i>High</i>			

11.3.2 VR EXPERIMENT

OVERVIEW. As part of the VR experiment, participants guided a virtual replica of an industrial robotic arm (Universal Robot e-Series UR5e) developed in Unity [511]. They were equipped with an HTC VIVE Pro Eye VR device and two VR controllers and guided the robotic arm shown in figure 11.2b through a pick-and-place, i.e., picking and placing a bolt into a box. They performed the task using two control systems (controller buttons and physical actions, i.e., moving their hands) and under two levels of workload (single-task and dual-task). In the dual-task, participants also performed simple arithmetic sums, typing the results on a virtual keyboard by pointing the controller. Further details about the task design can be found in [52]. In each condition, the young participants performed 40 trials, while the old participants performed 20 trials. In total, 35 participants executed this task (18 females, age mean = 39.33, SD = 14.21 – 17 males, age mean = 37.75, SD = 16.32). The following measures were continuously recorded through the device (120 Hz): position (in meters) in the three axes (x, y, z), rotation in Euler angles of both the VR headset and its controllers, pupil size (in millimeters), and eye openness (expressed from 0 to 1).

TASKS. From the experimental design, we identified the following tasks:







- *Controller-based Task - Low workload (CT-Low).* Participants guided the robot via controller buttons under a low workload;
- *Controller-based Task - High workload (CT-High).* Participants guided the robot via controller buttons under a high workload (i.e., while also calculating sums);
- *Action-based Task - Low workload (AT-Low).* Participants guided the robot via physical actions under a low workload;
- *Action-based Task - High workload (AT-High).* Participants guided the robot via physical actions under a low workload (i.e., while also calculating sums).

ACTIONS. From the tasks performed in VR, we extracted the following actions:

- *Idle.* Participants were only looking at the robot while it was executing either a pick or place automation, i.e., were not interacting with the virtual environment;
- *Pointing.* Participants were using the VR controller to point the numbers on the virtual keyboard to report the result of the arithmetical sums;
- *Button Interaction.* Participants guided the virtual robot through the pick-and-place task by only pressing specific buttons on the VR controller;
- *Physical interaction.* Participants physically touched the virtual robot and moved their arms to relocate it over the worktable.

In line with our previous findings [511], the actions performed concurrently with the arithmetic task were categorized under high workload. Differently, when performed without any additional task, they were categorized under the low workload. Table 11.3 represents the actions isolated in the VR environment.

Table 11.3: Virtual Reality actions organized per type of action and workload level.

<i>Workload</i>	<i>Action</i>	<i>Idle</i>	<i>Pointing</i>	<i>Button Interaction</i>	<i>Physical Interaction</i>
	<i>Low</i>				
<i>High</i>			-		

11.3.3 ETHICS

The data for this study come from our previous works, which were approved by the internal ethical committee of the University of Padova, Italy. Participants signed informed consent. The ethical committee approved the possibility of sharing anonymized data with other researchers to foster transparency, reproducibility, and further research.

11.4 EXPERIMENTAL SETTING

This section describes our experimental settings. Section 11.4.1 explains the targets of our profiling, while Section 11.4.2 describes the implementation (i.e., de-biasing, feature extraction, model selection).

11.4.1 PROFILING TARGETS

In our experiments, we are interested in the identification, age, and gender profiling processes. In each of them, we use the headset’s data the user generates when interacting with the XR environment (i.e., behavioral data) to predict a target (i.e., the user identity, gender, and age). These processes will be performed on each task and action presented in Section 11.3. We now describe in detail the three processes.

IDENTIFICATION. This process aims to identify a particular user among a group of known users. In this setting, every user appears in both training, validation, and testing data. Therefore, the training set contains the behavioral data of all the users. First, we train an ML model able to map a user’s behavioral data to their identity. Then, when we present the ML model with new (unknown) behavioral data, it identifies the user who generated them.

AGE PROFILING. This process aims to infer the user’s age starting from their behavioral data. In this setting, users appear only in one of the training, validation, and testing set. Therefore, the training set contains the behavioral data of only a subset of users. First, we train an ML model able to map a user’s behavioral data to their age. Then, when we present the ML model with the behavioral data of a new user (unknown), the model infers their age.

GENDER PROFILING. This process aims to infer the user’s gender starting from their behavioral data. In this setting, users appear only in one of the training, validation, and testing set. Therefore, the training set contains the behavioral data of only a subset of users. First, we train an ML model able to map a user’s behavioral data to their gender. Then, when we present the ML model with the behavioral data of a new user (unknown), the model infers their gender.

We remark that age and gender profiling are substantially different from identification. Indeed, in the identification process, the ML model works with data of *known* users, while in age and gender profiling, the aim is to infer the targets of *unknown* users.⁵ In other words, identification can be used only when the population is known (e.g., within a family context), while age and gender profiling can be used when the population is unknown (e.g., when a customer wears the device for the first time).

11.4.2 IMPLEMENTATION

DE-BIASING AND FEATURE EXTRACTION

AR and VR datasets contain different types of raw features acquired from the sensors. We now describe, for each category of sensors, the features and de-biasing techniques we applied.

- **Head Position (AR and VR)**, represented as a 3D coordinate (x, y, z) measuring the relative distance (in meters) of the user from a center point in the virtual environment. This feature might contain both session’s and users’ static traits (e.g., height). We thus derived different variants of this information, such as the movement, computed as the norm between two points at 5 timestamps of distance, and the vertical oscillation computed as the difference between two height values at 5 timestamps of distance.
- **Head Rotation (AR and VR)**, represented as a 3D value. For each axis, we compute its angular speed by considering points at 5 timestamps of distance. This transformation can remove information related to trials (e.g., specific positioning of objects with respect to the participant).
- **Eyes (VR)**, includes data on pupil size (in millimeters) and eye openness (0-1), for both left and right eyes. In order to overcome possible confounding variables [553] [554], it is usually appropriate to pre-process the raw eye data for flattening individual differences. However, as the aim of the present work was specifically to capture individual traits and behaviors for allowing identification/profiling, we opted for not pre-processing eye-tracking data. On the contrary, we leveraged the individual differences in pupil size and eye openness [555] [556] [557] for better identifying and profiling users. Further, we enhance this set of features by computing the symmetry among the eyes for both pupil dilatation and eye openness. On an applied level, using the raw output of the HTC Vive Pro Eye device speeds up the identification/profiling process and allows higher generability to multiple VR devices.
- **Controller Position (VR)**, represented as 3D coordinates (x, y, z) relative to the virtual environment center point. Similarly to the head position, this feature might contain both sessions and user traits.

⁵Having the same users in training and test data when performing private data inference causes overfitting, since it degenerates into an identification task.

We thus transform it in the movement, computed as the norm between two points at 5 timestamps of distance.

- Controller Rotation (VR) represented as 3D value. We conduct the same process of head rotation.

Finally, each feature of the previously described families is aggregated with `tsfresh`⁶. Given a time series, this library extracts more than 100 features, including average, standard deviation, quantile, and entropy. We further refined the features by keeping only the relevant ones.⁷ Thus, starting from the raw time series of a single action within a single task performed in a single trial by a single user, we extract a single aggregated data point. The process is repeated for all the users, trials, actions, and tasks, obtaining 9360 datapoints in AR, and 16520 datapoints in VR.

MODELS TRAINING AND VALIDATION

In our experiments, we test four different algorithms: logistic regression, ridge classifier, decision tree, and random forest. As a baseline, we defined a Dummy classifier that randomly predicts the outcome based on the training ground-truth distribution. For each experiment presented in Section 11.5, we adopt a common validation strategy: for each discussed model, we find the best hyper-parameters through a grid-search validation based on training, validation, and testing split of 70%, 10%, and 20% of samples, respectively. For private inferring tasks (i.e., age and gender), the splits contain different sets of users, i.e., users in training are not present in the validation and testing set. Similarly, users in validation are not present in both training and testing sets. Machine learning models are designed as a multiclass classification problem for the user identification task. On the opposite, we considered a binary classification problem for both age (i.e., young and old) and gender (i.e., male and female). Note that the young class correspond to users defined in [19 – 24] (AR) and [23 – 30]; the old class is defined in [25, 29] (AR) and [31 – 69]. We now report the parameter grids involved in the grid search:

- Logistic Regression (LR). C: 0.1, 1, 10.
- Ridge (RI). Alpha: 0.01, 0.1, 1., 10. Fit intercept: *False, True*.
- Decision Tree (DT). Max Depth: 3, 5, 7. Min samples leaf: 1, 3, 5.
- Random Forest (RF). N estimators: 50, 100, 150. Max Depth: 3, 5, 7. Min samples leaf: 1, 3, 5.

To provide accurate results, each experiment is repeated five times. We thus report both the mean and standard deviation of the F1-scores (with macro average). We implemented our experiments in Python 3.8.5 and we used Scikit-Learn [558] library for training models and validation algorithms.

11.5 RESULTS

In this section, we present the results of our experiments. We present both results for the task and action levels, in sections 11.5.1 and 11.5.2, respectively. We then conclude with an ablation study to determine the effect of

⁶<https://tsfresh.readthedocs.io/en/latest/index.html>

⁷We used `tfresh` `feature_selection` function: https://tsfresh.readthedocs.io/en/latest/api/tsfresh.feature_selection.html

different sensors on models' performance (Section 11.5.3),

11.5.1 TASK-LEVEL

In this section, we present profiling performance at a task-level. In particular, each presented experiment considers distinctly the tasks presented in sections 11.3.1 and 11.3.2. In more detail, we train, validate, and test our model only on the task under investigation, predicting each time the identity, age, and gender separately. For instance, we train a specific model to predict gender based only on the Mental Task.

IDENTIFICATION

Figure 11.3 shows the identification results in AR and VR environments. LR and RI achieved the highest (and comparable) performances in AR, whereas LR and RF performed best in VR. In general, all our algorithms outperform the baseline (Dummy). Looking at the results on the Overall Tasks, both in VR (OT-VR) and AR (OT-AR), we immediately notice that in VR identification, the performances remain pretty stable as the number of users increases, while AR degrades significantly. Indeed, AR best algorithms performance goes from nearly 0.90 F1-Score (two users) to slightly above 0.60 F1-Score (30 users). Instead, in VR, LR yields almost perfect prediction on two users, while the F1-Score is above 0.95 when performing identification over 30 users. This might reflect the different amount of sensors available in VR (headset, controller, and eye-related behaviors) compared to those available in AR (only headset-related behaviors). We further discuss the impact of each of the involved sensors in Section 11.5.3.

When looking at the individual tasks, we can see that the identification algorithm performs even better than the overall task, particularly in AR. For instance, we reached 0.70 F1-Score over 30 users in the NT-Low, which is roughly 0.10 higher than in the OT-AR. One reason for this result might be related to the nature of the performed task: in the NT-Low, participants were actively moving in the surroundings without performing any additional task. Therefore, their movements might have been more linear compared to the situation in which they performed the same task under a high workload (NT-High), thus revealing more identifiable movement patterns. The same does not apply to the VR scenario. Here, when looking at each of the identified actions, the higher the workload the better the performance of the identification algorithm. Indeed, the best performance was obtained at the AT-High and CT-High, where the F1-Score was around 0.95 and 0.97, respectively. Again, possible explanations might be related to the nature of the tasks and the number of sensors embedded in the devices. In the VR scenario, participants were only moving their upper body, and in the high workload conditions they were additionally engaged in a secondary mental task. We know from the literature that a higher workload is related to higher changes in eye behavior [511]. Therefore, the VR-embedded eye-tracker might have had an essential impact on the identification performance, mainly when users were under higher mental strain rather than when performing less demanding tasks (i.e., CT-Low, AT-Low).

AGE

Figure 11.4 shows the age classification results in AR and VR environments at task-level. Results from the age profiling clearly yielded better performance in the VR compared to the AR scenario. While in VR all models

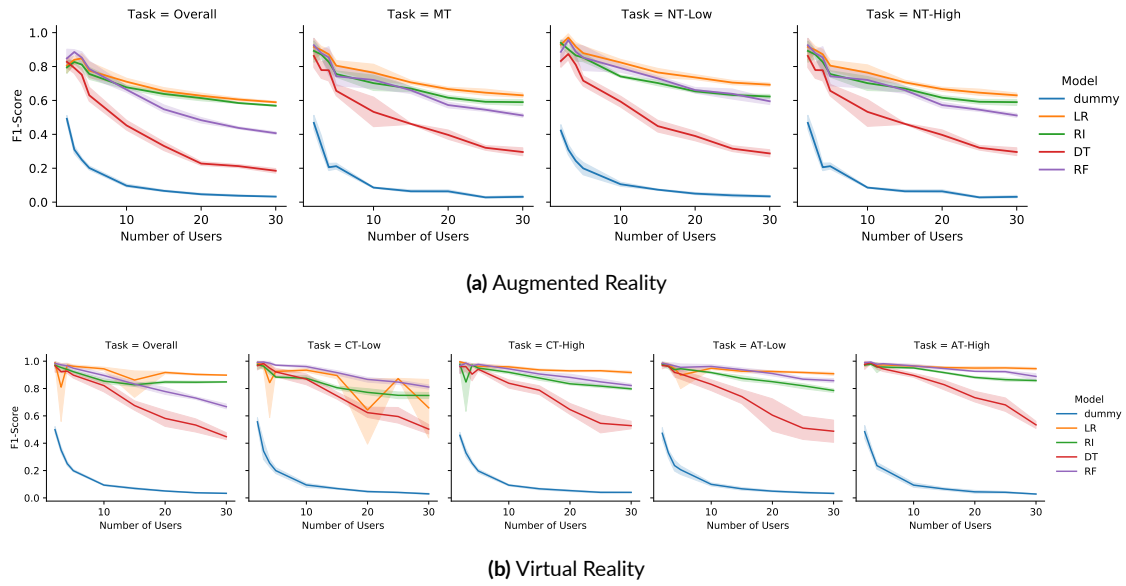


Figure 11.3: User Identification on task-level.

performed significantly better than the baseline, in AR the F1-Score was consistently lower than the baseline, in all tasks. This is likely to be related to the low age variability of participants that took part in the AR experiment, or the inadequacy of sensors (see Section 11.5.3). While this is a clear limitation of our study, such a result is still valuable, since it suggests that people of similar ages interact similarly with AR devices, meaning that age profiling may not be possible in every circumstance. Therefore, we focus our discussion mainly on age profiling performances in relation to the VR experiment.

In VR, the LR and RF algorithms appear to perform better than the other models in all tasks, but in the OT, where RI produced a higher F1-Score compared to LR. On the task-level, the users' age was profiled with higher accuracy when they performed the pick-and-place task via physical actions (AT-High and AT-Low, in which F1-Score was around 0.90 and 0.85 respectively) compared to controller buttons (CB-High, CB-Low, in which F1-Score was below 0.80 in both cases). A possible interpretation on this point is that the movement patterns of older users might have been quite different from younger users. Also, we know from the literature that robot teleoperation is significantly influenced by age [559]. In this view, our algorithm was particularly successful in detecting users' age during the pick-and-place task only when physical actions were involved.

GENDER

Figure 11.5 shows the gender classification results in AR and VR environments at task-level. When profiling users' gender, we obtained substantially better results in VR compared to AR. Indeed, in VR, all the tested algorithms performed above the baseline (dummy). More specifically, we can observe a better performance obtained through LR and RF, which reached a maximum F1-Score of 0.75. Differently, when detecting users' gender in the AR scenario, our algorithms performed only 0.5-0.10 above the baseline. This discrepancy could be explained by the inadequacy of sensors (see Section 11.5.3)

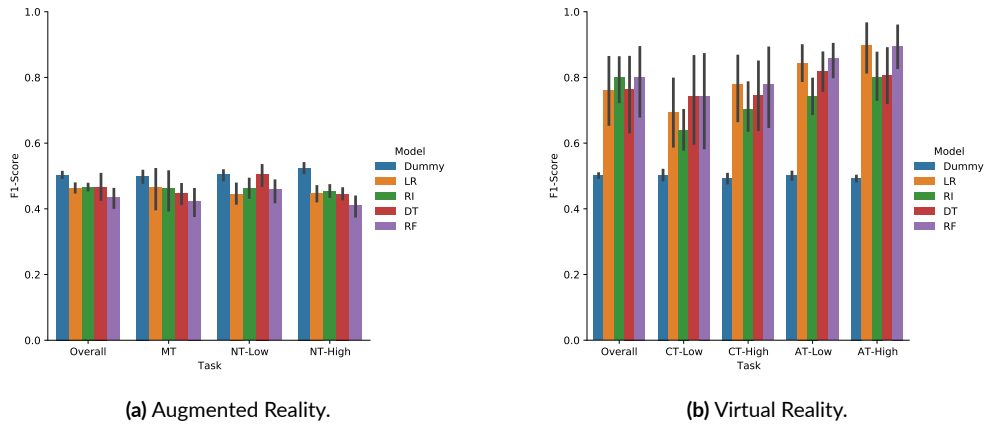


Figure 11.4: Age profiling on task-level.

In VR, we achieved better performance in tasks involving a higher workload (CT-High, AT-High) than those under a low workload (CT-Low, AT-Low). These results align with recent literature on behavioral gender differences in the VR pick-and-place task. For instance, our previous work [552] demonstrated how men outperformed women in the pick-and-place tasks in terms of task execution time, particularly when using controller buttons. These differences might have been even more marked when performing an additional mental task, thus allowing more precise gender profiling. We observe a similar trend in the AR scenario, where higher workloads (NT-High) result in better performance. This behavior reflects previous findings related to the different walking patterns between men and women [510]. Indeed, on average, the walking velocity of men is significantly higher than women’s one, particularly under high workloads. As we recorded the headset shifts in time, the different walking velocities might have been prominent in gender profiling.

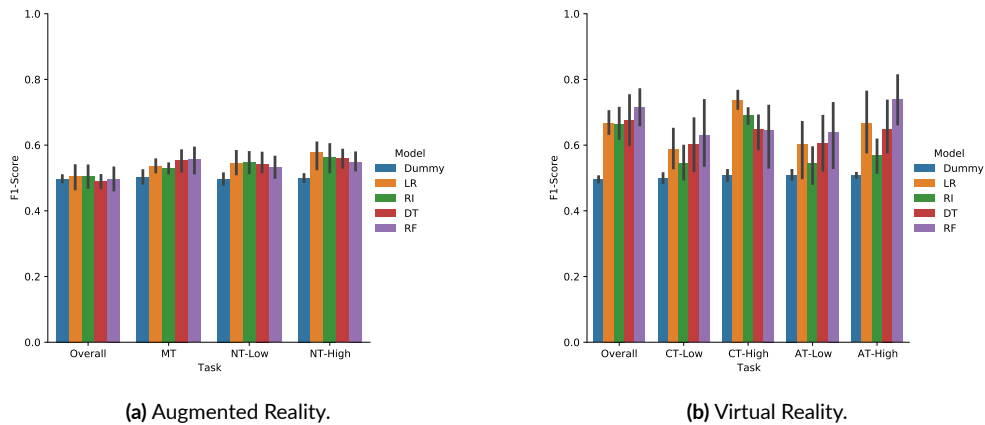


Figure 11.5: Gender profiling on task-level.

11.5.2 ACTION-LEVEL

Starting from the results obtained in the overall task, we investigated whether some actions had a particular effect on the identification and profiling performances. Specifically, we opted for leveraging the model that demonstrated better results, which was the Logistic Regression (LR). Each presented experiment considers distinctly the actions presented in sections 11.3.1 and 11.3.2. In more detail, we train, validate, and test our model only on the action under investigation, predicting each time the identity, age, and gender separately. For instance, we train a specific model to predict age based only on Button Interaction with Low Workload.

IDENTIFICATION

Table 11.4 shows the identification results in AR and VR environments at action-level. Previously at task-level we reached an F1-Score of about 0.60 in the AR and above 0.90 in the VR scenario. Looking at the action-level, specifically for AR, we see that the walking action reaches the highest performance (F1-Score is about 0.80 under low workload and 0.78 under high workload), while the search action and button interaction reveal F1-Scores below 0.70. This suggests that the walking action is prominent in identifying users in AR, possibly because the walking pattern is the most singular feature in such a use-case of AR. Differently, in VR, we observe higher F1-Scores for both button and physical interactions, specifically under high workload (F1-Score is about 0.96 in both cases). Also, the pointing action reached a very similar F1-Score (0.96), while the idle time intervals yield lower F1-Scores (below 0.80 both under high and low workloads). It seems that the most interactive actions (using controller buttons, pointing, and physically moving the upper body) yield better results compared to periods in which users were passively looking at the virtual surroundings.

Table 11.4: User identification on action-level organized per type of operation and workload level. Random guess at 0.03 for both AR and VR tasks. All the measures in F1-Score.

		<i>Augmented Reality</i>			<i>Virtual Reality</i>			
		<i>Button Interaction</i>	<i>Search</i>	<i>Walk</i>	<i>Idle</i>	<i>Pointing</i>	<i>Button Interaction</i>	<i>Physical Interaction</i>
<i>Workload</i>	<i>Low</i>	–	0.66±0.03	0.80±0.02	0.78±0.02	0.96±0.01	0.92±0.01	0.93±0.02
	<i>High</i>	0.61±0.02	0.69±0.01	0.78±0.02	0.86±0.01	–	0.96±0.00	0.96±0.01

AGE

Table 11.5 shows the age classification results in AR and VR environments at action-level. Users' age was profiled with an F1-Score of about 0.50 on the overall task executed in AR, and 0.80 in VR. As the age profiling was unsuccessful in AR, we will not pay close attention to the action-level results of this use case. These results confirm what we observed at task-level (see Figure 11.4). Regarding the VR scenario, we can note that, under low workload, the pointing (F1-Score = 0.88) and physical interactions (F1-Score = 0.82) were the most crucial in profiling users' age, compared to actions allowing less interactivity with the virtual environment (F1-Scores below 0.80). This might suggest a different movement and interaction pattern shown by older and younger users, especially when greater freedom of movement is allowed. This is also in line with what was observed on task-level. Moreover, this

trend becomes even more evident when the physical interactions are performed under a high workload (F1-Score = 0.90), likely reflecting the multitasking and motor difficulties related to age [560].

Table 11.5: Age profiling on action-level organized per type of operation and workload level. Random guess at 0.5 for both AR and VR tasks. All the measures in F1-Score.

		<i>Augmented Reality</i>			<i>Virtual Reality</i>			
<i>Workload</i> \ <i>Action</i>		<i>Button Interaction</i>	<i>Search</i>	<i>Walk</i>	<i>Idle</i>	<i>Pointing</i>	<i>Button Interaction</i>	<i>Physical Interaction</i>
Low		–	0.40±0.03	0.45±0.02	0.77±0.10	0.88±0.06	0.70±0.09	0.82±0.05
High		0.47±0.02	0.44±0.01	0.49±0.02	0.83±0.09	–	0.81±0.07	0.90±0.05

Table 11.6: Gender profiling on action-level organized per type of operation and workload level. Random guess at 0.5 for both AR and VR tasks. All the measures in F1-Score.

		<i>Augmented Reality</i>			<i>Virtual Reality</i>			
<i>Workload</i> \ <i>Action</i>		<i>Button Interaction</i>	<i>Search</i>	<i>Walk</i>	<i>Idle</i>	<i>Pointing</i>	<i>Button Interaction</i>	<i>Physical Interaction</i>
Low		–	0.50±0.02	0.45±0.06	0.60±0.10	0.82±0.09	0.62±0.05	0.66±0.11
High		0.54±0.03	0.58±0.03	0.60±0.06	0.63±0.05	–	0.74±0.06	0.66±0.08

GENDER

Table 11.6 shows the gender classification results in AR and VR environments at action-level. On task-level, our algorithms reached an F1-Score of about 0.50 in AR and above 0.70 in VR. Even though the gender profiling did not perform sufficiently well in AR, here we can observe that, under high workload, both walk (F1-Score = 0.60) and search (F1-Score = 0.58) had a significant influence in detecting the user gender compared to the same actions performed under the low workload. The button interaction was slightly better than the random classifier (F1-score = 0.54). These results align with task-level results, whereby the gender profiling performed better in the NT-high compared to NT-low. Additionally, we observe how the walking action has the largest influence on the accuracy of gender profiling. Again, it might be related to different walking velocities demonstrated by men and women, particularly under high workload [510].

When looking at the actions performed in VR, the pointing action stands out. With an F1-score of 0.82, it strongly contributes to gender profiling compared to all other actions. This might be related both to a singular movement pattern and/or to gender-related eye parameters' variations. Further, results obtained at task-level on a better performance achieved under high compared to low workload are here confirmed only for button interactions. Indeed, the F1-Score at button interactions is about 0.08 higher when users are under high rather than low workload. Again, this reflects results shown in previous studies demonstrating faster operation times in men compared to women specifically when using controller buttons, but not when acting via physical actions [511]. This suggests that profiling users' gender might be easier during tasks involving button interactions, but not in those allowing higher interactivity with the virtual environment.

11.5.3 SENSORS RELEVANCE - ABLATION STUDY

In this section, we conduct an ablation study to understand which sensors contribute the most in our identification, age, and gender predictions. In brief, we trained a Logistic Regression (LR) using only specific subsets of features. In the AR environment, we distinguish between Head Position and Head Rotation features. In VR, we also consider Eyes, Controller Position, and Controller Rotation features. The ablation study was carried out both at Task-Level (Section 11.5.3) and Action-Level (Section 11.5.3).

TASK-LEVEL

Table 11.7 and Table 11.8 show the results of the ablation study for AR and VR tasks, respectively. In the AR environment, Head Rotation features are predominant in the Mental Task for identification and gender prediction. Indeed, in this task, participants were standing still and were instructed not to move their heads; however, it was plausible that their heads oscillated in singular ways, which were detected by our algorithm and leveraged for their identification. In opposition, during the navigation task, Head Position had more impact on all the targets, given that it might be associated with walking patterns. Such a pattern was used in the literature to identify people [561], and could help in Age and Gender prediction as well.

In VR, the identification stage seems to be driven mainly by Eyes features, followed by Controller features. Reasonably, eyes blinking patterns and pupils' dilatation can be person-specific [555] [556] [557], and thus act as a biometric feature. The controllers, instead, were the main interface to interact with the virtual world. Thus, it is reasonable that how a person interacts within the environment helps in the identification. This result aligns with recent findings on video games using mice and keyboards to profile users [6]. Therefore, we could expect AR identification to achieve better performances if such sensors are available, particularly eyes trackers, as reasoned before in Section 11.5.1. In predicting the age, the Controller features yield the best performance. This finding can result from younger people being more familiar with joystick usage. When the workload is high, younger participants may pay more attention to the task rather than how to use the joystick. Moreover, in a low workload scenario, Head and Eyes features contribute similarly. On the other hand, in gender inference, the Head and Eyes features play the most significant role. Indeed, as shown in past literature, gender-based differences exist in how they visually explore a virtual world [562]. Controller features influence the prediction mainly in high workload controller-based tasks.

ACTION-LEVEL

Table 11.9 and Table 11.10 report the results of the ablation study for AR and VR Actions, respectively. In AR, the Head Position has more impact than Head Rotation in predicting our target actions, especially for the walk action. This is reasonable given that such a sensor mainly records the users' walking speed. Head Rotation becomes relevant in the Button Interaction action, in which the participants could only rotate their heads, and is quite helpful to distinguish between genders. As in previous results, the age was difficult to predict. The only case in which we surpass the baseline is in the Walk action with a high workload, but the improvement is too tiny to reason about it.

Looking at VR, we notice that Head Position remains relevant to predict the gender, particularly in scenarios with a low workload. However, most of the time, the Eyes features are the main discriminant to predict our

Table 11.7: Ablation study of sensor importance at task-level in AR. All the measures in F1-Score.

		<i>Identification</i>	<i>Age</i>	<i>Gender</i>
Guessing		0.03	0.5	0.5
Mental Task				
Head Position		0.38	0.46	0.51
Head Rotation		0.54	0.40	0.55
Low W.	Navigation Task			
	Head Position	0.64	0.45	0.56
	Head Rotation	0.46	0.40	0.45
High W.	Navigation Task			
	Head Position	0.65	0.45	0.51
	Head Rotation	0.48	0.44	0.52

Table 11.8: Ablation study of sensor importance at task-level in VR. All the measures in F1-Score.

		<i>Identification</i>	<i>Age</i>	<i>Gender</i>	
Guessing		0.03	0.5	0.5	
Low Workload	Controller Based Task				
	Head Position	0.41	0.68	0.64	
	Head Rotation	0.45	0.76	0.55	
	Eyes	0.83	0.75	0.59	
	Controller Position	0.39	0.69	0.57	
	Controller Rotation	0.59	0.69	0.58	
	Action Based Task				
	Head Position	0.50	0.76	0.62	
	Head Rotation	0.51	0.76	0.60	
	Eyes	0.83	0.74	0.54	
	Controller Position	0.51	0.76	0.58	
	Controller Rotation	0.68	0.81	0.55	
	High Workload	Controller Based Task			
		Head Position	0.48	0.73	0.61
Head Rotation		0.56	0.68	0.57	
Eyes		0.88	0.79	0.69	
Controller Position		0.45	0.78	0.60	
Controller Rotation		0.64	0.68	0.62	
Action Based Task					
Head Position		0.55	0.75	0.53	
Head Rotation		0.55	0.80	0.62	
Eyes		0.89	0.83	0.62	
Controller Position		0.57	0.86	0.50	
Controller Rotation		0.73	0.87	0.50	

targets. In identification, Eyes reached the highest F1-Score in six out of seven actions, suggesting that these features might be the main reason behind the higher identification performances in VR rather than AR. Further, Eyes are predominant in low workload scenarios to predict the users' age. Controller features are pretty helpful in inferring the user's age, especially in high workload actions, while only small differences appear in their usage from people of different genders. Regarding the identification task, the Controller Rotation appears more useful than Controller Position. Last, it is interesting to see how in the idle actions, the Eyes play a significant role, particularly in the high workload scenario, in which we identified a person with 0.81 of F1-Score.

Table 11.9: Ablation study of sensor importance at action-level in AR. All the measures in F1-Score.

		<i>Identification</i>	<i>Age</i>	<i>Gender</i>
Guessing		0.03	0.5	0.5
Low Workload	Search			
	Head Position	0.60	0.40	0.52
	Head Rotation	0.51	0.40	0.58
	Walk			
	Head Position	0.77	0.44	0.60
	Head Rotation	0.55	0.47	0.49
High Workload	Button Interaction			
	Head Position	0.38	0.46	0.52
	Head Rotation	0.56	0.40	0.56
	Search			
	Head Position	0.62	0.40	0.60
	Head Rotation	0.52	0.43	0.57
	Walk			
	Head Position	0.75	0.51	0.53
Head Rotation	0.55	0.43	0.47	

11.6 DISCUSSION

Literature offers some examples of profiling either in AR or VR, only on specific tasks, and through specific features (motion-based [505, 506], eye-tracking-based [507]). Furthermore, to the best of our knowledge, research work testing gender and age profiling in immersive technologies is scarce. In our work, we covered these points by combining all the above-mentioned aspects and performing users’ identification and profiling in two virtual-based scenarios, one involving AR and the other involving VR. The selected datasets present differences and similarities, offering a wide range of exemplary behaviors that can occur when immersed in XR. Indeed, we specifically aimed to propose a general framework that can accurately profile a user across diverse tasks, actions taken, and scenarios. We thus developed a generic pipeline and analyzed differences between profiling algorithms and features across different tasks. Specifically, we demonstrated to what extent users can be profiled during walking, searching for landmarks in the surroundings, pointing to a virtual keyboard for typing, and operating on a virtual robot both via controller-based interaction and physical actions. Remarkably, both virtual environments simulated highly realistic scenarios, and most of these behaviors were executed under high and low workloads, giving good insights into realistic applications of virtual technologies in the field.

The results show that users can be identified and profiled both in AR and VR, with higher VR accuracy. Specifically, in AR, user identification reached good results within the walking action at a low workload, while in VR, the identification algorithm was particularly successful when users performed more physical actions (i.e., pointing, physically interacting with the virtual robot) under a higher workload. As observed from the ablation study, this was mainly due to the additional eye-tracking sensors embedded in the VR but not in the AR headset. Indeed, while in VR the eye features had the most significant impact, the head movements were most influential on the AR users’ identification.

When detecting age, instead, our algorithms were not accurate in AR. This was plausibly related to the low age variability of the tested sample, as the age of participants included in the experiment ranged between 19 and 29. Differently, in VR, we worked on an experimental sample whose age ranged between 23 and 69 years old,

Table 11.10: Ablation study of sensor importance at action-level in VR. All the measures in F1-Score.

		<i>Identification</i>	<i>Age</i>	<i>Gender</i>
Guessing		0.03	0.5	0.5
Low Workload	Idle			
	Head Position	0.41	0.62	0.62
	Head Rotation	0.44	0.69	0.59
	Eyes	0.75	0.80	0.55
	Controller Position	0.38	0.69	0.58
	Controller Rotation	0.55	0.72	0.55
	Pointer			
	Head Position	0.67	0.80	0.57
	Head Rotation	0.73	0.83	0.62
	Eyes	0.91	0.86	0.71
	Controller Position	0.64	0.70	0.59
	Controller Rotation	0.83	0.81	0.51
	Button Interaction			
	Head Position	0.50	0.72	0.63
	Head Rotation	0.55	0.73	0.56
	Eyes	0.85	0.78	0.61
	Controller Position	0.47	0.72	0.58
	Controller Rotation	0.71	0.75	0.60
	Physical Interaction			
	Head Position	0.59	0.75	0.62
Head Rotation	0.56	0.81	0.63	
Eyes	0.87	0.80	0.61	
Controller Position	0.63	0.74	0.57	
Controller Rotation	0.75	0.85	0.56	
High Workload	Idle			
	Head Position	0.49	0.75	0.60
	Head Rotation	0.47	0.76	0.55
	Eyes	0.81	0.77	0.63
	Controller Position	0.46	0.79	0.50
	Controller Rotation	0.65	0.79	0.49
	Button Interaction			
	Head Position	0.57	0.69	0.56
	Head Rotation	0.65	0.66	0.55
	Eyes	0.93	0.83	0.67
	Controller Position	0.50	0.77	0.61
	Controller Rotation	0.73	0.72	0.61
	Physical Interaction			
	Head Position	0.63	0.82	0.54
	Head Rotation	0.63	0.81	0.62
	Eyes	0.71	0.87	0.66
Controller Position	0.66	0.90	0.45	
Controller Rotation	0.79	0.86	0.49	

resulting in detecting the user’s age reasonably accurately. Age detection performed better in most physical actions and interactions than those involving just joysticks and controller buttons, specifically under a higher workload. Interestingly, eye parameters had the most significant impact on age detection in all actions but in the physical interactions, in which the controller position and rotation were more important.

On gender profiling, instead, we observed how the walking activity was again the most prominent in helping detect the user’s gender in AR, with the head position being the most influential sensor. Differently, in VR, our algorithms performed better during the pointing action and under actions at a high workload. In this case, the eye-

related behaviors demonstrated the most considerable influence on gender detection during both these actions. In agreement with AR findings, the head position is quite relevant. Both findings align with the literature on the different eye and head movement behaviors between men and women.

11.6.1 LIMITATIONS

The intrinsic differences between AR and VR devices, and the different nature of the tasks that our participants executed, prevented us from directly comparing the accuracy of profiling between the two technologies. However, such a comparison was out of the scope of our investigation. In this work, we were mainly interested in building a general framework that could serve to profile users during different tasks and across different technologies and scenarios. Therefore, while such differences generated some methodological limitations, in turn, they also highlighted the generability of the proposed framework. Only secondarily, we leveraged some similarities between the tasks executed in AR and VR (i.e., button interaction, and two levels of workload as generated via dual-tasking) to reason about potential profiling differences in similar actions or workloads for these two technologies. However, we want to stress that, even if we adopted a single framework for XR, they remain substantially different. Future studies could focus on analyzing data coming from the same participants engaged in the same activities in both VR and AR, to assess whether differences related to the technical apparatus of XR affect the ease of profiling.

A second limitation was the narrow age range of AR participants, which resulted in poor classification performance. However, we believe such a result is still valuable since it demonstrates that people of similar ages interact similarly with AR devices; therefore, precise age profiling could require significant effort and might not be feasible in every situation.

11.7 CONCLUSION

In conclusion, our work thoroughly studied users' profiling in XR technologies. We proposed a general profiling framework that can potentially infer any private information in any virtual scenario, and could serve as a simple yet powerful baseline for future works. Our results highlight that VR profiling is more straightforward than AR. Through our ablation study, we found eye sensors to be particularly useful in all our predictions (i.e., identification, age, gender), explaining why AR and VR perform differently. Although we are aware of the technical challenges of accurately detecting eye movements in the real world, our findings highlight the importance of incorporating eye-tracking technologies into AR headsets. Our results strongly impact single application areas of XR technologies, such as VR-based industrial robotics and everyday use of wearable AR devices, but also more generally the fast-growing Metaverse. In fact, we pave the way for further researches in XR privacy, proposing a solid inference framework that can be adapted to different virtual technologies and contexts.

In the future, we plan to conduct more experiments on a higher participant pool, which will permit defining finer targets' granularity, including additional private information (e.g., personality traits). We will also focus on which sensors and activities led to higher risks of profiling, and design privacy-preserving techniques while maintaining data utility. Last, we intend to perform a more precise comparison between AR and VR technologies.

12

Conclusion and Future Work

This dissertation advanced Social Network Analytics in three aspects: (i) explaining influence and engagement mechanisms in trending OSNs; (ii) developing resilient SNA tools designed to function effectively in adversarial environments, and (iii) investigating security and privacy concerns in modern social platforms.

The first part of the thesis gave significant insight into how influence and engagement work on Instagram and TikTok, the most trending OSNs among millennials and Gen Z, often overlooked by the research community. The thesis also explored several facets of the recent phenomenon of virtual influencers, always more supported by the most innovating companies and brands, and presented a case study exemplifying information manipulation on Twitter orchestrated by social bots.

Inspired by the presence of adversarial activities in OSNs, the second part of the thesis focused on developing resilient SNA tools tailored for adversarial environments. First, it addressed the increasing occurrence of crowd-turfing on Instagram, creating efficient detectors for such ungenue activities. Then, it centered on identifying obfuscated content that eludes automated analysis. Last, it introduced the innovative concept of social honeypots, i.e., fully automated pages generating compelling content to engage real users, with the ultimate goal of studying their interactions.

The concluding part of this thesis centered on augmenting security and privacy within contemporary social platforms, such as video games and the metaverse. Through PvP, an identification framework tailored for online video games, it is possible to mitigate malicious activities like scams and cyberbullying. Additionally, it demonstrated that publicly available gaming data can be exploited to infer gamers' private information, with no easily adoptable countermeasures. Finally, by proposing a user profiling framework for augmented and virtual reality, it enhanced the security and privacy of these technologies while raising awareness about the potential uses of the data generated within these platforms.

This dissertation opens up numerous avenues for future research. In the scope of Influence and Engagement, there are several promising directions. For example, delving into Instagram stories, which represent a vital tool for influencers, presents an intriguing challenge due to their fleeting nature. On the TikTok front, conducting more

advanced studies, such as trend and challenge analysis, can unveil novel patterns of influence.

In the sphere of SNA tool development, there's room for enhancement in the crowdturfing detection domain, with opportunities to create more scalable algorithms. Concerning obfuscated content, some challenges identified in the proposed taxonomy persist, and the ever-evolving landscape may give rise to new obfuscation techniques. Further advancements in social honeypots can be explored, encompassing innovative engagement and content generation strategies to increase their appeal.

Within the domains of video games and the metaverse, many paths are yet to be investigated, given the novelty of these subjects. Potential areas of exploration include the deployment of profiling frameworks that function seamlessly across various games or platforms. Additionally, there's a pressing need to design more effective countermeasures for safeguarding user privacy while preserving data utility.

References

- [1] “Dendi - liquipedia dota2 wiki,” <https://liquipedia.net/dota2/Dendi>.
- [2] G. Ye, Z. Tang, D. Fang, Z. Zhu, Y. Feng, P. Xu, X. Chen, and Z. Wang, “Yet another text captcha solver: A generative adversarial network based approach,” in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 332–348.
- [3] I. T. B. Esports, “r/dota2 demographic survey,” <https://www.docdroid.net/ZeJTLar/rdota2-demographics-report-2021-pdf>, 2021, accessed: June, 2022.
- [4] A. Player, “Live player count and statistics,” <https://activeplayer.io/>, 2022.
- [5] Steam, “An ongoing analysis of steam’s concurrent players,” <https://steamcharts.com/>, 2022, accessed: July, 2022.
- [6] M. Conti and P. P. Tricomi, “PvP: Profiling Versus Player! Exploiting Gaming Data for Player Recognition,” in *Int. Conf. Inf. Secur.*, 2020.
- [7] howlongis.io, “Dota 2 playtime,” <https://howlongis.io/app/570/Dota+2>, 2022.
- [8] E. Earnings, “Top games awarding prize money.” <https://www.esportsearnings.com/games>, 2022, accessed: July, 2022.
- [9] P. Bedi and C. Sharma, “Community detection in social networks,” *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, vol. 6, no. 3, pp. 115–135, 2016.
- [10] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, “A survey of sentiment analysis in social media,” *Knowledge and Information Systems*, vol. 60, pp. 617–663, 2019.
- [11] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, “Information diffusion in online social networks: A survey,” *ACM Sigmod Record*, vol. 42, no. 2, pp. 17–28, 2013.
- [12] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, “Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network,” *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.
- [13] F. de Oliveira Santini, W. J. Ladeira, D. C. Pinto, M. M. Herter, C. H. Sampaio, and B. J. Babin, “Customer engagement in social media: a framework and meta-analysis,” *Journal of the Academy of Marketing Science*, vol. 48, pp. 1211–1228, 2020.
- [14] B. E. Weeks, A. Ardèvol-Abreu, and H. Gil de Zúñiga, “Online influence? social media use, opinion leadership, and political persuasion,” *International journal of public opinion research*, vol. 29, no. 2, pp. 214–239, 2017.

- [15] Insiderintelligence, “Instagram will net more us ad revenues than core facebook platform,” 2021, accessed: August 2022. [Online]. Available: <https://www.insiderintelligence.com/content/instagram-will-net-more-us-ad-revenues-than-core-facebook-platform>
- [16] A. E. Marwick and R. Lewis, *Media manipulation and disinformation online*. Data & Society Research Institute, 2017.
- [17] Y. Piao, K. Ye, and X. Cui, “Privacy inference attack against users in online social networks: a literature review,” *IEEE Access*, vol. 9, pp. 40 417–40 431, 2021.
- [18] N. Z. Gong and B. Liu, “You are who you know and how you behave: Attribute inference attacks via users’ social friends and behaviors,” in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 979–995.
- [19] M. Conti, J. Gathani, and P. P. Tricomi, “Virtual influencers in online social media,” *IEEE Communications Magazine*, vol. 60, no. 8, pp. 86–91, 2022.
- [20] P. P. Tricomi, M. Chilese, M. Conti, and A.-R. Sadeghi, “Follow us and become famous! insights and guidelines from instagram engagement mechanisms,” in *Proceedings of the 15th ACM Web Science Conference 2023*, 2023, pp. 346–356.
- [21] F. L. De Faveri, L. Cosuti, P. P. Tricomi, and M. Conti, “Twitter bots influence on the russo-ukrainian war during the 2022 italian general elections,” in *Security and Privacy in Social Networks and Big Data*. Singapore: Springer Nature Singapore, 2023, pp. 38–57.
- [22] P. P. Tricomi, S. Tarahomi, C. Cattai, F. Martini, and M. Conti, “Are we all in a truman show? spotting instagram crowdturfing through self-training,” in *2023 32nd International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 2023, pp. 1–10.
- [23] M. Conti, L. Pajola, and P. P. Tricomi, “Turning captchas against humanity: Captcha-based attacks in online social media,” *Online Social Networks and Media*, vol. 36, p. 100252, 2023.
- [24] S. Bardi, M. Conti, L. Pajola, and P. P. Tricomi, “Social honeypot for humans: Luring people through self-managed instagram pages,” in *Applied Cryptography and Network Security: 21st International Conference, ACNS 2023, Kyoto, Japan, June 19–22, 2023, Proceedings, Part I*. Springer, 2023, pp. 309–336.
- [25] P. P. Tricomi, L. Facciolo, G. Apruzzese, and M. Conti, “Attribute inference attacks in online multiplayer video games: A case study on dota2,” in *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, 2023, pp. 27–38.
- [26] P. P. Tricomi, F. Nenna, L. Pajola, M. Conti, and L. Gamberi, “You can’t hide behind your headset: User profiling in augmented and virtual reality,” *IEEE Access*, vol. 11, pp. 9859–9875, 2023.
- [27] M. Cardaioli, M. Conti, G. Orazi, P. P. Tricomi, and G. Tsudik, “Blufader: Blurred face detection & recognition for privacy-friendly continuous authentication,” *Pervasive and Mobile Computing*, vol. 92, p. 101801, 2023.
- [28] S. Mondini, V. Pucci, M. Pastore, O. Gaggi, P. P. Tricomi, and M. Nucci, “s-criq: the online short version of the cognitive reserve index questionnaire,” *Aging Clinical and Experimental Research*, pp. 1–8, 2023.

- [29] M. Conti, P. Vinod, and P. P. Tricomi, "Secure static content delivery for cdn using blockchain technology," in *International Workshop on Data Privacy Management*. Springer, 2021, pp. 301–309.
- [30] M. Cardaioli, M. Conti, P. P. Tricomi, and G. Tsudik, "Privacy-friendly de-authentication with blufade: Blurred face detection," in *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2022, pp. 197–206.
- [31] E. Moustakas, N. Lamba, D. Mahmoud, and C. Ranganathan, "Blurring lines between fiction and reality: Perspectives of experts on marketing effectiveness of virtual influencers," in *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*. IEEE, 2020, pp. 1–6.
- [32] N. Baklanov, "The top instagram virtual influencers in 2020," <https://hypeauditor.com/blog/the-top-instagram-virtual-influencers-in-2020>, 2020, accessed: March 12, 2021.
- [33] L. Dodgson, "The who has recruited a cgi influencer to get young people interested in safe practices around the coronavirus," <https://www.insider.com/who-using-cgi-influencer-to-spread-safe-coronavirus-practices-2020-4>, 2020, accessed: March 20, 2021.
- [34] D. Black, "The virtual idol: Producing and consuming digital femininity," *Idols and celebrity in Japanese media culture*, pp. 209–28, 2012.
- [35] —, "The virtual ideal: Virtual idols, cute technology and unclean biology," *Continuum*, vol. 22, no. 1, pp. 37–50, 2008.
- [36] J. Guga, "Virtual idol hatsune miku: New auratic experience of the performer as a collaborative platform," in *Arts and Technology: Fourth International Conference, ArtsIT 2014, Istanbul, Turkey, November 10-12, 2014, Revised Selected Papers 4*. Springer, 2015, pp. 36–44.
- [37] M. T. Tang, V. L. Zhu, and V. Popescu, "Alterecho: Loose avatar-streamer coupling for expressive vtubing," in *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2021, pp. 128–137.
- [38] I. M. Hub, "The state of influencer marketing 2021: Benchmark report," <https://influencermarketinghub.com/influencer-marketing-benchmark-report-2021/>, 2021, accessed: April 05, 2021.
- [39] J. Drenten and G. Brooks, "Celebrity 2.0: Lil miquela and the rise of a virtual star system," *Feminist Media Studies*, vol. 20, no. 8, pp. 1319–1323, 2020.
- [40] B. Robinson *et al.*, "Towards an ontology and ethics of virtual influencers," *Australasian Journal of Information Systems*, vol. 24, 2020.
- [41] V. Molin and S. Nordgren, "Robot or human? the marketing phenomenon of virtual influencers: A case study about virtual influencers' parasocial interaction on instagram," 2019.
- [42] D. Fowler, "The fascinating world of instagram's 'virtual' celebrities," <https://www.bbc.com/worklife/article/20180402-the-fascinating-world-of-instagrams-virtual-celebrities>, 2018, accessed: April 27, 2021.
- [43] Fabernovel, "In ten years, who'll know what's virtual and what's real? what matters will be the quality of the content," <https://www.fabernovel.com/en/article/cultures/interview-hirokuni-genie-miyaji-male-influencer-japan>, 2020, accessed: April 29, 2021.

- [44] C. Travers, “5 notable virtual influencer brand partnerships,” <https://www.virtualhumans.org/article/5-notable-virtual-influencer-brand-partnerships>, 2020, accessed: April 16, 2020.
- [45] T. B. of Fashion, “5 notable virtual influencer brand partnerships,” <https://www.businessoffashion.com/articles/media/meeting-fashions-first-computer-generated-influencer-lil-miquela-sousa>, 2018, accessed: May 03, 2021.
- [46] R. Ransaw, “The psychology behind why we share on social media,” 2021, accessed: August 2022. [Online]. Available: <https://www.shutterstock.com/blog/the-psychology-behind-why-we-share-on-social-media>
- [47] R. Martinez-Pecino and M. Garcia-Gavilán, “Likes and problematic instagram use: the moderating role of self-esteem,” *Cyberpsychology, Behavior, and Social Networking*, vol. 22, no. 6, pp. 412–416, 2019.
- [48] E. Maslowska, S. J. Kim, E. C. Malthouse, and V. Viswanathan, “Online reviews as customers’ dialogues with and about brands,” in *Handbook of research on customer engagement*. Cheltenham, UK: Edward Elgar Publishing, 2019.
- [49] Insiderintelligence, “Instagram in 2022: Global user statistics, demographics and marketing trends to know,” 2022, accessed: August 2022. [Online]. Available: <https://www.insiderintelligence.com/insights/instagram-user-statistics-trends/>
- [50] K. Zarei, R. Farahbakhsh, and N. Crespi, “How impersonators exploit instagram to generate fake engagement?” in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. Online Conference: IEEE, 2020, pp. 1–6.
- [51] Statusbrew, “Instagram algorithm 2022: How to conquer it,” 2021, accessed: August 2022. [Online]. Available: <https://statusbrew.com/insights/instagram-algorithm/>
- [52] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
- [53] M. S. Hee, R. K.-W. Lee, and W.-H. Chong, “On explaining multimodal hateful meme detection models,” in *Proceedings of the ACM Web Conference 2022*. Lyon, France: ACM, 2022, pp. 3651–3655.
- [54] C. Molnar, *Interpretable machine learning*. Online Publisher: Lulu.com, 2020.
- [55] M. Mazloom, R. Rietveld, S. Rudinac, M. Worring, and W. Van Dolen, “Multimodal popularity prediction of brand-related social media posts,” in *Proceedings of the 24th ACM international conference on Multimedia*. Amsterdam, Netherlands: ACM, 2016, pp. 197–201.
- [56] M. Mazloom, I. Pappi, and M. Worring, “Category specific post popularity prediction,” in *International Conference on Multimedia Modeling*. Bangkok, Thailand: Springer, 2018, pp. 594–607.
- [57] S. De, A. Maity, V. Goel, S. Shitole, and A. Bhattacharya, “Predicting the popularity of instagram posts for a lifestyle magazine using deep learning,” in *2017 2nd international conference on communication systems, computing and IT applications (CSCITA)*. Mumbai, India: IEEE, 2017, pp. 174–177.
- [58] A. Zohourian, H. Sajedi, and A. Yavary, “Popularity prediction of images and videos on instagram,” in *2018 4th International Conference on Web Research (ICWR)*. Iran: IEEE, 2018, pp. 1111–1117.

- [59] Z. Zhang, T. Chen, Z. Zhou, J. Li, and J. Luo, “How to become instagram famous: Post popularity prediction with dual-attention,” in *2018 IEEE international conference on big data (big data)*. Seattle, USA: IEEE, 2018, pp. 2383–2392.
- [60] K. Ding, K. Ma, and S. Wang, “Intrinsic image popularity assessment,” in *Proceedings of the 27th ACM International Conference on Multimedia*. Nice, France: ACM, 2019, pp. 1979–1987.
- [61] M. Gayberi and S. G. Oguducu, “Popularity prediction of posts in social networks based on user, post and image features,” in *Proceedings of the 11th International Conference on Management of Digital EcoSystems*. Limassol, Cyprus: ACM, 2019, pp. 9–15.
- [62] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Zurich, Switzerland: Springer, 2014, pp. 740–755.
- [63] C. Riis, D. K. Kowalczyk, and L. K. Hansen, “On the limits to multi-modal popularity prediction on instagram—a new robust, efficient and explainable baseline,” *arXiv preprint arXiv:2004.12482*, vol. 2004, no. 12482, pp. 1–10, 2020.
- [64] S. Carta, A. S. Podda, D. R. Recupero, R. Saia, and G. Usai, “Popularity prediction of instagram posts,” *Information*, vol. 11, no. 9, p. 453, 2020.
- [65] K. R. Purba, D. Asirvatham, and R. K. Murugesan, “Instagram post popularity trend analysis and prediction using hashtag, image assessment, and user history features.” *Int. Arab J. Inf. Technol.*, vol. 18, no. 1, pp. 85–94, 2021.
- [66] K. K. Aldous, J. An, and B. J. Jansen, “View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13. München, Germany: AAAI, 2019, pp. 47–57.
- [67] W. Geysler, “The state of influencer marketing 2021: Benchmark report,” 2021, accessed: August 2022. [Online]. Available: <https://influencermarketinghub.com/influencer-marketing-benchmark-report-2021/>
- [68] S. Sharma, F. Alam, M. Akhtar, D. Dimitrov, G. D. S. Martino, H. Firooz, A. Halevy, F. Silvestri, P. Nakov, T. Chakraborty *et al.*, “Detecting and understanding harmful memes: A survey,” *arXiv preprint arXiv:2205.04274*, vol. 2205, no. 04274, pp. 1–9, 2022.
- [69] H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven, and R. van Lier, *Explainable and interpretable models in computer vision and machine learning*. Cham, Switzerland: Springer, 2018.
- [70] S. Kim, J.-Y. Jiang, M. Nakada, J. Han, and W. Wang, “Multimodal post attentive profiling for influencer marketing,” in *Proceedings of The Web Conference 2020*. Taipei: ACM, 2020, pp. 2878–2884.
- [71] C. Newberry, “12 foolproof instagram growth strategies for 2022,” 2022, accessed: August 2022. [Online]. Available: <https://blog.hootsuite.com/instagram-growth/>

- [72] HypeAuditor, “State of influencer marketing 2022,” 2022, accessed: August 2022. [Online]. Available: <https://hypeauditor.com/blog/wp-content/uploads/2022/01/US-State-of-Influencer-Marketing-2022.pdf>
- [73] D. Maposa, E. Mudimu, and O. Ngwenya, “A multivariate analysis of variance (manova) of the performance of sorghum lines in different agro-ecological regions of zimbabwe,” *African Journal of Agricultural Research*, vol. 5, pp. 196–203, 02 2010.
- [74] B. G. Tabachnick, L. S. Fidell, and J. B. Ullman, *Using multivariate statistics*. Boston, MA: pearson, 2007, vol. 5.
- [75] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.
- [76] A. Paszke, S. Gross, Massa *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds. New York, USA: Curran Associates, Inc., 2019, pp. 8024–8035.
- [77] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-shot multi-level face localisation in the wild,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Virtual Conference: IEEE/CVF, 2020, pp. 5203–5212.
- [78] mowshon, “Age and gender,” 2022. [Online]. Available: <https://github.com/mowshon/age-and-gender>
- [79] C. Sherman and P. Quester, “The influence of product/nudity congruence on advertising effectiveness,” *Journal of Promotion Management*, vol. 11, no. 2-3, pp. 61–89, 2006.
- [80] notAI.tech, “Nudenet: Neural nets for nudity classification, detection and selective censoring,” 2022, accessed: August 2022. [Online]. Available: <https://github.com/notAI-tech/NudeNet>
- [81] S. Chandra Guntuku, D. Preotiuc-Pietro, J. C. Eichstaedt, and L. H. Ungar, “What twitter profile and posted images reveal about depression and anxiety,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, no. 01, pp. 236–246, Jul. 2019.
- [82] A. Mehrabian and J. A. Russell, *An approach to environmental psychology*. Cambridge, MA: the MIT Press, 1974.
- [83] P. Valdez and A. Mehrabian, “Effects of color on emotions.” *Journal of experimental psychology: General*, vol. 123, no. 4, p. 394, 1994.
- [84] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, “Photo aesthetics ranking network with attributes and content adaptation,” in *European conference on computer vision*. Amsterdam, Netherlands: Springer, 2016, pp. 662–679.
- [85] V. Campos, B. Jou, and X. Giro-i Nieto, “From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction,” *Image and Vision Computing*, vol. 65, pp. 15–22, 2017.

- [86] J. Kangasharju, “Cuteness detector,” <https://github.com/asharov/cute-animal-detector>, 2022.
- [87] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič, “Sentiment of emojis,” *PLOS ONE*, vol. 10, no. 12, pp. 1–22, 12 2015.
- [88] Google, “Google cloud platform,” 2022. [Online]. Available: <https://cloud.google.com/>
- [89] D. Zwillinger and S. Kokoska, *CRC standard probability and statistics tables and formulae*. USA: Crc Press, 1999.
- [90] J. Myers, A. Well, and R. Lorch, *Research Design and Statistical Analysis*, ser. Online access with subscription: Proquest Ebook Central. London, UK: Routledge, 2010.
- [91] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [92] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, “Effective prediction of three common diseases by combining smote with tokek links technique for imbalanced medical data,” in *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*. Chongqing, China: IEEE, 2016, pp. 225–228.
- [93] G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: <http://jmlr.org/papers/v18/16-365.html>
- [94] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, Nevada: IEEE, 2016, pp. 770–778.
- [95] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3982–3992.
- [96] F. Zarrinkalam, S. Faralli, G. Piao, E. Bagheri *et al.*, “Extracting, mining and predicting users’ interests from social media,” *Foundations and Trends® in Information Retrieval*, vol. 14, no. 5, pp. 445–617, 2020.
- [97] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International conference on machine learning*. Beijing, China: PMLR, 2014, pp. 1188–1196.
- [98] H. Face, 2022, accessed: August 2022. [Online]. Available: <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>
- [99] C. Niu, H. Shan, and G. Wang, “Spice: Semantic pseudo-labeling for image clustering,” *IEEE Transactions on Image Processing*, vol. 31, pp. 7264–7278, 2022.
- [100] L. Zhang, P. Jones, K. A. Otis, J. Gale, and E. Chan, “Trending topic extraction from social media,” Oct. 9 2018, uS Patent 10,095,686.
- [101] F. Pedregosa, G. Varoquaux *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [102] E. H. Simpson, “Measurement of diversity,” *nature*, vol. 163, no. 4148, pp. 688–688, 1949.

- [103] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 137–146.
- [104] S. Naseer, S. Hasan, J. Bhuiyan, and A. Prasad, “Current public trends in the discussion of dry eyes: a cross-sectional analysis of popular content on tiktok,” *Cureus*, vol. 14, no. 2, 2022.
- [105] M. Haenlein, E. Anadol, T. Farnsworth, H. Hugo, J. Hunichen, and D. Welte, “Navigating the new era of influencer marketing: How to be successful on instagram, tiktok, & co.” *California management review*, vol. 63, no. 1, pp. 5–25, 2020.
- [106] Amal Moursi, “The most from tiktok sponsored posts?” <https://www.hopperhq.com/blog/2022-tiktok-rich-list/>, 2022, accessed: 2023-09-08.
- [107] Y. Jiang, X. Jin, and Q. Deng, “Short video uprising: how# blacklivesmatter content on tiktok challenges the protest paradigm,” *arXiv preprint arXiv:2206.09946*, 2022.
- [108] Z. Tian, R. Dew, and R. Iyengar, “Mega or micro? influencer selection using follower elasticity,” *Influencer Selection Using Follower Elasticity (June 14, 2023)*, 2023.
- [109] T. Paksoy, S. C. Şen, G. Ustaoglu, and D. G. Bulut, “What do tiktok videos offer us about dental implants treatment?” *Journal of Stomatology, Oral and Maxillofacial Surgery*, vol. 124, no. 1, p. 101320, 2023.
- [110] S. I. Serengil and A. Ozpinar, “Hyperextended lightface: A facial attribute analysis framework,” in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2021, pp. 1–4. [Online]. Available: <https://doi.org/10.1109/ICEET53442.2021.9659697>
- [111] J. Feldkamp, “The rise of tiktok: The evolution of a social media platform during covid-19,” *Digital responses to Covid-19: Digital innovation, transformation, and entrepreneurship during pandemic outbreaks*, pp. 73–85, 2021.
- [112] Y. Li, M. Guan, P. Hammond, and L. E. Berrey, “Communicating COVID-19 information on TikTok: a content analysis of TikTok videos from official accounts featured in the COVID-19 information hub,” *Health Education Research*, vol. 36, no. 3, pp. 261–271, 2021. [Online]. Available: <https://doi.org/10.1093/her/cyab010>
- [113] A. M. Ostrovsky and J. R. Chen, “Tiktok and its role in covid-19 information propagation,” *Journal of adolescent health*, vol. 67, no. 5, p. 730, 2020.
- [114] L. Southwick, S. C. Guntuku, E. V. Klinger, E. Seltzer, H. J. McCalpin, and R. M. Merchant, “Characterizing covid-19 content posted to tiktok: public sentiment and response during the first phase of the covid-19 pandemic,” *Journal of Adolescent Health*, vol. 69, no. 2, pp. 234–241, 2021.
- [115] C. Ling, J. Blackburn, E. De Cristofaro, and G. Stringhini, “Slapping cats, bopping heads, and oreo shakes: Understanding indicators of virality in tiktok short videos,” in *Proceedings of the 14th ACM Web Science Conference 2022*, 2022, pp. 164–173.

- [116] D. Klug, Y. Qin, M. Evans, and G. Kaufman, "Trick and please. a mixed-method study on user assumptions about the tiktok algorithm," in *Proceedings of the 13th ACM Web Science Conference 2021*, ser. WebSci '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 84–92. [Online]. Available: <https://doi.org/10.1145/3447535.3462512>
- [117] A. Fiallos, C. Fiallos, and S. Figueroa, "Tiktok and education: Discovering knowledge through learning videos," in *2021 Eighth International Conference on EDemocracy & EGovernment (ICEDEG)*. IEEE, 2021, pp. 172–176.
- [118] Z. N. Khlaif and S. Salha, "Using tiktok in education: a form of micro-learning or nano-learning?" *Interdisciplinary Journal of Virtual Learning in Medical Sciences*, vol. 12, no. 3, pp. 213–218, 2021.
- [119] D. Vijay and A. Gekker, "Playing politics: How sabarimala played out on tiktok," *American behavioral scientist*, vol. 65, no. 5, pp. 712–734, 2021.
- [120] A. Neyaz, A. Kumar, S. Krishnan, J. Placker, and Q. Liu, "Security, privacy and steganographic analysis of faceapp and tiktok," *International journal of computer science and security*, vol. 14, no. 2, pp. 38–59, 2020.
- [121] K. E. Anderson, "Getting acquainted with social networks and apps: it is time to talk about tiktok," *Library hi tech news*, vol. 37, no. 4, pp. 7–12, 2020.
- [122] G. Weimann and N. Masri, "Research note: Spreading hate on tiktok," *Studies in conflict & terrorism*, vol. 46, no. 5, pp. 752–765, 2023.
- [123] S. Banerjee, M. Jenamani, and D. K. Pratihar, "A survey on influence maximization in a social network," *Knowledge and Information Systems*, vol. 62, pp. 3417–3455, 2020.
- [124] E. Lahuerta-Otero and R. Cordero-Gutiérrez, "Looking for the perfect tweet. the use of data mining techniques to find influencers on twitter," *Computers in Human Behavior*, vol. 64, pp. 575–583, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0747563216305258>
- [125] A. Arora, S. Bansal, C. Kandpal, R. Aswani, and Y. Dwivedi, "Measuring social media influencer index-insights from facebook, twitter and instagram," *Journal of retailing and consumer services*, vol. 49, pp. 86–101, 2019.
- [126] C. Hughes, V. Swaminathan, and G. Brooks, "Driving brand engagement through online social influencers: An empirical investigation of sponsored blogging campaigns," *Journal of marketing*, vol. 83, no. 5, pp. 78–96, 2019.
- [127] K. Sokolova and H. Kefi, "Instagram and youtube bloggers promote it, why should i buy? how credibility and parasocial interaction influence purchase intentions," *Journal of retailing and consumer services*, vol. 53, p. 101742, 2020.
- [128] W. Tafesse and B. P. Wood, "Followers' engagement with instagram influencers: The role of influencers' content and engagement strategy," *Journal of retailing and consumer services*, vol. 58, p. 102303, 2021.
- [129] L. V. Casalo, C. Flavián, and S. Ibáñez-Sánchez, "Be creative, my friend! engaging users on instagram by promoting positive emotions," *Journal of Business Research*, vol. 130, pp. 416–425, 2021.

- [130] S. Barta, D. Belanche, A. Fernández, and M. Flavián, “Influencer marketing on tiktok: The effectiveness of humor and followers’ hedonic experience,” *Journal of Retailing and Consumer Services*, vol. 70, p. 103149, 2023.
- [131] Y. Yang and L. Ha, “Why people use tiktok (douyin) and how their purchase intentions are affected by social media influencers in china: A uses and gratifications and parasocial relationship perspective,” *Journal of Interactive Advertising*, vol. 21, no. 3, pp. 297–305, 2021.
- [132] J. Yang, J. Zhang, and Y. Zhang, “Influencer video advertising in tiktok,” *Mit Initiative On The Digital Economy*, vol. 4, 2021.
- [133] J. A. Russell and A. Mehrabian, “Evidence for a three-factor theory of emotions,” *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.
- [134] speechbrain, “Emotion recognition with wav2vec2 base on iemocap,” <https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP>, 2021, accessed: 2023-07-10.
- [135] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, “SpeechBrain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.
- [136] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, “Dawn of the transformer era in speech emotion recognition: Closing the valence gap,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.
- [137] L. Shang, Z. Kou, Y. Zhang, and D. Wang, “A multimodal misinformation detector for covid-19 short videos on tiktok,” in *2021 IEEE international conference on big data (big data)*. IEEE, 2021, pp. 899–908.
- [138] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-shot multi-level face localisation in the wild,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.
- [139] P. Ekman and W. V. Friesen, “Facial action coding system,” *Environmental Psychology & Nonverbal Behavior*, 1978.
- [140] J. H. Cheong, E. Jolly, T. Xie, S. Byrne, M. Kenney, and L. J. Chang, “Py-feat: Python facial expression analysis toolbox,” *Affective Science*, pp. 1–16, 2023.
- [141] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, “The menlo report,” *IEEE Security & Privacy*, vol. 10, no. 2, pp. 71–75, 2012.
- [142] C. Fiesler, N. Beard, and B. C. Keegan, “No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service,” in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 187–196.
- [143] G. U. della Repubblica Italiana, “Decreto legge 28 febbraio 2022,” <https://www.gazzettaufficiale.it/eli/gu/2022/02/28/49/sg/pdf>, 2022.

- [144] M. Ludovico, “Cybersecurity, 2022 annus horribilis: 13mila attacchi, +138%,” *Il Sole 24 Ore*, 2023.
- [145] K. T. Gaubatz, *Elections and war: the electoral incentive in the democratic politics of war and peace*. Stanford University Press, 1999.
- [146] F. Caravaca, J. González-Cabañas, Á. Cuevas, and R. Cuevas, “Estimating ideology and polarization in european countries using facebook data,” *EPJ Data Science*, vol. 11, no. 1, p. 56, 2022.
- [147] S. C. Woolley, “Automating power: Social bot interference in global politics,” *First Monday*, 2016.
- [148] V. Vasilkova and N. Legostaeva, “Social bots in political communication,” *RUDN Journal of Sociology*, vol. 19, no. 1, pp. 121–133, 2019.
- [149] H. S. Dutta and T. Chakraborty, “Blackmarket-driven collusion among retweeters—analysis, detection, and characterization,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1935–1944, 2019.
- [150] F. Schäfer, S. Evert, and P. Heinrich, “Japan’s 2014 general election: Political bots, right-wing internet activism, and prime minister shinzō abe’s hidden nationalist agenda,” *Big data*, vol. 5, no. 4, pp. 294–309, 2017.
- [151] D. L. Linvill, B. C. Boatwright, W. J. Grant, and P. L. Warren, ““the russians are hacking my brain!” investigating russia’s internet research agency twitter tactics during the 2016 united states presidential campaign,” *Computers in Human Behavior*, vol. 99, pp. 292–300, 2019.
- [152] H.-C. H. Chang, E. Chen, M. Zhang, G. Muric, and E. Ferrara, “Social bots and social media manipulation in 2020: the year in review,” *arXiv preprint arXiv:2102.08436*, 2021.
- [153] K.-C. Yang, E. Ferrara, and F. Menczer, “Botometer 101: Social bot practicum for computational social scientists,” *Journal of Computational Social Science*, pp. 1–18, 2022.
- [154] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, “A survey of twitter research: Data model, graph structure, sentiment analysis and attacks,” *Expert Systems with Applications*, vol. 164, p. 114006, 2021.
- [155] M. Arias, A. Arratia, and R. Xuriguera, “Forecasting with twitter data,” *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 1, jan 2014. [Online]. Available: <https://doi.org/10.1145/2542182.2542190>
- [156] A. Khan, H. Zhang, N. Boudjellal, A. Ahmad, J. Shang, L. Dai, and B. Hayat, “Election prediction on twitter: a systematic mapping study,” *Complexity*, vol. 2021, pp. 1–27, 2021.
- [157] D. P. Giakatos, P. Sermpezis, and A. Vakali, “Pypoll: A python library automating mining of networks, discussions and polarization on twitter,” *arXiv preprint arXiv:2303.06478*, 2023.
- [158] I. Weber, V. R. K. Garimella, and A. Batayneh, “Secular vs. islamist polarization in egypt on twitter,” in *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, 2013, pp. 290–297.
- [159] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo, “People are strange when you’re a stranger: Impact and influence of bots on social networks,” in *International conference on web and social media (ICWSM)*, vol. 6, 2012, pp. 10–17.

- [160] O. Varol, “Should we agree to disagree about twitter’s bot problem?” *arXiv preprint arXiv:2209.10006*, 2022.
- [161] E. Alothali, N. Zaki, E. A. Mohamed, and H. Alashwal, “Detecting social bots on twitter: a literature review,” in *2018 International conference on innovations in information technology (IIT)*. IEEE, 2018, pp. 175–180.
- [162] N. Chavoshi, H. Hamooni, and A. Mueen, “Identifying correlated bots in twitter,” in *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part II 8*. Springer, 2016, pp. 14–21.
- [163] Z. Gilani, R. Farahbakhsh, G. Tyson, L. Wang, and J. Crowcroft, “Of bots and humans (on twitter),” in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 349–354.
- [164] L. Mannocci, S. Cresci, A. Monreale, A. Vakali, and M. Tesconi, “Mulbot: Unsupervised bot detection based on multivariate time series,” *arXiv preprint arXiv:2209.10361*, 2022.
- [165] S. Cresci, “A decade of social bot detection,” *Communications of the ACM*, vol. 63, no. 10, pp. 72–83, 2020.
- [166] B. Insider, “Percentage of Bots on Twitter,” <https://www.businessinsider.com/twitter-bots-comprise-less-than-5-but-tweet-more-2022-9>.
- [167] T. N. Y. Rob Dubbin, “Percentage of Bots on the early stages of Twitter,” <https://www.newyorker.com/tech/annals-of-technology/the-rise-of-twitter-bots>.
- [168] Z. Weng and A. Lin, “Public opinion manipulation on social media: Social network analysis of twitter bots during the covid-19 pandemic,” *International Journal of Environmental Research and Public Health*, vol. 19, no. 24, p. 16376, 2022.
- [169] M. Mazza, M. Avvenuti, S. Cresci, and M. Tesconi, “Investigating the difference between trolls, social bots, and humans on twitter,” *Computer Communications*, vol. 196, pp. 23–36, 2022.
- [170] I. Alsmadi and M. J. O’Brien, “How many bots in russian troll tweets?” *Information Processing & Management*, vol. 57, no. 6, p. 102303, 2020.
- [171] S. Cresci, F. Lillo, D. Regoli, S. Tardelli, and M. Tesconi, “Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter,” *ACM Transactions on the Web (TWEB)*, vol. 13, no. 2, pp. 1–27, 2019.
- [172] M. Singh, D. Bansal, and S. Sofat, “Behavioral analysis and classification of spammers distributing pornographic content in social media,” *Social Network Analysis and Mining*, vol. 6, pp. 1–18, 2016.
- [173] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze, “Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate,” *American journal of public health*, vol. 108, no. 10, pp. 1378–1384, 2018.

- [174] J. Pastor-Galindo, M. Zago, P. Nespoli, S. L. Bernal, A. H. Celdrán, M. G. Pérez, J. A. Ruipérez-Valiente, G. M. Pérez, and F. G. Mármol, “Spotting political social bots in twitter: A use case of the 2019 spanish general election,” *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2156–2170, 2020.
- [175] “Social feed manager,” 2016.
- [176] J. Fernquist, L. Kaati, and R. Schroeder, “Political bots and the swedish general election,” in *2018 IEEE international conference on intelligence and security informatics (isi)*. IEEE, 2018, pp. 124–129.
- [177] A. Bessi and E. Ferrara, “Social bots distort the 2016 us presidential election online discussion,” *First Monday*, vol. 21, no. 11-7, 2016.
- [178] E. Ferrara, H. Chang, E. Chen, G. Muric, and J. Patel, “Characterizing social media manipulation in the 2020 us presidential election,” *First Monday*, 2020.
- [179] E. Ferrara, “Bots, elections, and social media: a brief overview,” *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pp. 95–114, 2020.
- [180] B. James Reynolds, “Italy pm conte vows more united italy as salvini leaves power,” [ItalyPMContevows moreunitedItalyasSalvinileavespower](#).
- [181] H. Messia and C. Angela Dewan, “Italian prime minister giuseppe conte resigns, in calculated move amid coronavirus crisis,” 2021. [Online]. Available: <https://edition.cnn.com/2021/01/26/europe/italy-giuseppe-conte-resignation-intl/index.html>
- [182] T. Developer, “Twitter api platform,” <https://developer.twitter.com/en/docs>.
- [183] J. Filter, “clean-text,” 2022. [Online]. Available: <https://github.com/jfilter/clean-text>
- [184] G. Diaz, “stopwords-it,” 2022. [Online]. Available: <https://github.com/stopwords-iso/stopwords-it>
- [185] L. Oesper, D. Merico, R. Isserlin, and G. D. Bader, “Wordcloud: a cytoscape plugin to create a visual semantic summary of networks,” *Source code for biology and medicine*, vol. 6, no. 1, p. 7, 2011.
- [186] N. Reimers and I. Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. [Online]. Available: <https://arxiv.org/abs/2004.09813>
- [187] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.
- [188] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering.” *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017.
- [189] M. Grootendorst, “Bertopic: Neural topic modeling with a class-based tf-idf procedure,” *arXiv preprint arXiv:2203.05794*, 2022.
- [190] Wikipedia, “Timeline of the 2022 russian invasion of ukraine,” https://en.wikipedia.org/wiki/Timeline_of_the_2022_Russian_invasion_of_Ukraine, 2023.

- [191] I. for the study of war, “Ukraine conflict updates 2022,” <https://www.understandingwar.org/background/ukraine-conflict-updates-2022>, 2022.
- [192] F. Stefanoni, “Un anno di guerra in ucraina, la risposta della politica italiana: le posizioni (e le evoluzioni) dei partiti,” *Corriere della Sera*, 2023.
- [193] F. Martini, P. Samula, T. R. Keller, and U. Klinger, “Bot, or not? comparing three methods for detecting social bots in five political discourses,” *Big data & society*, vol. 8, no. 2, p. 20539517211033566, 2021.
- [194] M. Mattei, G. Caldarelli, T. Squartini, and F. Saracco, “Italian twitter semantic network during the covid-19 epidemic,” *EPJ Data Science*, vol. 10, no. 1, p. 47, 2021.
- [195] A. Shevtsov, C. Tzagkarakis, D. Antonakaki, and S. Ioannidis, “Identification of twitter bots based on an explainable machine learning framework: The us 2020 elections case study,” in *International conference on web and social media (ICWSM)*, vol. 16, 2022, pp. 956–967.
- [196] L. Lorenzo-Luaces, J. Howard, A. Edinger, H. Y. Yan, L. A. Rutter, D. Valdez, J. Bollen *et al.*, “Sociodemographics and transdiagnostic mental health symptoms in social (studies of online cohorts for internalizing symptoms and language) i and ii: Cross-sectional survey and botometer analysis,” *JMIR Formative Research*, vol. 6, no. 10, p. e39324, 2022.
- [197] E. Sartori, L. Pajola, G. Da San Martino, and M. Conti, “The impact of covid-19 on online discussions: the case study of the sanctioned suicide forum,” in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 4060–4064.
- [198] F. Tahmasbi, L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, and S. Zannettou, ““go eat a bat, chang!”: On the emergence of sinophobic behavior on web communities in the face of covid-19,” in *Proceedings of the web conference 2021*, 2021, pp. 1122–1133.
- [199] R. Rehurek and P. Sojka, “Gensim–python framework for vector space modelling,” *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.
- [200] Wikipedia, “Cosine similarity,” https://en.wikipedia.org/wiki/Cosine_similarity, 2023.
- [201] M. Bastian, S. Heymann, and M. Jacomy, “Gephi: an open source software for exploring and manipulating networks,” in *International conference on web and social media (ICWSM)*, vol. 3, 2009, pp. 361–362.
- [202] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, “Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software,” *PLoS one*, vol. 9, no. 6, p. e98679, 2014.
- [203] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.
- [204] A. Rauchfleisch and J. Kaiser, “The false positive problem of automatic bot detection in social science research,” *PLoS one*, vol. 15, no. 10, p. e0241045, 2020.
- [205] Statista, “Most popular social networks worldwide as of january 2022,” <https://statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>, acc. Oct 2022.

- [206] W. Geysler, “The state of influencer marketing 2022: Benchmark report,” <https://influencermarketinghub.com/influencer-marketing-benchmark-report/>, 2022, acc. Nov 2022.
- [207] X. Liao *et al.*, “Should we trust influencers on social networks? on instagram sponsored post analysis,” in *ICCCN*, 2021.
- [208] P. K. Roy and S. Chahar, “Fake profile detection on social networking websites: a comprehensive review,” *IEEE TAI*, 2020.
- [209] T. Abbas and U. Gadiraju, “Goal-setting behavior of workers on crowdsourcing platforms: An exploratory study on mturk and prolific,” in *AAAI Conference on Human Computation and Crowdsourcing*, 2022.
- [210] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao, “Serf and turf: crowdturfing for fun and profit,” in *21st WWW*, 2012, pp. 679–688.
- [211] T. Elmas, R. Overdorf, A. F. Özkalay, and K. Aberer, “Ephemeral astroturfing attacks: The case of fake twitter trends,” in *2021 IEEE EuroS&P*. IEEE, 2021, pp. 403–422.
- [212] M. Orabi *et al.*, “Detection of bots in social media: A systematic review,” *Information Processing & Management*, vol. 57, no. 4, p. 102250, 2020.
- [213] J. Song, S. Lee, and J. Kim, “Crowdtarget: Target-based detection of crowdturfing in online social networks,” in *CCS*, 2015.
- [214] N. Shi *et al.*, “Semi-supervised random forest for intrusion detection network.” in *MAICS*, 2017, pp. 181–185.
- [215] X. Yang, Q. Yang, and C. Wilson, “Penny for your thoughts: Searching for the 50 cent party on sina weibo,” in *ICWISM*, 2015.
- [216] K. Lee, S. Webb, and H. Ge, “The dark side of micro-task marketplaces: Characterizing fiverr and automatically detecting crowdturfing,” in *ICWISM*, 2014.
- [217] A. Chetan, B. Joshi, H. S. Dutta, and T. Chakraborty, “Corerank: Ranking to detect users involved in blackmarket-based collusive retweeting activities,” in *Proceedings WSDM*, 2019, pp. 330–338.
- [218] H. S. Dutta, K. Aggarwal, and T. Chakraborty, “Decife: Detecting collusive users involved in blackmarket following services on twitter,” in *32nd ACM conference on hypertext and social media*, 2021.
- [219] H. S. Dutta, U. Arora, and T. Chakraborty, “Abome: A multi-platform data repository of artificially boosted online media entities,” in *ICWISM*, 2021.
- [220] G. Voronin, A. Baumann, and S. Lessmann, “Crowdturfing on instagram-the influence of profile characteristics on the engagement of others,” in *Twenty-Sixth European Conference on Information Systems 2016*, 2018.
- [221] H. S. Dutta and T. Chakraborty, “Blackmarket-driven collusion on online media: a survey,” *ACM/IMS Transactions on Data Science (TDS)*, vol. 2, no. 4, pp. 1–37, 2022.
- [222] H. S. Dutta, M. Jobanputra, H. Negi, and T. Chakraborty, “Detecting and analyzing collusive entities on youtube,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 5, pp. 1–28, 2021.

- [223] H. S. Dutta, N. Diwan, and T. Chakraborty, “Weakening the inner strength: Spotting core collusive users in youtube blackmarket network,” in *ICWSSM*, vol. 16, 2022, pp. 147–158.
- [224] J. Wise, “How much time do people spend on social media 2022?” <https://earthweb.com/how-much-time-do-people-spend-on-social-media/>, 2022, acc. Nov 2022.
- [225] L. Manikonda, V. V. Meduri, and S. Kambhampati, “Tweeting the mind and instagramming the heart: Exploring differentiated content sharing on social media,” in *ICWSSM*, 2016.
- [226] D. C. Hernandez-Bocanegra, A. Borchert, F. Brünker, G. K. Shahi, and B. Ross, “Towards a better understanding of online influence: Differences in twitter communication between companies and influencers,” *ACIS 2020 Proceedings*, 2020.
- [227] S. Stieglitz, F. Brachten, B. Ross, and A. Jung, “Do social bots dream of electric sheep? a categorisation of social media bot accounts,” in *27th Australian Conference on Information Systems*, 2018, pp. 1–11.
- [228] G. Thejas *et al.*, “Learning-based model to fight against fake like clicks on instagram posts,” in *2019 SoutheastCon*. IEEE, 2019, pp. 1–8.
- [229] F. C. Akyon and M. E. Kalfaoglu, “Instagram fake and automated account detection,” in *IEEE ASYU*, 2019.
- [230] S. Sheikhi, “An efficient method for detection of fake accounts on the instagram platform.” *Rev. d’Intelligence Artif.*, vol. 34, no. 4, pp. 429–436, 2020.
- [231] K. R. Purba, D. Asirvatham, and R. K. Murugesan, “Classification of instagram fake users using supervised machine learning algorithms,” *IJECE*, vol. 10, no. 3, p. 2763, 2020.
- [232] S. Kim and J. Han, “Detecting engagement bots on social influencer marketing,” in *International Conference on Social Informatics*, 2020.
- [233] N. Z. Gong *et al.*, “Sybilbelief: A semi-supervised learning approach for structure-based sybil detection,” *IEEE TIFS*, 2014.
- [234] M. Al-Qurishi *et al.*, “Sybiltrap: A graph-based semi-supervised sybil defense scheme for online social networks,” *Concurrency and Computation: Practice and Experience*, vol. 30, no. 5, p. e4276, 2018.
- [235] A. Dorri, M. Abadi, and M. Dadfarnia, “Socialbothunter: Botnet detection in twitter-like social networking services using semi-supervised collective classification,” in *DASC/PiCom/DataCom/CyberSciTech*, 2018.
- [236] H. Alvari, E. Shaabani, and P. Shakarian, “Semi-supervised causal inference for identifying pathogenic social media accounts,” in *Identification of Pathogenic Social Media Accounts*. Springer, 2021, pp. 51–61.
- [237] Instagram, “Reducing inauthentic activity on instagram,” <https://about.instagram.com/blog/announcements/reducing-inauthentic-activity-on-instagram>, 2018, acc. Feb 2023.
- [238] —, “Introducing new authenticity measures on instagram,” <https://about.instagram.com/blog/announcements/introducing-new-authenticity-measures-on-instagram/>, 2020, acc. Feb 2023.

- [239] E. Morales, “Instagram bots in 2022 — the best bots and everything else you need to know,” <https://bettermarketing.pub/instagram-bots-in-2021-everything-you-need-to-know-b57fba3b8e9>, 2021, acc. 03-28-2022.
- [240] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. A. Roundy, ““Real Attackers Don’t Compute Gradients”: Bridging the Gap between Adversarial ML Research and Practice,” in *Proceedings of the 1st IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.
- [241] Hashatgsforlikes, “Instagram followers: How many does the average person have?” <https://hashtagsforlikes.co/blog/instagram-followers-how-many-does-the-average-person-have/>, 2020, acc. Mar 2022.
- [242] W. Luo, J. Liu, J. Liu, and C. Fan, “An analysis of security in social networks,” in *2009 IEEE DASC*, 2009.
- [243] K. Lakshmi, “Bot comments on instagram are becoming horrendous,” <https://eatmy.news/2020/11/bot-comments-on-instagram-are-becoming.html>, 2020, acc. Mar 2022.
- [244] B. Chacon, “5 things to know about the instagram algorithm,” <https://later.com/blog/instagram-algorithm/>, 2017, acc. Oct 2022.
- [245] M. Bhargava, P. Mehndiratta, and K. Asawa, “Stylometric analysis for authorship attribution on twitter,” in *International Conference on Big Data Analytics*. Springer, 2013, pp. 37–47.
- [246] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, “Enhancing topic modeling for short texts with auxiliary word embeddings,” 2016.
- [247] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson, “{Trafficking} fraudulent accounts: The role of the underground market in twitter spam and abuse,” in *USENIX Security*, 2013, pp. 195–210.
- [248] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft, “In the mood for being influential on twitter,” in *2011 IEEE PST/SCSM*, 2011.
- [249] C. L. Hanson *et al.*, “Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students,” *Journal of medical Internet research*, vol. 15, no. 4, p. e2503, 2013.
- [250] R. H. Franke and J. D. Kaul, “The hawthorne experiments: First statistical interpretation,” *American sociological review*, pp. 623–643, 1978.
- [251] Statista, “Media usage in an internet minute as of august 2020,” <https://www.statista.com/study/12393/social-networks-statista-dossier>, 2020, accessed on 2021-03-18.
- [252] A. Arshat and D. Etcovitch, “The human cost of online content moderation,” *Harvard Law Review Online, Harvard University, Cambridge, MA, USA*. Retrieved from <https://jolt.law.harvard.edu/digest/the-human-cost-ofonline-content-moderation>, 2018.
- [253] W. Warner and J. Hirschberg, “Detecting hate speech on the world wide web,” in *Proceedings of the second workshop on language in social media*, 2012, pp. 19–26.
- [254] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings,” in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 29–30.

- [255] J. Vincent, “Facebook is now using ai to sort content for quicker moderation,” www.theverge.com/2020/11/13/21562596/facebook-ai-moderation, 2020, accessed on 2020-11-13.
- [256] Instagram, “Introducing sensitive content control,” about.fb.com/news/2021/07/introducing-sensitive-content-control/, 2021, accessed on 2021-08-10.
- [257] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, “Evasion attacks against machine learning at test time,” in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.
- [258] Q. Xiao, Y. Chen, C. Shen, Y. Chen, and K. Li, “Seeing is not believing: Camouflage attacks on image scaling algorithms,” in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 443–460.
- [259] L. Pajola and M. Conti, “Fall of giants: How popular text-based mlapps fall against a simple evasion attack,” in *2021 IEEE European Symposium on Security and Privacy (EuroSP)*, 2021, pp. 198–211.
- [260] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2015.
- [261] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, “Can machine learning be secure?” in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, 2006, pp. 16–25.
- [262] B. I. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, S. Rao, N. Taft, and J. D. Tygar, “Antidote: understanding and defending against poisoning of anomaly detectors,” in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, 2009, pp. 1–14.
- [263] Q. Xiao, K. Li, D. Zhang, and W. Xu, “Security risks in deep learning implementations,” in *IEEE Security and Privacy Workshops*, 2018, pp. 123–128.
- [264] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, “Adversarial attacks on deep-learning models in natural language processing: A survey,” *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 3, 2020.
- [265] L. Chen, J. Sun, and W. Xu, “Fawa: Fast adversarial watermark attack on optical character recognition (ocr) systems,” in *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2021, pp. 547–563.
- [266] L. Chen and W. Xu, “Attacking optical character recognition (ocr) systems with adversarial watermarks,” 2020.
- [267] T. Gillespie, “Content moderation, ai, and the question of scale,” *Big Data & Society*, vol. 7, no. 2, p. 2053951720943234, 2020.
- [268] S. Sun, Y. Liu, and L. Mao, “Multi-view learning for visual violence recognition with maximum entropy discrimination and deep features,” *Information Fusion*, vol. 50, pp. 43–53, 2019.
- [269] F. Mayer and M. Steinebach, “Forensic image inspection assisted by deep learning,” in *Proceedings of the 12th International Conference on Availability, Reliability and Security*. Association for Computing Machinery, 2017.

- [270] A. Schmidt and M. Wiegand, “A survey on hate speech detection using natural language processing,” in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Apr. 2017, pp. 1–10.
- [271] D. Kiela, H. Firouz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, “The hateful memes challenge: Detecting hate speech in multimodal memes,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 2611–2624.
- [272] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, “Exploring hate speech detection in multimodal publications,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, March 2020.
- [273] R. Velioglu and J. Rose, “Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge,” *arXiv preprint arXiv:2012.12975*, 2020.
- [274] M. J. Wolf, K. Miller, and F. S. Grodzinsky, “Why we should have seen that coming: Comments on microsoft’s tay ”experiment,” and wider implications,” *SIGCAS Comput. Soc.*, vol. 47, no. 3, p. 54–64, Sep. 2017.
- [275] K. Yuan, D. Tang, X. Liao, X. Wang, X. Feng, Y. Chen, M. Sun, H. Lu, and K. Zhang, “Stealthy porn: Understanding real-world adversarial images for illicit online promotion,” in *Symposium on Security and Privacy*. IEEE, 2019, pp. 952–966.
- [276] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan, “All you need is “love”: Evading hate speech detection,” in *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. Association for Computing Machinery, 2018, p. 2–12.
- [277] E. Stewart, “Detecting fake news: Two problems for content moderation,” *Philosophy & technology*, vol. 34, no. 4, pp. 923–940, 2021.
- [278] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford, “Captcha: Using hard ai problems for security,” in *International conference on the theory and applications of cryptographic techniques*. Springer, 2003, pp. 294–311.
- [279] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, “Multi-digit number recognition from street view imagery using deep convolutional neural networks,” *arXiv preprint arXiv:1312.6082*, 2013.
- [280] V. P. Singh and P. Pal, “Survey of different types of captcha,” *International Journal of Computer Science and Information Technologies*, vol. 5, no. 2, pp. 2242–2245, 2014.
- [281] M. Guerar, L. Verderame, M. Migliardi, F. Palmieri, and A. Merlo, “Gotta captcha’em all: A survey of twenty years of the human-or-computer dilemma,” *arXiv preprint arXiv:2103.01748*, 2021.
- [282] K. Krol, S. Parkin, and M. A. Sasse, “Better the devil you know: A user study of two captchas and a possible replacement technology,” in *NDSS Workshop on Usable Security*, vol. 10, 2016.
- [283] K. Chellapilla, K. Larson, P. Y. Simard, and M. Czerwinski, “Computers beat humans at single character recognition in reading based human interaction proofs.” in *CEAS*, 2005.

- [284] G. Mori and J. Malik, "Recognizing objects in adversarial clutter: Breaking a visual captcha," in *Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1. IEEE, 2003, pp. I-I.
- [285] J. Yan and A. S. El Ahmad, "A low-cost attack on a microsoft captcha," in *Proceedings of the 15th ACM conference on Computer and communications security*, 2008, pp. 543-554.
- [286] E. Bursztein, J. Aigrain, A. Moscicki, and J. C. Mitchell, "The end is nigh: Generic solving of text-based captchas," in *8th {USENIX} Workshop on Offensive Technologies*, 2014.
- [287] J. Yan, "A simple generic attack on text captchas," 2016.
- [288] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161-184, 2021.
- [289] Hatebase, "Hatebase," hatebase.org/, 2021, accessed on 2021-03-22.
- [290] J. Memon, M. Sami, R. Khan, and M. Uddin, "Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr)," *IEEE Access*, pp. 1-1, 07 2020.
- [291] H. Lin, P. Yang, and F. Zhang, "Review of scene text detection and recognition," *Archives of computational methods in engineering*, vol. 27, no. 2, pp. 433-454, 2020.
- [292] G. V. Cormack *et al.*, "Email spam filtering: A systematic review," *Foundations and Trends® in Information Retrieval*, vol. 1, no. 4, pp. 335-455, 2008.
- [293] B. Biggio, G. Fumera, I. Pillai, and F. Roli, "Improving image spam filtering using image text features," in *Proc of the fifth conf on email and anti-spam*, 2008.
- [294] A. Agiollo, M. Conti, P. Kaliyar, T.-N. Lin, and L. Pajola, "Detonar: Detection of routing attacks in rpl-based iot," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1178-1190, 2021.
- [295] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19-31, 2016.
- [296] J. C. Gomez, "pinterest_dataset v3," [10.17632/fs4k2zcc5j5.3](https://www.kaggle.com/jc-gomez/pinterest-dataset-v3), 2018.
- [297] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 308-317.
- [298] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64-73, 2016.
- [299] T. Guo, J. Dong, H. Li, and Y. Gao, "Simple convolutional neural network on image classification," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*. IEEE, 2017, pp. 721-724.
- [300] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097-1105, 2012.
- [301] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

- [302] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [303] N. Inkawich, "Finetuning torchvision models," pytorch.org/tutorials/beginner/finetuning_torchvision_models_tutorial.html, 2021, accessed on 2021-09-10.
- [304] Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, and G. H. Chen, "Ecod: Unsupervised outlier detection using empirical cumulative distribution functions," *arXiv preprint arXiv:2201.00382*, 2022.
- [305] Y. Zhao, Z. Nasrullah, and Z. Li, "Pyod: A python toolbox for scalable outlier detection," *Journal of Machine Learning Research*, vol. 20, no. 96, pp. 1–7, 2019. [Online]. Available: <http://jmlr.org/papers/v20/19-011.html>
- [306] B. Biggio, G. Fumera, I. Pillai, and F. Roli, "A survey and experimental evaluation of image spam filtering techniques," *Pattern recognition letters*, vol. 32, no. 10, pp. 1436–1446, 2011.
- [307] E. B. Smith, R. A. Brands, M. E. Brashears, and A. M. Kleinbaum, "Social networks and cognition," *Annual Review of Sociology*, vol. 46, no. 1, pp. 159–174, 2020.
- [308] D. Fisher and A. McAdam, "Social traits, social networks and evolutionary biology," *Journal of Evolutionary Biology*, vol. 30, no. 12, pp. 2088–2103, 2017.
- [309] L. Hagen, T. Keller, S. Neely, N. DePaula, and C. Robert-Cooperman, "Crisis communications in the age of social media: A network analysis of zika-related tweets," *Social science computer review*, vol. 36, no. 5, pp. 523–541, 2018.
- [310] R. E. Kim and L. J. Kotzé, "Planetary boundaries at the intersection of earth system law, science and governance: A state-of-the-art review," *Review of European, Comparative & International Environmental Law*, vol. 30, no. 1, pp. 3–15, 2021.
- [311] Alexa, "Alexa top websites," <https://www.expireddomains.net/alexa-top-websites/>, 2022, accessed: Sep. 2022.
- [312] Karl, "The 15 biggest social media sites and apps," <https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/>, 2022, accessed: Sep. 2022.
- [313] F. Richter, "Social networking is the no. 1 online activity in the u.s." <https://www.statista.com/chart/1238/digital-media-use-in-the-us/>, 2022, accessed: Sep. 2022.
- [314] P. Rani and J. Shokeen, "A survey of tools for social network analysis." *Int. J. Web Eng. Technol.*, vol. 16, no. 3, pp. 189–216, 2021.
- [315] Infographic, "Data never sleeps 5.0," <https://www.domo.com/learn/infographic/data-never-sleeps-5>, 2022, accessed: Oct. 2022.
- [316] P. Brooker, J. Barnett, T. Cribbin, and S. Sharma, "Have we even solved the first 'big data challenge?' practical issues concerning data collection and visual representation for social media analytics," in *Digital methods for social science*. Springer, 2016, pp. 34–50.

- [317] D. Boyd and K. Crawford, “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,” *Information, communication & society*, vol. 15, no. 5, pp. 662–679, 2012.
- [318] A. K. Jain, S. R. Sahoo, and J. Kaubiya, “Online social networks security and privacy: comprehensive review and analysis,” *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2157–2177, 2021.
- [319] X. Hu, J. Tang, and H. Liu, “Online social spammer detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, 2014.
- [320] Y. Zhu, X. Wang, E. Zhong, N. Liu, H. Li, and Q. Yang, “Discovering spammers in social networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, 2012, pp. 171–177.
- [321] N. S. Murugan and G. U. Devi, “Detecting spams in social networks using ml algorithms-a review,” *International Journal of Environment and Waste Management*, vol. 21, no. 1, pp. 22–36, 2018.
- [322] S. Webb, J. Caverlee, and C. Pu, “Social honeypots: Making friends with a spammer near you.” in *CEAS*. San Francisco, CA, 2008, pp. 1–10.
- [323] K. Lee, J. Caverlee, and S. Webb, “Uncovering social spammers: social honeypots+ machine learning,” in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 435–442.
- [324] G. Stringhini, C. Kruegel, and G. Vigna, “Detecting spammers on social networks,” in *Proceedings of the 26th annual computer security applications conference*, 2010, pp. 1–9.
- [325] M. for Developers, “Instagram api,” <https://developers.facebook.com/docs/instagram-api/guides/insights>, 2021, accessed on: Oct. 2022.
- [326] M. Robertson, *Instagram Marketing: How to Grow Your Instagram Page And Gain Millions of Followers Quickly With Step-by-Step Social Media Marketing Strategies*. CreateSpace Independent Publishing Platform, 2018.
- [327] M. Hub, “The state of influencer marketing 2021: Benchmark report,” <https://influencermarketinghub.com/influencer-marketing-benchmark-report-2021>, 2021, accessed on: Oct. 2022.
- [328] W. Stallings, L. Brown, M. D. Bauer, and M. Howard, *Computer security: principles and practice*. Pearson Upper Saddle River, 2012, vol. 2.
- [329] “Heat-seeking honeypots: design and experience,” in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 207–216.
- [330] V. Yegneswaran, J. T. Giffin, P. Barford, and S. Jha, “An architecture for generating semantic aware signatures.” in *USENIX security symposium*, 2005, pp. 97–112.
- [331] C. Kreibich and J. Crowcroft, “Honeycomb: creating intrusion detection signatures using honeypots,” *ACM SIGCOMM computer communication review*, vol. 34, no. 1, pp. 51–56, 2004.
- [332] A. Moshchuk, T. Bragin, S. D. Gribble, and H. M. Levy, “A crawler-based study of spyware in the web.” in *NDSS*, vol. 1, 2006, p. 2.

- [333] Y.-M. Wang, D. Beck, X. Jiang, and R. Roussev, “Automated web patrol with strider honeymonkeys: Finding web sites that exploit browser vulnerabilities,” in *IN NDSS*. Citeseer, 2006.
- [334] E. De Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq, “Paying for likes? understanding facebook like fraud using honeypots,” in *Proceedings of the 2014 Conference on Internet Measurement Conference*, 2014, pp. 129–136.
- [335] K. Lee, B. Eoff, and J. Caverlee, “Seven months with the devils: A long-term study of content polluters on twitter,” in *Proceedings of the international AAAI conference on web and social media*, vol. 5, 2011, pp. 185–192.
- [336] C. Yang, J. Zhang, and G. Gu, “A taste of tweets: Reverse engineering twitter spammers,” in *Proceedings of the 30th annual computer security applications conference*, 2014, pp. 86–95.
- [337] Y. Zhang, H. Zhang, and X. Yuan, “Toward efficient spammers gathering in twitter social networks,” in *Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy*, 2019, pp. 157–159.
- [338] C. Campbell, C. Ferraro, and S. Sands, “Segmenting consumer reactions to social network marketing,” *European Journal of Marketing*, 2014.
- [339] S. D. McClurg, “Social networks and political participation: The role of social interaction in explaining political participation,” *Political research quarterly*, vol. 56, no. 4, pp. 449–464, 2003.
- [340] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, “The spreading of misinformation online,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016.
- [341] W. Ahmed, J. Vidal-Alaball, J. Downing, F. L. Seguí *et al.*, “Covid-19 and the 5g conspiracy theory: social network analysis of twitter data,” *Journal of medical internet research*, vol. 22, no. 5, p. e19458, 2020.
- [342] N. Dey, S. Borah, R. Babo, and A. S. Ashour, *Social network analytics: computational research methods and techniques*. Academic Press, 2018.
- [343] A. Singh, M. N. Halgamuge, and B. Moses, “An analysis of demographic and behavior trends using social media: Facebook, twitter, and instagram,” *Social Network Analytics*, p. 87, 2019.
- [344] N. M. Ferreira, “300+ best instagram captions and selfie quotes for your photos,” <https://www.oberlo.com/blog/instagram-captions>, 2022, accessed: Sep. 2022.
- [345] J. Petriska, <https://gist.github.com/JakubPetriska/060958fd744ca34f099e947cdo80b540>, 2022, accessed: Oct. 2022.
- [346] C. Laurence, “Call to action instagram: 13 creative ctas to test on your account,” <https://www.planthat.com/call-to-action-instagram/>, 2022, accessed: Sep. 2022.
- [347] K. McCormick, “23 smart ways to get more instagram followers in 2022,” <https://www.wordstream.com/blog/ws/get-more-instagram-followers>, 2022, accessed: Sep. 2022.
- [348] A. Daugherty, <https://aigrow.me/follow-unfollow-instagram/>, 2022, accessed: Oct. 2022.

- [349] R. Aditya, D. Prafulla, N. Alex, C. Casey, and C. Mark, <https://openai.com/product/dall-e-2>, 2022, accessed: Mar. 2023.
- [350] OpenAI, <https://openai.com/blog/chatgpt>, 2022, accessed: Mar. 2023.
- [351] L. Meyer, “How often to post on social media: 2022 success guide,” <https://louisem.com/144557/often-post-social-media>, 2022, accessed: Oct. 2022.
- [352] SocialBuddy, “How often to post on social media: 2022 success guide,” <https://socialbuddy.com/how-often-should-you-post-on-instagram/>, 2022, accessed: Oct. 2022.
- [353] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [354] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [355] H. Face, “Keytotext,” <https://huggingface.co/gagan3012/k2t>, 2022, accessed on: Oct. 2022.
- [356] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer.” *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- [357] B. Dayma, S. Patil, P. Cuenca, K. Saifullah, T. Abraham, P. Le Khac, L. Melas, and R. Ghosh, “Dalle-mini,” 7 2021. [Online]. Available: <https://github.com/borisdavyma/dalle-mini>
- [358] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” 2019.
- [359] R. Mushtaq, “Augmented dickey fuller test,” *Econometrics: Mathematical Methods & Programming eJournal*, 2011.
- [360] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, “Low-quality product review detection in opinion summarization,” in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007, pp. 334–342.
- [361] H. HQ, “How to get followers on instagram,” <https://www.hopperhq.com/blog/how-to-get-followers-instagram-2021/>, 2022, accessed: Jan. 2023.
- [362] AppsUK, “How long does it take to get 1000 followers on instagram?” <https://apps.uk/how-long-1000-followers-on-instagram/>, 2022, accessed: Jan. 2023.
- [363] H. Macready, “The only instagram metrics you really need to track in 2023,” <https://blog.hootsuite.com/instagram-metrics>, 2022, accessed: Jan. 2023.
- [364] I. Me, “How to get your first 1000 followers on instagram,” <https://www.epidemicsound.com/blog/how-to-get-your-first-1000-followers-on-instagram/>, 2022, accessed: Jan. 2023.

- [365] Lavanya, “How to avoid-stop spam comments on instagram posts?” <https://versionweekly.com/news/instagram/how-to-avoid-stop-spam-comments-on-instagram-posts-easy-method/>, 2021, accessed on: Oct. 2022.
- [366] N. Pereira, “5 different tiers of influencers and when to use each,” <https://zerogravitymarketing.com/the-different-tiers-of-influencers-and-when-to-use-each/>, 2022, accessed: Oct. 2022.
- [367] Y. Xiao, Y. Jia, X. Cheng, S. Wang, J. Mao, and Z. Liang, “I know your social network accounts: A novel attack architecture for device-identity association,” *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2022.
- [368] N. Vishwamitra, Y. Li, H. Hu, K. Caine, L. Cheng, Z. Zhao, and G.-J. Ahn, “Towards automated content-based photo privacy control in user-centered social networks,” in *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*. Association for Computing Machinery, 2022.
- [369] S. Raponi, Z. Khalifa, G. Oligeri, and R. Di Pietro, “Fake news propagation: A review of epidemic models, datasets, and insights,” *ACM Trans. Web*, vol. 16, no. 3, 2022.
- [370] W. Zhang and H.-M. Sun, “Instagram spam detection,” in *2017 IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC)*. IEEE, 2017, pp. 227–228.
- [371] S. Kuhn, “How to stop instagram spam?” <https://www.itgears.com/how-to-stop-instagram-spam/>, 2022, accessed: Jan. 2023.
- [372] N. A. Haqimi, N. Rokhman, and S. Priyanta, “Detection of spam comments on instagram using complementary naïve bayes,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 3, pp. 263–272, 2019.
- [373] D. Dittrich, “The ethics of social honeypots,” *Research Ethics*, vol. 11, no. 4, pp. 192–210, 2015.
- [374] J. Howarth, “How many gamers are there? (new 2023 statistics),” 2021, september, 2023. [Online]. Available: <https://explodingtopics.com/blog/number-of-gamers>
- [375] Statista, “Video games - worldwide,” 2023, september, 2023. [Online]. Available: <https://www.statista.com/outlook/dmo/digital-media/video-games/worldwide/>
- [376] E. McDonald, “Aci and newzoo whitepaper: Turning players into payers | understanding the gaming payments experience,” <https://newzoo.com/insights/articles/aci-and-newzoo-whitepaper-turning-players-into-payers/>, 2018, accessed: March 8, 2020.
- [377] E. Chung, “Playstation data breach deemed in ‘top 5 ever’,” *CBC News*, 2011.
- [378] Valve, “Security and trading,” <https://store.steampowered.com/news/19618/>, 2015, accessed: March 11, 2020.
- [379] A. Blog, “League of legends gamers targeted by phishing scam | avast,” <https://securityboulevard.com/2018/10/league-of-legends-gamers-targeted-by-phishing-scam-avast/>, 2018, accessed: March 8, 2020.
- [380] C. D’anastasio, “What’s really going on with all those hacked fortnite accounts,” <https://kotaku.com/whats-really-going-on-with-all-those-hacked-fortnite-ac-1823965781>, 2018, accessed: March 10, 2020.

- [381] Z. Whittaker, “Nintendo now says 300,000 accounts breached by hackers,” <https://techcrunch.com/2020/06/09/nintendo-accounts-affected-breach/>, 2020, accessed: March 8, 2020.
- [382] R. Stanton, “Bandai namco confirms ransomware hack of internal servers,” <https://www.pcgamer.com/bandai-namco-confirms-ransomware-hack-of-internal-servers/>, 2022, accessed: September 20, 2023.
- [383] M. Fryling, J. L. Cotler, J. Rivituso, L. Mathews, and S. Pratico, “Cyberbullying or normal game play? impact of age, gender, and experience on cyberbullying in multi-player online gaming environments: Perceptions from one gaming forum,” *Journal of Information Systems Applied Research*, vol. 8, no. 1, p. 4, 2015.
- [384] H. Kwak, J. Blackburn, and S. Han, “Exploring cyberbullying and other toxic behavior in team competition online games,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 3739–3748.
- [385] M. Conti and P. P. Tricomi, “Pvp: Profiling versus player! exploiting gaming data for player recognition,” in *Information Security*, W. Susilo, R. H. Deng, F. Guo, Y. Li, and R. Intan, Eds. Cham: Springer International Publishing, 2020, pp. 393–408.
- [386] D. Gong, H. He, D. Liu, W. Ma, L. Dong, C. Luo, and D. Yao, “Enhanced functional connectivity and increased gray matter volume of insula related to action video game playing,” *Scientific reports*, vol. 5, p. 9763, 2015.
- [387] J. P. Gee, “Video games and embodiment,” vol. 3, no. 3-4, 2008, pp. 253–263.
- [388] S. S. Kim, K. M. Huang-Isherwood, W. Zheng, and D. Williams, “The art of being together: How group play can increase reciprocity, social capital, and social status in a multiplayer online game,” *Computers in Human Behavior*, vol. 133, p. 107291, 2022.
- [389] V. Kovess-Masfety, K. Keyes, A. Hamilton, G. Hanson, A. Bitfoi, D. Golitz, C. Koç, R. Kuijpers, S. Lesinskiene, Z. Mihova *et al.*, “Is time spent playing video games associated with mental health, cognitive and social skills in young children?” *Social psychiatry and psychiatric epidemiology*, vol. 51, no. 3, pp. 349–357, 2016.
- [390] K. Squire, “Video games in education,” *Int. J. Intell. Games & Simulation*, vol. 2, no. 1, pp. 49–62, 2003.
- [391] D. W. Shaffer, K. R. Squire, R. Halverson, and J. P. Gee, “Video games and the future of learning,” *Phi delta kappan*, vol. 87, no. 2, pp. 105–111, 2005.
- [392] M. Griffiths, “Video games and health,” pp. 122–123, 2005.
- [393] I. Granic, A. Lobel, and R. C. Engels, “The benefits of playing video games.” *American psychologist*, vol. 69, no. 1, p. 66, 2014.
- [394] J. B. Funk, H. B. Baldacci, T. Pasold, and J. Baumgardner, “Violence exposure in real-life, video games, television, movies, and the internet: is there desensitization?” *Journal of adolescence*, vol. 27, no. 1, pp. 23–39, 2004.

- [395] F. Öztütüncü Doğan, “Video games and children: Violence in video games,” 2006.
- [396] P. M. Markey, C. N. Markey, and J. E. French, “Violent video games and real-world violence: Rhetoric versus data.” *Psychology of Popular Media Culture*, vol. 4, no. 4, p. 277, 2015.
- [397] A. Smuts, “Are video games art?” vol. 3, 2005. [Online]. Available: https://digitalcommons.risd.edu/liberalarts_contempaesthetics/vol3/iss1/6/
- [398] A. Drachen, A. Canossa, and G. N. Yannakakis, “Player modeling using self-organization in tomb raider: Underworld,” in *2009 IEEE Symposium on Computational Intelligence and Games*, Sep. 2009, pp. 1–8.
- [399] T. Mahlmann, A. Drachen, J. Togelius, A. Canossa, and G. N. Yannakakis, “Predicting player behavior in tomb raider: Underworld,” in *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games*, Aug 2010, pp. 178–185.
- [400] S. Müller, M. Kapadia, S. Frey, S. Klinger, R. P. Mann, B. Solenthaler, R. W. Sumner, and M. Gross, “Statistical analysis of player behavior in minecraft,” in *Proceedings of the 10th International Conference on the Foundations of Digital Games*. Society for the Advancement of the Science of Digital Games, 2015.
- [401] A. Smerdov, A. Kiskun, R. Shaniiazov, A. Somov, and E. Burnaev, “Understanding cyber athletes behaviour through a smart chair: Cs:go and monolith team scenario,” in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, 2019, pp. 973–978.
- [402] J. Newman and J. Jerome, “Press start to track privacy and the new questions posed by modern video game technology,” *AIPLA QJ*, vol. 42, p. 527, 2014.
- [403] N. C. Russell, J. R. Reidenberg, and S. Moon, “Privacy in gaming,” *Fordham Intell. Prop. Media & Ent. LJ*, vol. 29, p. 61, 2018.
- [404] L. Jędrzejczyk, B. A. Price, A. K. Bandara, B. Nuseibeh, W. Hall, and M. Keynes, “I know what you did last summer: risks of location data leakage in mobile and social computing,” *Department of Computing Faculty of Mathematics, Computing and Technology The Open University*, pp. 1744–1986, 2009.
- [405] M. Conti, M. Nati, E. Rotundo, and R. Spolaor, “Mind the plug! laptop-user recognition through power consumption,” in *Proceedings of the 2nd ACM International Workshop on IoT Privacy, Trust, and Security*, 2016, pp. 37–44.
- [406] L. Gao, J. Judd, D. Wong, and J. Lowder, “Classifying dota 2 hero characters based on play style and performance,” *Univ. of Utah Course on ML*, 2013.
- [407] C. Eggert, M. Herrlich, J. Smeddinck, and R. Malaka, “Classification of player roles in the team-based multi-player game dota 2,” 09 2015, pp. 112–125.
- [408] OpenAI, C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. de Oliveira Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang, “Dota 2 with large scale deep reinforcement learning,” 2019. [Online]. Available: <https://arxiv.org/abs/1912.06680>

- [409] X. Qian, R. Sifa, X. Liu, S. Ganguly, B. Yadamsuren, D. Klabjan, A. Drachen, and S. Demediuk, "Anomaly detection in player performances in multiplayer online battle arena games," in *Proceedings of the 2022 Australasian Computer Science Week*, 2022, pp. 23–30.
- [410] Y. J. Ding, "Profiling smurfs and boosters on dota 2 using k-means," Ph.D. dissertation, UTAR, 2021.
- [411] K. Yamamoto and V. McArthur, "Digital economies and trading in counter strike global offensive: How virtual items are valued to real world currencies in an online barter-free market," in *2015 IEEE Games Entertainment Media Conference (GEM)*, 2015, pp. 1–6.
- [412] M. N. Rizani and H. Iida, "Analysis of counter-strike: Global offensive," in *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 2018, pp. 373–378.
- [413] P. Xenopoulos, H. Doraiswamy, and C. Silva, "Valuing player actions in counter-strike: Global offensive," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 1283–1292.
- [414] I. Makarov, D. Savostyanov, B. Litvyakov, and D. I. Ignatov, "Predicting winning team and probabilistic ratings in "dota 2" and "counter-strike: Global offensive" video games," in *Analysis of Images, Social Networks and Texts: 6th International Conference, AIST 2017, Moscow, Russia, July 27–29, 2017, Revised Selected Papers 6*. Springer, 2018, pp. 183–196.
- [415] M. Rusk, Fredrik & Stahl, "A ca perspective on kills and deaths in counter-strike: Global offensive video game play." in *Social Interaction - Video-Based Studies of Human Sociality*, 2020.
- [416] Sasmoko, J. Harsono, Y. Udjaja, Y. Indrianti, and J. Moniaga, "The effect of game experience from counter-strike: Global offensive," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIIT)*, 2019, pp. 374–378.
- [417] M. Ståhl and F. Rusk, "Player customization, competence and team discourse: Exploring player identity (co) construction in counter-strike: Global offensive," *Game Studies*, vol. 20, no. 4, 2020.
- [418] M. Hoogendoorn and B. Funk, "Machine learning for the quantified self," *On the art of learning from sensory data*, 2018.
- [419] N. K. Manaswi and N. K. Manaswi, "Rnn and lstm," *Deep Learning with Applications Using Python: Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras*, pp. 115–126, 2018.
- [420] J. Connor, R. Martin, and L. Atlas, "Recurrent neural networks and robust time series prediction," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 240–254, 1994.
- [421] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [422] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbdo53c1c4a845aa-Paper.pdf

- [423] activePlayer, “Cs:go active players,” 2021, 30, May, 2023. [Online]. Available: <https://activeplayer.io/counter-strike-global-offensive/>
- [424] —, “Dota 2 active players,” 2021, 30, May, 2023. [Online]. Available: <https://activeplayer.io/dota-2/>
- [425] —, “Fortnite active players,” 2021, 30, May, 2023. [Online]. Available: <https://activeplayer.io/fortnite/>
- [426] —, “Lol active players,” 2021, 30, May, 2023. [Online]. Available: <https://activeplayer.io/league-of-legends/>
- [427] —, “Minecraft active players,” 2021, 30, May, 2023. [Online]. Available: <https://activeplayer.io/minecraft/>
- [428] —, “PUBG active players,” 2021, 30, May, 2023. [Online]. Available: <https://activeplayer.io/pubg/>
- [429] steam charts, “Steam charts,” 2021, 30, May, 2023. [Online]. Available: <https://steamcharts.com/>
- [430] Y. Zhang, Y. Huang, L. Wang, and S. Yu, “A comprehensive study on gait biometrics using a joint CNN-based method,” *Pattern Recognition*, vol. 93, pp. 228–236, 2019.
- [431] Z. Liang, F. Tan, and Z. Chi, “Video-based biometric identification using eye tracking technique,” in *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*. IEEE, 2012, pp. 728–733.
- [432] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [433] A. Bendale and T. E. Boult, “Towards open set deep networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.
- [434] E. Conroy, M. Kowal, A. J. Toth, and M. J. Campbell, “Boosting: Rank and skill deception in esports,” *Entertainment Computing*, vol. 36, p. 100393, 2021.
- [435] V. L. Rubin and S. C. Camm, “Deception in video games: examining varieties of griefing,” *Online Information Review*, vol. 37, no. 3, pp. 369–387, 2013.
- [436] C. J. Lennings, K. L. Amon, H. Brummert, and N. J. Lennings, “Grooming for terror: The internet and young people,” *Psychiatry, Psychology and Law*, vol. 17, no. 3, pp. 424–437, 2010.
- [437] T. Wijman, “The games market and beyond in 2021: The year in numbers,” <https://newzoo.com/insights/articles/the-games-market-in-2021-the-year-in-numbers-esports-cloud-gaming>, 2021, accessed: June 2022.
- [438] M. Barr and A. C. Stewart, “Playing video games during the COVID-19 pandemic and effects on players’ well-being,” *Games & Culture*, 2022.
- [439] A. C. Tally, Y. R. Kim, K. Boustani, and C. Nippert-Eng, “Protect and project: Names, privacy, and the boundary negotiations of online video game players,” *Proc. ACM Human-Comp. Inter.*, 2021.
- [440] J. Hamari and M. Sjöblom, “What is esports and why do people watch it?” *Internet research*, 2017.

- [441] C. Cough, “Share of gamers who want to become professional gamers in the future worldwide in 2020, by gender,” <https://www.statista.com/statistics/1132968/professionals-gamers-gender/>, 2020, accessed: June, 2022.
- [442] M. Kaytoue, A. Silva, L. Cerf, W. Meira Jr, and C. Raïssi, “Watch me playing, i am a professional: a first study on video game live streaming,” in *Proc. Int. Conf. World Wide Web*, 2012.
- [443] “The international,” https://liquipedia.net/dota2/The_International, 2022.
- [444] W. Guo, X. Wu, S. Huang, and X. Xing, “Adversarial policy learning in two-player competitive games,” in *Int. Conf. Machin. Learn.*, 2021.
- [445] M. D. Griffiths, “The psychosocial impact of professional gambling, professional video gaming & eSports,” *Casino & Gaming International*, 2017.
- [446] Stratz, “Accounts and matches analyzed by stratz,” <https://stratz.com/welcome>, 2022, accessed: July, 2022.
- [447] E. Commission, “Sensitive data,” https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data_en.
- [448] U. D. of the Treasury, “Sensitive personal data,” <https://home.treasury.gov/taxonomy/term/7651>.
- [449] S. Tekofsky, J. Van Den Herik, P. Spronck, and A. Plaat, “Psyops: Personality assessment through gaming behavior,” in *In Proceedings of the International Conference on the Foundations of Digital Games*. Citeseer, 2013.
- [450] D. Martinovic, V. Ralevich, J. McDougall, and M. Perklin, ““you are what you play”: Breaching privacy and identifying users in online gaming,” in *Proc. IEEE Ann. Int. Conf. Priv. Secur. Trust*, 2014.
- [451] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” *Pattern Recognition*, vol. 84, pp. 317–331, 2018.
- [452] J. M. Spring, T. Moore, and D. Pym, “Practicing a science of security: a philosophy of science perspective,” in *New Secur. Paradig. Workshop*, 2017.
- [453] P. Karkallis, J. Blasco, G. Suarez-Tangil, and S. Pastrana, “Detecting video-game injectors exchanged in game cheating communities,” in *Europ. Symp. Res. Comp. Secur.*, 2021.
- [454] Chadlantis, “How to improve in any video game,” <https://medium.com/@chadlantistv/how-to-improve-in-any-video-game-7doefe5edo53>, 2019.
- [455] U. Weinsberg, S. Bhagat, S. Ioannidis, and N. Taft, “Blurme: Inferring and obfuscating user gender based on ratings,” in *Proceedings of the sixth ACM conference on Recommender systems*, 2012, pp. 195–202.
- [456] I. Jarin and B. Eshete, “Pricure: privacy-preserving collaborative inference in a multi-party setting,” in *Proc. ACM Workshop Secur. Privacy Analytics*, 2021.
- [457] D. Zhong, H. Sun, J. Xu, N. Gong, and W. H. Wang, “Understanding disparate effects of membership inference attacks and their countermeasures,” in *Proc. ACM AsiaCCS*, 2022.

- [458] Y. Cheng, J. Park, and R. Sandhu, "Preserving user privacy from third-party applications in online social networks," in *Int. Conf. World Wide Web*, 2013.
- [459] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Ioannidis, "Face/off: Preventing privacy leakage from photos in social networks," in *Proc. ACM CCS*, 2015.
- [460] J. Morris, S. Newman, K. Palaniappan, J. Fan, and D. Lin, "'do you know you are tracked by photos that you didnt take: large-scale location-aware multi-party image privacy protection,'" *IEEE TDSC*, 2021.
- [461] G. Apruzzese et al, "The role of machine learning in cybersecurity," *ACM Digital Threats: Research and Practice*, 2022.
- [462] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *CHI—Human Factors in Computing Systems*, 2011, pp. 253–262.
- [463] D. Jurgens, T. Finethy, J. McCorriston, Y. Xu, and D. Ruths, "Geolocation prediction in twitter using social networks: A critical analysis and review of current practice," in *Proc. Int. AAAI Conf. Web Social Media*, 2015.
- [464] N. Z. Gong and B. Liu, "Attribute inference attacks in online social networks," *ACM T. Privacy Secur.*, 2018.
- [465] Y. Zhang, N. Gao, and J. Chen, "A practical defense against attribute inference attacks in session-based recommendations," in *IEEE Int. Conf. Web Serv.*, 2020.
- [466] B. A. Pijani, A. Imine, and M. Rusinowitch, "You are what emojis say about your pictures: language-independent gender inference attack on Facebook," in *Proc. ACM Symp. Appl. Comp.*, 2020.
- [467] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Academy Sciences*, 2013.
- [468] T. Chen, R. Boreli, M.-A. Kaafar, and A. Friedman, "On the effectiveness of obfuscation techniques in online social networks," in *Int. Priv. Enhancing Techn. Symp.*, 2014.
- [469] T. Yo and K. Sasahara, "Inference of personal attributes from tweets using machine learning," in *IEEE Int. Conf. Big Data*, 2017.
- [470] S. Eidizadehakhcheloo, B. Alipour Pijani, A. Imine, and M. Rusinowitch, "Divide-and-learn: A random indexing approach to attribute inference attacks in online social networks," in *IFIP Ann. Conf. Data Appl. Secur. Privacy*, 2021.
- [471] J. Oggins and J. Sammis, "Notions of video game addiction and their relation to self-reported addiction among players of world of warcraft," *International Journal of Mental Health and Addiction*, vol. 10, no. 2, pp. 210–230, 2012.
- [472] K. L. Nowak and C. Rauh, "The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction," *Journal of Computer-Mediated Communication*, vol. 11, no. 1, pp. 153–178, 2005.

- [473] P. Likarish, O. Brdiczka, N. Yee, N. Ducheneaut, and L. Nelson, “Demographic profiling from mmog gameplay,” in *11th Privacy Enhancing Technologies Symposium. Waterloo, Canada*. Citeseer, 2011.
- [474] C. Symborski, G. M. Jackson, M. Barton, G. Cranmer, B. Raines, and M. M. Quinn, “The use of social science methods to predict player characteristics from avatar observations,” in *Predicting real world behaviors from virtual world data*. Springer, 2014, pp. 19–37.
- [475] P. Spronck, I. Balemans, and G. Van Lankveld, “Player profiling with fallout 3,” in *Artif. Intell. Interactive Dig. Entertainment Conf.*, 2012.
- [476] R. Sifa, A. Drachen, and C. Bauckhage, “Profiling in games: Understanding behavior from telemetry,” *Social interactions in virtual worlds: An interdisciplinary perspective*, 2018.
- [477] A. Drachen, C. Thureau, R. Sifa, and C. Bauckhage, “A comparison of methods for player clustering via behavioral telemetry,” *arXiv preprint arXiv:1407.3950*, 2014.
- [478] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, and M. Colajanni, “Modeling realistic adversarial attacks against network intrusion detection systems,” *ACM Digital Threats: Research and Practice*, 2021.
- [479] D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, and K. Rieck, “Dos and don’ts of machine learning in computer security,” in *USENIX Security*, 2022.
- [480] K. S. Wilson and M. A. Kiy, “Some fundamental cybersecurity concepts,” *IEEE access*, 2014.
- [481] S. Mehnaz et al, “Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models,” in *USENIX Security*, 2022.
- [482] H. Mohajeri Moghaddam, G. Acar, B. Burgess, A. Mathur, D. Y. Huang, N. Feamster, E. W. Felten, P. Mittal, and A. Narayanan, “Watching you watch: The tracking ecosystem of over-the-top tv streaming devices,” in *Proc. ACM Conf. Comp. Commun. Secur.*, 2019.
- [483] B. Schneier, *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company, 2015.
- [484] M. Fryling, J. L. Cotler, J. Rivituso, L. Mathews, and S. Pratico, “Cyberbullying or normal game play? impact of age, gender, and experience on cyberbullying in multi-player online gaming environments: Perceptions from one gaming forum,” *J. Inf. Syst. Appl. Res.*, 2015.
- [485] O. Richman, “Hashinshin responds to accusations of grooming a minor,” <https://win.gg/news/hashinshin-in-responds-to-accusations-of-grooming-a-minor/>, 2020, accessed: June 2022.
- [486] B. News, “Fortnite predator ’groomed children on voice chat’,” <https://www.bbc.com/news/technology-46923789>, 2019, accessed: June 2022.
- [487] S. Zhang, H. Yin, T. Chen, Z. Huang, L. Cui, and X. Zhang, “Graph embedding for recommendation against attribute inference attacks,” in *Proc. Web Conf.*, 2021.
- [488] J. S. Wiggins, *The five-factor model of personality: Theoretical perspectives*. Guilford Press, 1996.

- [489] B. Rammstedt and O. P. John, “Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german,” *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.
- [490] J. Kotrlik and C. Higgins, “Organizational research: Determining appropriate sample size in survey research,” *Inf. Tech. Learn. Perf. J.*, 2001.
- [491] S. Bunian, A. Canossa, R. Colvin, and M. S. El-Nasr, “Modeling individual differences in game behavior using hmm,” in *Artif. Intell. Interact. Digit. Entert. Conf.*, 2017.
- [492] H. Akoglu, “User’s guide to correlation coefficients,” *Turkish journal of emergency medicine*, 2018.
- [493] G. J. Meyer, S. E. Finn, L. D. Eyde, G. G. Kay, K. L. Moreland, R. R. Dies, E. J. Eisman, T. W. Kubiszyn, and G. M. Reed, “Psychological testing and psychological assessment: A review of evidence and issues.” *American psychologist*, vol. 56, no. 2, p. 128, 2001.
- [494] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [495] D. L. Wilson, “Asymptotic properties of nearest neighbor rules using edited data,” *IEEE Transactions on Systems, Man, and Cybernetics*, no. 3, pp. 408–421, 1972.
- [496] F. Al Zamal, W. Liu, and D. Ruths, “Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors,” in *Proc. AAAI Int. Conf. Web Social Media*, 2012.
- [497] Q. Fang, J. Sang, C. Xu, and M. S. Hossain, “Relational user attribute inference in social media,” *IEEE T. Multimedia*.
- [498] B. Mei, Y. Xiao, R. Li, H. Li, X. Cheng, and Y. Sun, “Image and attribute based convolutional neural network inference attacks in social networks,” *IEEE T. Netw. Sci. Eng.*
- [499] E. U. A. for Fundamental Rights, “Child participation in research,” <https://fra.europa.eu/en/publication/2019/child-participation-research>, 2014.
- [500] S. Mystakidis, “Metaverse,” *Encyclopedia*, vol. 2, no. 1, pp. 486–497, 2022.
- [501] K. Kamenov, “Immersive experience—the 4th wave in tech: Learning the ropes,” <https://www.linkedin.com/pulse/immersive-experience-4th-wave-tech-learning-ropes-kamen-kamenov/>, 2017, accessed: Sep 2022.
- [502] M. R. Miller, F. Herrera, H. Jun, J. A. Landay, and J. N. Bailenson, “Personal identifiability of user tracking data during observation of 360-degree vr video,” *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020.
- [503] K. Pfeuffer, M. J. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt, “Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.
- [504] C. E. Rogers, A. W. Witt, A. D. Solomon, and K. K. Venkatasubramanian, “An approach for user identification for head-mounted displays,” in *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, 2015, pp. 143–146.

- [505] S. Li, A. Ashok, Y. Zhang, C. Xu, J. Lindqvist, and M. Gruteser, “Whose move is it anyway? authenticating smart wearable devices using unique head movement patterns,” in *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2016, pp. 1–9.
- [506] T. Mustafa, R. Matovu, A. Serwadda, and N. Muirhead, “Unsure how to authenticate on your vr headset? come on, use your head!” in *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, 2018, pp. 23–30.
- [507] J. Steil, I. Hagedstedt, M. X. Huang, and A. Bulling, “Privacy-aware eye tracking using differential privacy,” in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–9.
- [508] J. Liebers, M. Abdelaziz, L. Mecke, A. Saad, J. Auda, U. Gruenefeld, F. Alt, and S. Schneegass, “Understanding user identification in virtual reality through behavioral biometrics and the effect of body normalization,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–11.
- [509] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, “Augmented reality: A class of displays on the reality-virtuality continuum,” in *Telemanipulator and telepresence technologies*, vol. 2351. Spie, 1995, pp. 282–292.
- [510] F. Nenna, M. Zorzi, and L. Gamberini, “Augmented reality as a research tool: Investigating cognitive-motor dual-task during outdoor navigation,” *International Journal of Human-Computer Studies*, vol. 152, p. 102644, 2021.
- [511] F. Nenna, V. Orso, D. Zanardi, and L. Gamberini, “The virtualization of human–robot interactions: a user-centric workload assessment,” *Virtual Reality*, pp. 1–19, 2022.
- [512] A. Berni and Y. Borgianni, “Applications of virtual reality in engineering and product design: Why, what, how, when and where,” *Electronics*, vol. 9, no. 7, p. 1064, 2020.
- [513] A. A. Malik, T. Masood, and A. Bilberg, “Virtual reality in manufacturing: immersive and collaborative artificial-reality in design of human-robot workspace,” *International Journal of Computer Integrated Manufacturing*, vol. 33, no. 1, pp. 22–37, 2020.
- [514] C. Linn, S. Bender, J. Prosser, K. Schmitt, and D. Werth, “Virtual remote inspection—a new concept for virtual reality enhanced real-time maintenance,” in *2017 23rd International Conference on Virtual System & Multimedia (VSMM)*. IEEE, 2017, pp. 1–6.
- [515] J. Xiao, P. Wang, H. Lu, and H. Zhang, “A three-dimensional mapping and virtual reality-based human–robot interaction for collaborative space exploration,” *International Journal of Advanced Robotic Systems*, vol. 17, no. 3, 2020.
- [516] F. G. Praticò and F. Lamberti, “Towards the adoption of virtual reality training systems for the self-tuition of industrial robot operators: A case study at kuka,” *Computers in Industry*, vol. 129, p. 103446, 2021.
- [517] J. J. Roldán, E. Crespo, A. Martín-Barrio, E. Peña-Tapia, and A. Barrientos, “A training system for industry 4.0 operators in complex assemblies based on virtual reality and process mining,” *Robotics and computer-integrated manufacturing*, vol. 59, pp. 305–316, 2019.

- [518] Q. Liu, Y. Wang, Q. Tang, and Z. Liu, "Do you feel the same as i do? differences in virtual reality technology experience and acceptance between elderly adults and college students," *Frontiers in Psychology*, vol. 11, p. 573673, 2020.
- [519] B. Chavez and S. Bayona, "Virtual reality in the learning process," in *World conference on information systems and technologies*. Springer, 2018, pp. 1345–1356.
- [520] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," *Computers & Education*, vol. 147, p. 103778, 2020.
- [521] S. Z. A. Ansari, V. K. Shukla, K. Saxena, and B. Filomeno, "Implementing virtual reality in entertainment industry," in *Cyber Intelligence and Information Retrieval*. Springer, 2022, pp. 561–570.
- [522] K. Vasista, "Augmented reality vs. virtual reality," *Central asian journal of mathematical theory and computer sciences*, vol. 3, no. 3, pp. 1–4, 2022.
- [523] M. A. M. Abdelmaged, "Implementation of virtual reality in healthcare, entertainment, tourism, education, and retail sectors," *Dissertation*, 2021.
- [524] S. Sharma and A. Bumb, "Product placement in entertainment industry: a systematic review," *Quarterly Review of Film and Video*, vol. 39, no. 1, pp. 103–119, 2022.
- [525] A. W. K. Yeung, A. Tosevska, E. Klager, F. Eibensteiner, D. Laxar, J. Stoyanov, M. Glisic, S. Zeiner, S. T. Kulnik, R. Crutzen *et al.*, "Virtual and augmented reality applications in medicine: analysis of the scientific literature," *Journal of medical internet research*, vol. 23, no. 2, p. e25499, 2021.
- [526] M. Birlo, P. E. Edwards, M. Clarkson, and D. Stoyanov, "Utility of optical see-through head mounted displays in augmented reality-assisted surgery: a systematic review," *Medical Image Analysis*, p. 102361, 2022.
- [527] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario, "Medical robotics and computer-integrated surgery," in *Springer handbook of robotics*. Springer, 2016, pp. 1657–1684.
- [528] M. G. Maggio and R. S. Calabrò, "Virtual reality and cognitive rehabilitation after traumatic brain injury," in *Diagnosis and Treatment of Traumatic Brain Injury*. Elsevier, 2022, pp. 497–506.
- [529] N. M. van der Kolk, N. M. de Vries, R. P. Kessels, H. Joosten, A. H. Zwinderman, B. Post, and B. R. Bloem, "Effectiveness of home-based and remotely supervised aerobic exercise in parkinson's disease: a double-blind, randomised controlled trial," *The Lancet Neurology*, vol. 18, no. 11, pp. 998–1008, 2019.
- [530] D. Kim and Y. Choi, "Applications of smart glasses in applied sciences: A systematic review," *Applied Sciences*, vol. 11, no. 11, p. 4956, 2021.
- [531] Y. Zhao, E. Kupferstein, B. V. Castro, S. Feiner, and S. Azenkot, "Designing ar visualizations to facilitate stair navigation for people with low vision," in *Proceedings of the 32nd annual ACM symposium on user interface software and technology*, 2019, pp. 387–402.
- [532] M. Z. Iqbal and A. G. Campbell, "Adopting smart glasses responsibly: potential benefits, ethical, and privacy concerns with ray-ban stories," *AI and Ethics*, pp. 1–3, 2022.

- [533] IBM, “Big data analytics,” <https://www.ibm.com/analytics/big-data-analytics>, 2022, accessed: Sep 2022.
- [534] P. R. Center, “Americans and privacy: Concerned, confused and feeling lack of control over their personal information,” <https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/>, 2019, accessed: Sep 2022.
- [535] E. Ozturk, “Privacy in emerging technologies,” Ph.D. dissertation, UC Irvine, 2021.
- [536] I. consulting, “General data protection regulation,” <https://gdpr-info.eu/>, 2018, accessed: Sep 2022.
- [537] D. Adams, A. Bah, C. Barwulor, N. Musaby, K. Pitkin, and E. M. Redmiles, “Ethics emerging: the story of privacy and security perceptions in virtual reality,” in *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, 2018, pp. 427–442.
- [538] M. Abraham, P. Saeghe, M. McGill, and M. Khamis, “Implications of xr on privacy, security and behaviour: Insights from experts,” in *Nordic Human-Computer Interaction Conference*, 2022, pp. 1–12.
- [539] R. Di Pietro and S. Cresci, “Metaverse: Security and privacy issues,” in *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications*, 2021, pp. 281–288.
- [540] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T. H. Luan, and X. Shen, “A survey on metaverse: Fundamentals, security, and privacy,” *IEEE Communications Surveys & Tutorials*, 2022.
- [541] V. Nair, G. M. Garrido, and D. Song, “Going incognito in the metaverse,” *arXiv preprint arXiv:2208.05604*, 2022.
- [542] M. Barni, T. Bianchi, D. Catalano, M. Di Raimondo, R. Donida Labati, P. Failla, D. Fiore, R. Lazzeretti, V. Piuri, F. Scotti *et al.*, “Privacy-preserving fingercode authentication,” in *Proceedings of the 12th ACM workshop on Multimedia and security*, 2010, pp. 231–240.
- [543] D. J. Liebling and S. Preibusch, “Privacy considerations for a pervasive eye tracking world,” in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 2014, pp. 1169–1177.
- [544] S. Berkovsky, R. Taib, I. Koprinska, E. Wang, Y. Zeng, J. Li, and S. Kleitman, “Detecting personality traits using eye-tracking data,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.
- [545] G. Fernández, F. Manes, L. E. Politi, D. Orozco, M. Schumacher, L. Castro, O. Agamennoni, and N. P. Rotstein, “Patients with mild alzheimer’s disease fail when using their working memory: evidence from the eye tracking technique,” *Journal of Alzheimer’s Disease*, vol. 50, no. 3, pp. 827–838, 2016.
- [546] T. Kinnunen, F. Sedlak, and R. Bednarik, “Towards task-independent person authentication using eye movement signals,” in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 2010, pp. 187–190.
- [547] J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling, “Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features,” in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–10.

- [548] E. Bozkir, O. Günlü, W. Fuhl, R. F. Schaefer, and E. Kasneci, “Differential privacy for eye tracking with temporal correlations,” *Plos one*, vol. 16, no. 8, p. e0255979, 2021.
- [549] B. David-John, K. Butler, and E. Jain, “For your eyes only: Privacy-preserving eye-tracking datasets,” in *2022 Symposium on Eye Tracking Research and Applications*, 2022, pp. 1–6.
- [550] J. L. Kröger, O. H.-M. Lutz, and F. Müller, “What does your gaze reveal about you? on the privacy implications of eye tracking,” in *IFIP International Summer School on Privacy and Identity Management*. Springer, 2019, pp. 226–241.
- [551] A. Nadeem, D. Vos, C. Cao, L. Pajola, S. Dieck, R. Baumgartner, and S. Verwer, “Sok: Explainable machine learning for computer security applications,” *arXiv preprint arXiv:2208.10605*, 2022.
- [552] F. Nenna and L. Gamberini, “The influence of gaming experience, gender and other individual factors on robot teleoperations in vr,” in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, 2022, pp. 945–949.
- [553] A. F. Kramer, “Physiological metrics of mental workload: A review of recent progress,” *Multiple-task performance*, pp. 279–328, 2020.
- [554] S. Mathôt, “Pupillometry: Psychology, physiology, and function,” *Journal of Cognition*, vol. 1, no. 1, 2018.
- [555] G. Bargary, J. M. Bosten, P. T. Goodbourn, A. J. Lawrance-Owen, R. E. Hogg, and J. Mollon, “Individual differences in human eye movements: An oculomotor signature?” *Vision research*, vol. 141, pp. 157–169, 2017.
- [556] C. Fawcett, E. Nordenswan, S. Yrttiaho, T. Häikiö, R. Korja, L. Karlsson, H. Karlsson, and E.-L. Kataja, “Individual differences in pupil dilation to others’ emotional and neutral eyes with varying pupil sizes,” *Cognition and Emotion*, pp. 1–15, 2022.
- [557] S. Aminihajibashi, T. Hagen, M. D. Foldal, B. Laeng, and T. Espeseth, “Individual differences in resting-state pupil size: Evidence for association between working memory capacity and pupil size variability,” *International Journal of Psychophysiology*, vol. 140, pp. 1–7, 2019.
- [558] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [559] A. Grabowski, J. Jankowski, and M. Wodzyński, “Teleoperated mobile robot with two arms: the influence of a human-machine interface, vr training and operator age,” *International Journal of Human-Computer Studies*, vol. 156, p. 102707, 2021.
- [560] K. Z. Li, R. T. Krampe, and A. Bondar, “An ecological approach to studying aging and dual-task performance,” *Cognitive limitations in aging and psychopathology*, pp. 190–218, 2005.
- [561] R. Katiyar, V. K. Pathak, and K. Arya, “A study on existing gait biometrics approaches and challenges,” *International Journal of Computer Science Issues (IJCSI)*, vol. 10, no. 1, p. 135, 2013.
- [562] B. A. Sargezeh, N. Tavakoli, and M. R. Daliri, “Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study,” *Physiology & behavior*, vol. 206, pp. 43–50, 2019.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Prof. Mauro Conti, for his support, guidance, and mentorship throughout my doctoral journey. His expertise and dedication were truly inspiring, and he offered me countless opportunities to grow as a researcher and as a person. Thanks to his support, I had the opportunity to meet and collaborate with exceptional people in the SPRITZ group and outstanding professors worldwide, including V.S. Subrahmanian, Giovanni Apruzzese, and Gene Tsudik. I want to express my sincere appreciation to all of them for everything they taught me and the incredible experiences they made possible. A particular thanks to Elena, for her kindness and incredible help to make these experiences real.

Special thanks go to my lab mates, who made this challenging journey far more enjoyable. In particular, Luca and Federico for being great friends and especially worthy rivals during the breaks; Matteo for his real-life hacks and immense wisdom; Stefano for knowing Linux better than the creator; Denis and Gabriele for keeping me on track; Tommaso and Francesco for the nerdy stuff; Prof. Losiouk and Prof. Brighente for being tireless and inspiring; and last but not least, Prof. Pasa for teaching us the concept of “Bella Vita”.

A warm thanks to my other Ph.D. colleagues Guglielmo, Federica, Veronica, Giulia, Silvia, and Ying, with whom we shared joys and sorrows, but mostly spritz to forget the sorrows.

A special thought goes to my friends at Northwestern University, Saurabh and Tonmoay, for the moments shared in the lab, and Lirika, Federico, Atmn, and Dennis for the incredible memories we have together. You made my period abroad unforgettable!

I also want to thank all my friends who sustained me during this long and difficult journey. My best friend, Simeone, for always being ready to support me, for sharing passions and travels, for his wisdom, and for the great things we will do together. My “maestro”, food mentor, and best friend Antonio, from whom I will never stop learning. The Sanca gang Alberto, Ila, Davide, Ida, Pietro, Fede, and Fra, for always being there and brightening my days. My Dota 2 squad Joe, Pain, Comis, and Mene, always ready to lose a game together and add more stress. My fellow nerd Anthony, my BBQ mates Giglio, Silvio, and Guab, my piano teacher Nicola, my Erasmus mate Eddy, the lifelong friend Nicola, and all my Sicilian friends, particularly Donia, Emanuele, Federico, Gabriele, Domenico, and Sonica.

My warmest gratitude goes to my family. First and foremost, my parents, who always supported, encouraged, and motivated me to achieve my goals and be the best version of myself. I know I can always count on them, which is incredibly precious. To my sister, kind and generous, whom I know I can rely on. To my cousins Alessia and Roberto, we will never stop having fun together. To my grandparents, for their love and affection, who helped me grow and taught me so much. I wish you were here to celebrate this important milestone with me.

Lastly, my immense gratitude goes to my girlfriend, Susanna, who fills my days with joy and love. You have been my rock during challenging times and my partner in my best life’s adventures. You always encourage me to follow my dreams, even if it involves us making sacrifices, but we both know it is for a greater purpose. Though far away, we are always close, and your presence is my greatest comfort. Your way of seeing the world, totally different

from mine, has really helped me grow. Even when I'm down, your little head (TDC) makes me smile and keep moving forward. I am endlessly grateful for every moment we share together, and I look forward to a future filled with more love, laughter, and precious memories.