UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Dipartimento di
Economia e Management

CORSO DI DOTTORATO DI RICERCA IN: Economia e Management
CURRICOLO: Economics
CICLO XXIX°

TITOLO TESI

The Information Content of Financial Textual Data: Creating News Measures for Volatility Modeling and for the Analysis of Price Jumps

Coordinatore: Ch.mo Prof. Antonio Nicolò
Supervisore: Ch.mo Prof. Massimiliano Caporin

Dottorando : Francesco Poli

The Information Content of Financial Textual Data: Creating News Measures for Volatility Modeling and for the Analysis of Price Jumps

Contents

Introduction

In the first chapter, we retrieve news stories and earnings announcements of the S&P 100 constituents from two professional news providers, along with ten macroeconomic indicators. We also gather data from Google Trends about these firms' assets as an index of retail investors' attention. Thus, we create an extensive and innovative database that contains precise information with which to analyze the link between news and asset price dynamics. We detect the sentiment of news stories using a dictionary of sentiment-related words and negations and propose a set of more than five thousand information-based variables that provide natural proxies for the information used by heterogeneous market players. We first shed light on the impact of information measures on daily realized volatility and select them by penalized regression. Then we use these measures to forecast volatility and obtain superior results with respect to the results of models that omit them.

In the second chapter, we detect intraday price jumps in the S&P 100 constituents' stocks. Then, we build high frequency news indicators from news stories released by two professional news providers, earnings announcements, and twenty-three US macroeconomic indicators. We investigate the extent to which statistically significant intraday jumps are associated with the news indicators and select them by penalized logistic regression. Finally, we compare the economic significance of jumps. We find effects on returns and volatility at both high frequency and daily level, and that these effects vary depending on the type of news to which jumps are associated. We also find that future quarterly and yearly returns seem to be exposed to jump risk measures built using jumps related to macro-announcements.

A common method to detect the sentiment of a text is the so-called bag-of-words approach. In the third chapter, we extend the method in three directions, by using: 1) an extended negations list of single words, two-word sequences, and three-word sequences; 2) lists of sentiment-related expressions (e.g., "high quality"); 3) lists of sentiment-related words combinations (e.g., "increase" and "dividend"). The aim is creating a general method suitable for detecting the sentiment of a financial text of any type.

Chapter 1

News Measures from Textual Data and an Application to Volatility Forecasting

Massimiliano Caporin and Francesco Poli

1.1 Introduction

Traditional "efficient markets" thinking suggests that asset prices should completely and instantaneously reflect movements in underlying fundamentals, while an opposite view indicates that asset prices and fundamentals are continuously disconnected. One hypothesis that explains the success of the GARCH class of models is the mixture of distributions hypothesis (MDH). (See Clark (1973), Epps and Epps (1976), Tauchen and Pitts (1983), and Lamoureux and Lastrapes (1990), among many others.) According to the MDH, a serially correlated mixing variable that measures the rate at which information arrives to the market explains the GARCH effects on asset returns. The validity of the MDH remains an open debate; there is no agreement about how quickly and in what form responses to news occur. We shed light on the link between news information and volatility, focusing on three questions: What is the relative importance of types of news? Are investors more influenced by the volume of information or by variations if it? Do news and an index of investors' attention help to forecast volatility?

Our first contribution is to create an extensive and innovative database that contains information useful in answering the three questions. From two news providers, Factset-StreetAccount and Thomson Reuters-Thomson One, we retrieve news stories and earnings announcements of the S&P 100 constituents, along with ten macroeconomic announcements. Both news providers assign news stories a topic—Thomson Reuters also gives its news stories a level of importance—while earnings and macro-announcements report both released figures and consensus forecasts, allowing diversions from expectations to be computed. In addition, we gather Google Trends[1] information about the assets and use them as a proxy for retail investors' attention. Google restricts access to daily data for intervals longer than ten months but allows daily data to be gathered for shorter intervals. Exploiting the series of daily data associated with each month and the series of monthly data associated with the whole sample, we reconstruct the daily relative search volume for the whole sample. The collected news reports are dated with to-the-minute precision, while Google Trends are aggregated by day, so the dataset has to-the-minute precision for news and daily precision for Google Trends. The sample contains data for the ten-year period from February 2005 to February 2015.

As a second contribution, we detect the sentiment of news stories using the sentiment-related word lists developed by Loughran and McDonald (2011) and introduce a set of negations, both with the aim of creating a method that can be used to extract the sentiment of a financial text with more confidence and independent of its type, length, and audience.

Our third contribution is to propose a set of news measures that provide natural proxies for retail investors' attention and for the information heterogeneous market players use. This study goes beyond how information has been used so far: starting from the reasoning that investors' perception and, as a consequence, their reaction to news disclosures can differ based on how information varies over time and the reasoning that investors digest and react to news at differing speeds, we look at how the information stream fluctuates over the day and across days, weeks, and months. We end with a large set of news measures, each representing a different type of information that can cause a different market reaction.

As final contribution, we shed light on the impact of news on volatility and address the three questions posed above using the information-related variables we develop. We perform an application using the database

[1] Google Trends is a public web facility of Google Inc. based on Google Search that shows how often a particular search-term is entered relative to the total search volume across various regions of the world and in various languages.

to explain realized volatility and selecting the most important indicators with LASSO (least absolute shrinkage and selection operator), an estimation method for linear models that is commonly employed in big data analysis which performs variable selection and shrinks coefficients. Then we employ news and Google Trends to forecast volatility in an out-of-sample analysis.

Empirical analyses favor the MDH and show that macroeconomic news and earnings announcements are the most important drivers of daily realized volatility, followed by news stories and Google Trends, and that earnings and upgrades/downgrades are the topics of news stories that are most relevant to explaining volatility. In addition, the analyses show that it is important both to look at variations of the volume of information across time and to build measures based on the aggregation of information over various time horizons since the measures imply varying reactions from market players. By including news-based information, we can improve volatility forecasting substantially.

1.2 Literature Review

1.2.1 Mixture of Distributions Hypothesis

The MDH is a classic topic in the finance literature. Clark (1973), Epps and Epps (1976), and Tauchen and Pitts (1983) use different approaches to test the relationship between returns variance and trading volume for the same interval of time. Lamoureux and Lastrapes (1990) show that trading volume has been used as a proxy for information arrival, that they have significant explanatory power regarding the variance of the daily returns, and that ARCH effects tend to disappear when volume is included in the variance equation. More recently, Kalev et al. (2004) find results that are consistent with the MDH by employing firm-specific announcements and lagged volume as a proxy for information flows and investigating the information-volatility relationship using high-frequency data from the Australian Stock Exchange. Martens et al. (2009) evaluate the forecasting performance of time series models for realized volatility and account for the effects of macroeconomic news announcements, arguing that allowing volatility to differ on days that contain news releases can disentangle calendar and announcement effects. McMillan and García (2013) forecast intra-day volatility for the IBEX 35 Index futures using volume and the number of transactions as proxies for information flows and show that introducing the proxy improves the volatility forecast for several volatility models at various frequencies. Zhang et al. (2014) employ the number of news stories that appeared in *Baidu News*[2] as a proxy for information arrival and use a sample of SME Price Index[3] in China to validate the MDH. Their empirical results reveal a positive impact of internet information on the conditional volatility of stock returns. This link has also been documented for the US stock market (Kim and Kon (1994); Gallo and Pacini (2000)), the UK stock market (Omran and McKenzie (2000)), and the Australian stock market (Brailsford (1996)).

More generally with regard to the relationship between news flows and asset price dynamics, the last few decades of research have produced a tremendous number of empirical studies, but these studies have by no means reached consensus. While some of these papers focus on the impact of macroeconomic news, others explore the idea that assets react to firm-specific news releases.

1.2.2 Macroeconomic News

The finance literature began analyzing the relationship between news and market movements with studies like Cutler et al. (1989), who report a faint relationship among macroeconomic news, world political events, and stock market activity, and Schwert (1989), who finds weak evidence that macroeconomic volatility can explain stock return volatility. A stream of the literature addresses the volatility reaction to news released on announcement days, focusing on the dynamics of conditional volatility based on the ARCH/GARCH framework introduced by Engle (1982) and Bollerslev (1986). For example, Li and Engle (1998) compare the degree of persistence associated with scheduled macroeconomic announcement days and non-announcement days in the Treasury futures market and find heterogeneous persistence. Jones et al. (1998) present a similar analysis for the Treasury bond market, and show U-shaped day-of-the-week effects and calm-before-the-storm effects for bond returns' volatility. In contrast to Li and Engle (1998), Jones et al. (1998) find that announcement-day shocks do not persist at all, as they are purely transitory. Flannery and Protopapadakis (2002) also use a GARCH model to detect a consistent influence of monetary and macroeconomic variables on stock market indices. Bomfim (2003), whose work is based on Jones et al.'s (1998) framework, examines the effect of monetary policy announcements on the volatility of stock returns, finding that unexpected monetary policy decisions tend to boost volatility significantly in the short run. Using conditional variance modelling, Janssen (2004) demonstrates an intertemporal relationship between the arrival of public information (measured as the

[2]Baidu News is a service of the Chinese web services company Baidu. Baidu News provides links to a selection of local, national, and international news, and presents news stories in a searchable format within minutes of their publication on the web.

[3]the SME Price Index functions as the market indicator of China's small and medium-size enterprises listed on the SME Board.

daily number of economic news headlines) and volatility persistence of US stocks, Treasury bills, bonds, and the dollar. Engle and Rangel (2008) find a strong relationship between the low-frequency component of market volatility, represented by an exponential spline, and macroeconomic variables like inflation, growth, and macroeconomic volatility. Vrugt (2009) analyzes the impact of US and Japanese macroeconomic news on stock market volatility in Japan, Hong Kong, South Korea, and Australia and employ a GARCH model that allows for multiplicative announcement effects and asymmetries to find that overnight conditional variances are higher on announcement days than on days before and after announcements, especially for US news, while the impact of announcements on implied volatilities is weak. Brenner et al. (2009) find that US macroeconomic information drives the level, volatility, and co-movement of the US stock, Treasury, and corporate bond markets. Hautsch et al. (2011) find that the arrival of macroeconomic news has an impact on the bid and ask dynamics of the German Bund futures. Rangel (2011) examines the effect of macroeconomic releases on stock market volatility through a Poisson-Gaussian-GARCH process with time-varying jump intensity, which is allowed to respond to information, and finds that macroeconomic surprises impact both volatility and jump intensity. Birz and Lott (2011) choose newspaper stories about GDP and unemployment as a measure of news and find that macroeconomic news affects S&P 500 returns. Savor and Wilson (2013) document higher average excess market returns on days with important macroeconomic news releases compared to non-announcement days.

1.2.3 Firm-Specific News and Sentiment

With regard to firm-specific news, Mitchell and Mulherin (1994), Berry and Howe (1994), and Roll (1988) are among the first to report a weak relationship between stock market activity and news. Kalev et al. (2004) document a positive relationship between the number of intraday news articles and the Australian stock market volatility, and in another intraday study Busse and Green (2002) consider the impact of news released via television on more than 300 stocks to test market efficiency. However, both studies show that the impact of news on intraday trading activity is weak, that it disappears altogether if earnings announcements are discarded, and that news stories have to be aggregated to reduce the influence of noisy and non-informative news. Fang and Peress (2009) explore news coverage and predictability of returns and find that stocks with no media coverage earn higher returns than do stocks with high media coverage. Baklaci et al. (2011) explore the relationship between intraday firm-specific news announcements and return volatility in the Turkish stock market and find that the persistence of volatility diminishes with the inclusion of news, suggesting that news is rapidly incorporated into prices.

Several studies investigate the relationship between news sentiment and changes in asset price dynamics. Antweiler and Frank (2004) are the first to develop news sentiment measures to explain stock returns. Using a Naive Bayes algorithm based on the number of times certain words occur, they infer trading signals from posts on internet message boards and find that, while such signals can predict market volatility, their effect on stock returns is small. Zhang et al. (2012) incorporate several methodological improvements and create news sentiment indices that are significant directional indicators. Tetlock (2007) undertakes the so-called bag-of-words approach, which has become widespread in the literature. This approach consists of building lists (bags) of words and associating each list with a category (e.g., positive or negative). Classifying words based on categories from the *Harvard Psychosocial Dictionary*, Tetlock quantifies optimism and pessimism from the *Wall Street Journal*'s "Abreast of the Market" column and reports that high levels of media pessimism predict declining market prices, which are followed by price reversals. Using a similar technique, Tetlock et al. (2008) use the *Harvard IV-4 Psychological Dictionary* and find that the fraction of negative words in Dow Jones News Service and Wall Street Journal stories forecasts firm earnings because the linguistic content of news messages captures the hard-to-quantify aspects of fundamentals that are quickly incorporated into stock prices. Thanks to recent advances in technology, software packages like *Reuters NewsScope Sentiment Engine*, the more recent *Thomson Reuters News Analytics*[4], and *Ravenpack News Analytics*[5] have been developed. These packages use advanced algorithms and assign sentiment indicators to firm-specific newswire releases, enabling investors who pay for the service to employ "real-time" trading signals from textual analysis in quantitative trading strategies. Gloß-Klußmann and Hautsch (2011) employ the trading signals from *Reuters NewsScope Sentiment Engine* to find that high-frequency responses in market activity and volatility are significant, especially after the release of intraday company-specific news, and that classifying news according to relevance helps to filter noise and identify significant effects. Using sentiment scores generated at high frequency by *RavenPack News Analytics*, Ho et al. (2013) find a significant impact of firm-specific news sentiment on intraday volatility persistence, even after controlling for the potential effects of macroeconomic news. Firm-specific news sentiment apparently accounts for a greater proportion of overall volatility persistence than macroeconomic news sentiment does, and negative

[4]Reuters NewsScope Sentiment Engine and Thomson Reuters News Analytics are tools that provide sentiment and linguistic analytics, such as novelty and relevance indicators, for each news article. The indicators are produced based on automated linguistic pattern recognition of news texts.

[5]RavenPack News Analytics is a service of RavenPack.com, a provider of news analytics and machine-readable content, that provides event and sentiment information to financial services clients.

news has a greater impact on volatility than positive news does. Riordan et al. (2013) suggest that negative newswire messages from *Reuters NewsScope Sentiment Engine*, compared to positive ones, are associated with higher adverse selection costs, are more informative, and have a more significant impact on high-frequency asset price discovery and liquidity. Smales (2015) use *Thomson Reuters News Analytics* sentiment scores to create aggregate daily news sentiment indicators and find that positive and negative news result in above and below average returns, respectively, and that neutral news days are indistinguishable from days without news. In the field of bag-of-words methods in financial contexts, Loughran and McDonald (2011) show that word lists developed for other disciplines misclassify common words in financial texts and develop alternative positive and negative word lists and four other word lists that reflect tone in financial texts. They show that the proportion of negative words in annual 10-Ks reports[6] is associated with lower returns.

Differently from previous studies that use or focus on only macroeconomic or firm-specific information, Bajgrowicz et al. (2016) consider macro, pre-scheduled company-specific announcements and stories from news agencies like Reuters and Dow Jones News Service and relate them to jumps in the US stock market.

1.2.4 Google Trends

Quantitative data on internet use will soon be an invaluable source for economic analysis since they capture investors' attention and information demand. Ginsberg et al. (2009), in the first article to use Google data, estimate the weekly influenza activity in the US using an index of health-seeking behavior that is equal to the incidence of influenza-related internet queries. Since then, the use of internet search data has been extended rapidly in estimating economic variables. For instance, Baker and Fradkin (2011) develop a job-search activity index to analyze the reaction of job-search intensity to changes in the duration of unemployment benefits in the US, and D'Amuri and Marcucci (2012) suggest that the Google index (GI), based on internet job searches performed through Google, is the best leading indicator of the US monthly unemployment rate. Recent studies have shown that online search activity is also associated with volatility and returns in the financial, commodity, and exchange-rate markets. (See Da et al. (2011) and Vlastakis and Markellos (2012) for individual stocks; Andrei and Hasler (2015), Dimpfl and Jank (2016), and Hamid and Heiden (2015) for stock indexes; Vozlyublennaia (2014) for stock and bond indices, gold, and crude oil; Da et al. (2015) for stock indices, the VIX volatility index, and equity and Treasury bonds mutual funds; Guo and Ji (2013) for crude oil; and Smith (2012) and Goddard et al. (2015) for exchange rates.)

1.3 Database Construction

Our first contribution lies in the extraction of information collected from two news providers, FactSet-StreetAccount and Thomson Reuters-Thomson One, and from Google Trends. Here we describe our novel dataset and the procedures used to extract the data.

1.3.1 Dataset

A large set of firm-specific and macroeconomic news is available from the two news providers FactSet-StreetAccount and Thomson Reuters-Thomson One. As Gloß-Klußmann and Hautsch (2011) point out, recording and analyzing the overall news flow for a specific asset is challenging since the amount of news, the number of news sources, and the speed of information dissemination are all rapidly increasing. Because of the huge amount of information published in all modern media, news are overlaid with substantial noise from irrelevant information. Since we rely on two professional news providers that provide only firm-specific news classified by their professionals as relevant to the firm, we assume that relevant news stories are effectively disentangled from irrelevant ones and that the impact of noise is adequately reduced. Our approach differs substantially in this regard from work that analyzes newspapers articles that are not selected *a priori*.[7]

StreetAccount, owned by the financial data and software company FactSet, is a news provider that supplies investment professionals with news summaries. StreetAccount data includes real-time company updates, portfolio and sector filtering, email alerts, and market summaries. Content can be customized for portfolio, index, sector, market, time of day (e.g., overnight summaries), and category (e.g., top stories, market summaries, economic stories, M&A stories). Writers, all of whom are financial professionals, include former portfolio managers, traders, analysts, and economists who use their collective market expertise to scan all possible sources for corporate news and report only those stories that they consider new and material. Comprehensive U.S. and

[6] A Form 10-K is an annual report required by the U.S. Securities and Exchange Commission (SEC), that gives a comprehensive summary of a company's financial performance.

[7] From the pioneering works of Tetlock (2007) and Tetlock et al. (2008) to the more recent studies of, for example, Birz and Lott (2011), Dougal et al. (2012), García (2013), Solomon et al. (2014), and Kraussl and Mirgorodskaya (2016), authors have employed general economic or company-specific news articles from newspapers or specific sections/columns to explain asset price dynamics but have not made selections based on articles' relevance or novelty.

European company coverage and coverage of a smaller but relevant list of Canadian and Asia Pacific companies extend to thousands of companies. Firm-specific and macroeconomic news are available in StreetAccount.

Thomson Reuters is a world-leading source of information for businesses and professionals, and Thomson One, one of its core products, is a database that provides financial market news from Reuters and leading third-party sources. Thomson One data results from the incorporation of 400 real-time global sources and newswires and more than 6,000 global and regional sources, including *The Economist*, *Barron's*, *Le Monde*, *The Washington Post*, *PR Newswire*, *Business Wire*, and *The Wall Street Journal*. Comprehensive global coverage of 57,000 publicly listed companies spanning more than 120 markets tracked and corresponding to 99 percent of global market capitalization includes the constituents of all major indices and extends to the frontier/emerging markets of Central and Eastern Europe, Asia, the Middle East, and Africa. Firm-specific news is available in a variety of formats, each corresponding to a type of information: *Significant Developments*, *Company Events*, and *Earnings Surprises*. Macroeconomic news is available as well from Thomson One.

Following the growing popularity of the internet as a search tool, the use of such sources as Google to find information on a certain stock seems to be closely linked to stock market participation. (See, e.g., Preis et al. (2010).) However, as Da et al. (2011) point out, Google is likely to be representative of general internet search behavior, so the quantity of queries for a term is a measure for retail investors' activity, rather than for professional investors' activity. Therefore, we use Google Trends' public data as a proxy for retail investors' attention.

We gather news about ten US macroeconomic indicators and firm-specific news and Google Trends for the S&P 100 Index companies since they are highly capitalized and attention-grabbing companies. We excluded from the database: 1) stocks whose news stories were not available from either provider for the period February 2005 – February 2015, and 2) stocks that entered the S&P 100 Index or were created after February 2005. Eleven stocks were excluded, and the remaining eighty-nine stocks are listed in Table A.1 in Appendix A.1.

The information that constitutes the database can be classified into five types:

1. **StreetAccount news stories**: firm-specific news stories released by StreetAccount
2. **Thomson Reuters news stories**: firm-specific news stories released by Thomson Reuters
3. **Earnings announcements**: firm-specific EPS earnings per share (announcement and forecast) released by StreetAccount
4. **Macro-announcements**: ten macroeconomic indicators (announcement and forecast) released by Thomson Reuters
5. **Google Trends**: firm-specific relative indicators of internet search volume available from Google

Firm-specific StreetAccount news stories include trading-floor conjectures, court rulings, FDA and EU drug approvals, FTC antitrust decisions, SEC filings, brokerage firm upgrades and downgrades, newspaper and television stories, stories released by social media, and company press releases, including perspectives, corporate conference calls, and presentations. News are classified along eleven topics, which are listed in Table A.2 in Appendix A.1. News are filtered for relevancy and redundancy so each news story is included only once.

Firm-specific Thomson Reuters news stories are available from Significant Developments, a news analysis, tagging, and filtering service of Thomson One that screens press releases and provides concise summaries and categorizations of important company events on a near real-time basis. Customized reports can be created for a portfolio of companies, regions, industries, and news topics. Each story is organized into one or more of thirty-six topics and is given one of four levels of significance/importance: *low*, *medium*, *high*, and *top*, where each level implies a filter which eliminates all news stories with a lower significance; for instance, *low* corresponds to all news stories, while *medium* corresponds to news stories from *medium* to *top*. The thirty-six topics are listed in Table A.2 in Appendix A.1. Assignment of degree of significance is based on the expected effect that the event will have on the company's operational and/or financial performance. As for StreetAccount news, Thomson Reuters news stories are also filtered for relevance and redundancy. Firm-specific company events are also available from Thomson One and consist of a comprehensive list of current and past events—primarily earnings releases, conference calls, news conferences, and shareholders' meetings. While they are not categorized by topic, they are short descriptions of the events that do not allow sentiment to be extracted. For these reasons we do not use company events to construct news measures, but they are part of our database and are reported here.

Firm-specific earnings announcements incorporate both the company's reported actual EPS and the consensus forecast figure, given as the mean of a set of surveys at the time of reporting, so investors and analysts can determine whether the company has met, exceeded, or fallen short of the street's expectations. Earnings announcements are recovered by StreetAccount news stories that contain the quarterly EPS announcements and their consensus forecast, which we compare to compute earnings surprises. Thomson One reports earnings

surprises too; although the figures are highly reliable since they are computed by the provider, data are available with day precision instead of minute precision and are limited to the period July 2013 – June 2015. As a consequence of these limitations, we do not use Thomson Reuters earnings surprises in this study.

Ten US macroeconomic indicators are available from Thomson One, and they are listed in Table 1.1.

Table 1.1: Macro indicators.

Abbreviation	**Complete Name**
CCONF	Consumer Confidence
CPI	Consumer Price Index
FOMC	FOMC Rate Decisions
GDP	Gross Domestic Product
INDPROD	Industrial Production
BOP	Balance of Payments
JOBLESS	Jobless Claims
NFP	Non-Farm Payrolls
PPI	Producer Price Index
RSALES	Retail Sales

Google Trends summarizes the searches performed through the Google website and shows how many web searches have been done for a particular keyword in a particular period of time in a particular geographical area relative to the total number of web searches performed through Google in the same period and area. Absolute values of the index are not publicly available since Google normalizes the index to 100 in the period in which it reaches the maximum level. Data are gathered using IP addresses only if the number of searches exceeds a certain threshold. Repeated queries from a single IP address within a short time are eliminated. Google Trends have been available almost in real time since the end of January 2004. For each stock, we look at the number of search queries for the name of the company but do not include search queries for the company's products or other related expressions since it is likely that investors search for the company's name when they look for information about it. We also exclude search queries for tickers since, in many cases, they correspond to acronyms for other institutions or have other meanings. Google restricts the access to daily data for intervals longer than ten months but allows daily data (relative to the maximum) to be gathered for shorter intervals. For the ten-year period, we reconstruct the daily search volume for the whole sample from the set of the daily series for each month and the monthly aggregated series for the whole sample, following a procedure detailed in subsection 1.5.3.

The dataset's time range is February 4, 2005, to February 25, 2015, and all data is available with minute precision, except for Google Trends, which are daily.

News is available in various data formats, depending on the provider. Using the software $python^{TM}$, we extracted from each news story a set of elements that depend on the type of news, its data format, and its provider. For StreetAccount and Thomson Reuters news stories we obtain stock, date with minute precision time, headline, topic (also importance for Thomson Reuters news stories), and text. For company events we derive stock, date, time, and event description. For earnings announcements we extrapolate stock, date, time, actual EPS, and consensus forecast EPS. For macro-announcements we isolate type of macro-indicator (e.g., GDP), date, time, actual figure, and consensus forecast. With regard to Google Trends, we collect for each stock the set of the daily series for each month and the monthly aggregated series for the whole sample.

1.3.2 Topics and Importance for News Classification

StreetAccount news stories are classified into six of the eleven available topics—*earnings-related*, *litigation* (court disputes), *M&A*, *newspapers*, *regulatory*, and *upgrades/downgrades*—or *all* (all news, no filter by topic). The other topics were discarded because they lacked in either importance or frequency.[8]

Thomson Reuters news stories are classified by both importance and topic. We use all four levels of importance (*low*, *medium*, *high*, and *top*) and build six topics from the thirty-six available: *all*, *earnings pre-announcements*, *financial*, *litigation*, *M&A*, and *regulatory/company investigation* (events concerning regulatory agencies, internal investigations, and any type of charges brought by regulatory bodies). The *earnings pre-announcements* topic merges three topics: *positive earnings pre-announcements* (higher than expected), *negative earnings pre-announcements* (lower than expected), and *other earnings pre-announcements* (neutral with respect to expectations). The *financial* topic merges *equity issues*, *bond issues*, *share repurchases*, and

[8] *Guidance* news is almost coincident with *earnings-related* news; *conjecture* news describes possible and uncertain events and are presumably perceived as not important; *corporate actions* news is about companies' internal events, which usually have minor relevance to investors; *management changes* and *syndicate* news stories are rare and even non-existent for some stocks.

equity investments, all of which are events that have an impact on the company's balance sheet. The other topics were discarded for reasons similar to those that led to our discarding some of StreetAccount news' topics. By jointly exploiting topics and importance, we get (n. importance) x (n. topics) = 4 x 6 = 24 classifications for Thomson Reuters news stories.

Topics from different providers that appear to have the same meaning usually have similarities but can also differ significantly because they depend on the criteria the analysts use for news categorization, and topics often have different meanings. For instance, StreetAccount's *earnings-related* news is a more broad concept than Thomson Reuters' *earnings pre-announcements*, since the former is comprehensive of earnings pre-announcements released by the company, consensus forecasts released by the provider, and EPS announcements, while the latter consists only of the company's earnings pre-announcements. As another example, StreetAccount's *regulatory* news topic apparently does not include company-internal investigations, unlike Thomson Reuters' *regulatory/company investigation*.

Table 1.2 lists the topics that are included in our dataset for each data provider.

Table 1.2: Selected topics by provider

StreetAccount	**Thomson Reuters**
all	all
earnings related	earnings pre-announcements
M&A	M&A
litigation	litigation
regulatory	regulatory/company investigation
newspapers	financial
up/downgrades	

1.3.3 News stories' Summary Stats and Provider Comparison

Table 1.3 presents the summary statistics of the basic variables *number of news stories per day* and *number of words per day*; for StreetAccount's news stories the *all* topic; and for Thomson Reuters news stories the *all* topic with *low* importance. (*Low* importance means that there is no filter; that is, all news categorized from *low* to *top* is included.) The tables report the cross-sectional median of min, 5% quantile, median, 95% quantile, max, mean, standard deviation for each measure. StreetAccount releases more news than Thomson Reuters on average—more than one news story every five days versus one news story every ten days. In addition, StreetAccount news stories are longer than those of Thomson Reuters: more than eighteen words per day versus fewer than nine.

Table 1.3: Summary statistics of news stories from the two providers.

Measure	**Min**	**Quant 5 %**	**Median**	**Quant 95 %**	**Max**	**Mean**	**Std Dev**
SA n. news stories per day	0.00	0.00	0.00	1.00	7.00	0.23	0.60
TR n. news stories per day	0.00	0.00	0.00	1.00	3.00	0.10	0.33
SA n. words per day	0.00	0.00	0.00	102.00	1079.00	18.18	66.68
TR n. words per day	0.00	0.00	0.00	68.90	390.00	8.23	31.80

Notes: Summary statistics of Street Account news stories topic *all* and Thomson Reuters news stories topic *all*, importance *low*.

In order to understand to what degree the information supplied by the two providers is similar, we compute, for a series of topics (pooling all stocks), the ratio between the number of days in which both providers report at least one news story and the number of days in which at least one provider reports at least one news story, naming the result *Coincident/Total Ratio*. The higher the ratio, the greater the similarity of the information released by the two providers. We compare the topics *all* (no filter), *earnings* (StreetAccount's *earnings-related* vs. Thomson Reuters' *earnings pre-announcements*), *litigation*, *M&A*, and *regulatory* (StreetAccount's *regulatory* vs. Thomson Reuters' *regulatory/company investigation*). Table 1.4 reports the Coincident/Total Ratio, the percentage of days with at least one news release by StreetAccount, and the percentage of days with at least one news release by Thomson Reuters. Even when the news occurrence is aggregated on a daily level, it is clear that StreetAccount and Thomson Reuters supply different information. Therefore, we use news stories released by both providers in the rest of this study.

Table 1.4: Coincident/Total Ratio, percentage of StreetAccount news days, percentage of Thomson Reuters news days.

Topic	Coincident/Total Ratio	% SA News Days	% TR News Days
all	26.27	19.95	13.14
earnings	34.72	3.55	1.46
litigation	13.38	0.86	0.84
M&A	22.83	2.25	1.65
regulatory	3.90	1.27	0.28

1.4 Sentiment Detection

Our second contribution consists of detecting the sentiment of news stories. *Sentiment* indicates whether the content of a document—in our case, a news story—is good, bad, or neutral in relation to the issue it addresses. We use the sentiment-related word lists developed by Loughran and McDonald (2011) and introduce a set of negations, with the aim of creating a method for extracting the sentiment of a financial text with more confidence and independent of its type, length, and audience.

Loughran and McDonald (2011) develop six word lists (*negative*, *positive*, *uncertainty*, *litigious*, *strong modal*, and *weak modal*) and show that a higher proportion of negative words is associated with lower returns. Their lists are tailored for financial texts; for example, they do not contain words like *liability*, *earnings*, and *tax*, which are expected to appear in both positive and negative contexts. The authors account for negation with six words (*no*, *not*, *none*, *neither*, *never*, and *nobody*), but only if they precede words that are classified as positive. The methodology is applied to US companies' 10-K filings and these texts have a formal tone and are unlikely to contain many negations. We deal instead with news created by news providers, which we expect to be less limited in the use of language compared to company filings of 10-Ks. Loughran and McDonald's (2011) procedure is not adequate for extracting the sentiment of news stories from news providers because, unlike 10-Ks that are given to the SEC, news stories do not necessarily have a formal tone; in addition, 10-Ks are long enough that, if negated words occur and their sentiment is incorrectly identified, the effect is negligible in the whole, long document. We deal, instead, with news stories that are seldom longer than a few dozen words.

Negations can appear in various forms and can invert the meaning of whole phrases, as well as single words. The phrase whose meaning is changed is called the negation scope. Negations can also flip the meaning of sentences, as in "the company has invented a new product for the first and last time." Identifying negation scopes, implicit negations, and linguistic peculiarities like sarcasm and irony still presents many problems. Approaches like heuristic rules and machine-learning that perform natural language processing can bring significant improvements, but they are out of the scope of the present work.

Remaining in the field of the bag-of-words and avoiding the numerical complexities implied by the aforementioned approaches, we invert the sentiment each time a word, whether positive or negative, is preceded by a negation; and in place of the short list of six single negative words, we use twenty-eight single words, twenty-four sequences of two words, and six sequences of three words. We believe that this modification allows the sentiment of a financial text to be extracted with more confidence and independent of its type, length, and audience:

- single words: *no, not, none, never, nothing, nobody, nowhere, neither, nor, hardly, scarcely, seldom, barely, few, little, rarely, instead, can't, cannot, don't, doesn't, didn't, mustn't, won't, despite, overly, too, less*
- two-word sequences: *can not, do not, did not, short of, not every, not all, not much, not many, not always, not so, instead of, far from, not to, never to, no way, out of, not very, not enough, too few, too little, no big, not big, no significant, not significant*
- three-word sequences: *not at all, by no means, in no way, in place of, in spite of, in lieu of*

The procedure we develop works as follows:

1. Positive words are given a value of 1 and negative words a value of -1; the value is inverted in case of negation.
2. The values of all words with a sentiment are summed to get the sentiment sum (*Sent_Sum*):

$$Sent\_Sum = \sum_{i=1}^{N} s_i \tag{1.1}$$

where i is the word index, N is the number of words with a sentiment in a text and, s_i is the sentiment of the word indexed by i.

3. *Sent_Sum* is divided by the number of words with a sentiment to obtain a standardized quantity that we call relative sentiment (*Rel_Sent*) and that is between -1 and 1 by construction:

$$Rel\_Sent = \frac{Sent\_Sum}{N} \tag{1.2}$$

4. If *Rel_Sent* is larger than 0.05 or smaller than -0.05, we associate a positive (1) or a negative sentiment (-1) to the news, respectively; otherwise a neutral sentiment (0) is given:

$$Text\_Sent = \begin{cases} -1 & \text{if} \quad Rel\_Sent < -0.05 \\ 0 & \text{if} \quad -0.05 \leq Rel\_Sent \leq 0.05 \\ 1 & \text{if} \quad Rel\_Sent > 0.05 \end{cases} \tag{1.3}$$

Different from the mainstream of text-analysis techniques, which look either only at headlines or only at text, we use both headlines and text by applying the sentiment extraction to the headline. The procedure stops if a positive or negative sentiment is detected; otherwise, the whole text is analyzed. This method is more complete than looking at headlines only while also being more efficient than looking directly at the text since it allows us to use small pieces of text rather than long ones when it is possible to infer sentiment from headlines only.

1.5 Creating News Measures

Our third contribution consists of going beyond the standard techniques to assign numbers to textual information, as we identify a set of concepts/events that are based on how news is released over various time horizons with the aim of identifying the portions of information on which market players base their decisions. In our view, explaining price dynamics using only news measures based on a single time horizon creates an omitted-variable bias.

The following subsections describe the procedures we followed to build news-related variables from the dataset, and consist of: 1) the concepts to be used to build variables from news stories, 2) the standardized surprises obtained from earnings and macro-announcements, 3) the daily Google Search Index reconstruction for the whole sample, and 4) the news measures we propose for a daily analysis of asset price dynamics.

1.5.1 Concepts for Variables Related to News Stories

We built the variables using unique concepts in terms of the reaction the concept may cause in the market. All concepts refer to a reference period and to previous periods of equal or longer length. For instance, if the reference period is day *t*, the variables built depend on the information released during day *t*, day *t-1*, last week, and so on. We consider nine concepts:

- **standard measures**: number of news stories, number of words, sentiment. Number of news stories and number of words are proxies for the amount of information.
- **abnormal quantity**: number of news stories above a certain threshold. Investors' reaction could be triggered by the release of an unusual amount of information.
- **uncertainty**: occurrence of news stories with opposite sentiments during the reference period. Information is released, but investors are likely unable to detect whether it is good or bad.
- **news burst index**: a measure of the amount of information released during the reference period that takes into account the possibility that a sudden, abnormal burst of information can affect market activity differently from the same information released gradually. Developed from the notion of realized volatility of an asset's intraday returns, the news burst index is computed as the sum of the k-th power of the number of news stories (or words) disclosed over a series of time intervals:

$$News\_BI_t(M,k) = \sum_{j=1}^{M} n_{t,j}^k \qquad k \geq 1 \tag{1.4}$$

where t is the time period over which the measure is computed, M is the number of subintervals into which t can be split, and $n_{t,j}$ is the number of news stories disclosed within *(t-1 + (j-1)/M)* and *(t-1 +*

j/M)—that is, in each subinterval. t can range from few minutes to a day or a series of days. If t is a day, it will be split in a series of intraday intervals, such as five minutes, ten minutes, or fifteen minutes. If t is a longer period—say, a week or a month—it is reasonable to divide it into a series of days, such as one-day, two-day, or five-day intervals.

- **quantity variation**: variation across periods of the quantity of news stories (or words). This concept takes into account the chance that investors' reactions are triggered not only by the release of information, but more generally by increases in the quantity of information. The market can become accustomed to news releases such that it perceives them as informative only when they are released at a higher (lower) rate than usual, in which case they wait for the rate of information arrival to increase (decrease) before making a decision.

- **news persistence/interaction**: when the quantity of news is above a threshold in each of two consecutive periods. Since providers do not supply redundant news[9], this event denotes persistence in the release of news stories that are related in each period to a different issue.

- **sentiment inversion**: when the sentiment of the reference period is opposite to that of previous periods.

- **quantity variation conditional on sentiment**: positive quantity variation conditional on the sentiment of the reference period and negative quantity variation conditional on the sentiment of the previous period. The sentiment of the period with a higher quantity of information is likely to have the greater influence on investors' attention.

- **sentiment conditional on quantity**: sentiment of the reference period conditional on the quantity of information released during the same and during longer periods. Investors may base their decisions on the sentiment of the reference period, but their attention may depend on the quantity of information that is released during periods of the same duration or during longer periods.

1.5.2 Standardized Surprises of Earnings and Macro-Announcements

Earnings and macro-surprises are constructed using techniques widespread in the literature.

With regard to earnings announcements, from actual and consensus forecasts of EPS we compute the Standardized Unexpected Earnings score (SUE), which measures the number of standard deviations by which the reported actual earnings per share differ from the consensus forecast.

$$SUE_t = \frac{EPS_t^{actual} - EPS_t^{forecast}}{\sigma(EPS_t^{actual} - EPS_t^{forecast})} \tag{1.5}$$

where $\sigma(EPS_t^{actual} - EPS_t^{forecast})$ is the standard deviation of $(EPS_t^{actual} - EPS_t^{forecast})$.

With regard to macro-announcements, we compute from actual and consensus forecasts of the indicators the standardized surprise, *Std_Macro*, as we did for earnings.

$$Std\_Macro_t = \frac{Macro_t^{actual} - Macro_t^{forecast}}{\sigma(Macro_t^{actual} - Macro_t^{forecast})} \tag{1.6}$$

where Macro generically stands for any of the ten indicators listed in Table 1.1 and $\sigma(Macro_t^{actual} - Macro_t^{forecast})$ is the standard deviation of $(Macro_t^{actual} - Macro_t^{forecast})$.

1.5.3 Google Search Index

Google restricts access to daily data for intervals longer than ten months but allows daily data to be gathered for shorter intervals. The series covering our sample period of ten years is available with a monthly aggregation only. We reconstruct the daily search volume series for the whole sample, which we call Google Search Index[10] (GSI), where all observations are rescaled in order to be comparable to each other and the maximum observation over the series is equal to 100. In reconstructing this series, we use:

- the set of daily series for each month $GT\_Daily_{d,m}$ (121 series, one for each month from February 2005 to February 2015, with the observations in each series equaling the number of days in the month), where for each series the observations are relative to the maximum of 100.

- the monthly-aggregated series for the whole sample $GT\_Monthly_m$ (one series having 121 observations), where the observations are relative to the maximum of 100.

[9] News providers claim to supply only novel news stories, so we expect them not to report the same information more than once.

[10] The term "Google Search Index" is consistent with the recent literature.

We employ a three-step procedure:

1. Compute the relative contribution of day d to the search volume of month m $GT\_DailyRel_{d,m}$, by dividing the daily observation of day d (relative to the maximum of month m) $GT\_Daily_{d,m}$ by the sum of all the daily observations of that month:[11]

$$GT\_DailyRel_{d,m} = \frac{GT\_Daily_{d,m}}{\sum_{d=1}^{M_m} GT\_Daily_{d,m}} \tag{1.7}$$

   where M_m is the number of days of month m.

2. Compute the daily observations relative to the whole sample $GT_{d,m}$, by multiplying the relative contribution of day d to the search volume of month m $GT\_DailyRel_{d,m}$ by the monthly observation of month m $GT\_Monthly_m$:

$$GT_{d,m} = GT\_DailyRel_{d,m} \cdot GT\_Monthly_m \tag{1.8}$$

3. Find $GSI_{d,m}$ by dividing by the maximum and multiplying by 100:

$$GSI_{d,m} = \frac{GT_{d,m}}{max(GT_{d,m})} \cdot 100 \tag{1.9}$$

   where $max(GT_{d,m})$ is the max of $GT_{d,m}$ over the series.

1.5.4 Proposed Measures Based on Various Time Horizons

We propose a set of news measures that can be linked to daily asset price dynamics. In order to build news-related variables that are linked to heterogeneous market players who assimilate and react to news disclosure at differing speeds, we consider the information released during four time horizons:

- **Daily**: information from the market closing time of day *t-1* to the market closing time of day t[12]
- **Overnight**: information from the market closing time of day *t-1* to the market opening time of day t
- **Weekly**: the most recent five trading days
- **Monthly**: the most recent twenty-two trading days

As a last step, with the aim of identifying possible non-linearities in the relationship between market activity and the indicators, we extend the variables along a series of monotonic transformations, which are detailed in Table 1.5. Based on the values the original measure x can assume, we apply either all transformations or only a subsample of them. For instance, if x can only be non-negative (e.g., *n. news stories*), *flag if* $x > 0$ and *flag if* $x < 0$ are not applied; if x is the *sentiment*, only x, *flag if* $x > 0$, and *flag if* $x < 0$ are applied; if x is the *day-to-day* Δ *n. news stories*, all transformations are applied.

Table 1.5: Measures transformations.

Transformation	**Formula**
original measure	x
flag if $x \neq 0$	$\begin{cases} 1 & \text{if} \quad x \neq 0 \\ 0 & \text{otherwise} \end{cases}$
flag if $x > 0$	$\begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{otherwise} \end{cases}$
flag if $x < 0$	$\begin{cases} 1 & \text{if} \quad x < 0 \\ 0 & \text{otherwise} \end{cases}$
signed square root(x)	$sign(x) \cdot \sqrt{\lvert x \rvert}$
signed log(x)	$sign(x) \cdot log(1 + \lvert x \rvert)$
signed square(x)	$sign(x) \cdot x^2$

[11] *Rel* refers to the daily observations' being divided by their sum over the month in such a way that their sum over each month is equal to 1. Monthly observations are divided by their sum over the whole sample period.

[12] *t*, *t-1*, and so on refer to trading days only, so information released during holidays and weekends is considered part of the daily information of the first following trading day, as well as part of its overnight information.

Tables 1.6 to 1.11 report the measures built from news stories. Tables 1.6 to 1.9 correspond to one table for each time horizon, while tables 1.10 and 1.11 list the measures based on the aggregation or comparison of the information across more than one time horizon. Tables 1.12 and 1.13 report the measures built from EPS and macro-news. Information is aggregated over daily, overnight, weekly, and monthly time horizons. EPS and macro-announcements are released with frequencies ranging from one week to several months, so measures based on their comparison across periods would either represent lagged announcements or zero. Therefore, differently from news stories' measures, EPS and macro-news are not compared across periods. Table 1.14 reports the measures built from Google Trends. Summing news-related variables for StreetAccount news stories, Thomson Reuters news stories, earnings announcements, macro-announcements, and Google Trends, we have 5,159 news measures for each asset.

Table 1.6: Daily news stories measures.

Variable	**N. Transf.**
STANDARD	
n. news stories	5^a
n. words	4
sentiment	3^b
ABNORMAL QUANTITY	
n. news stories ≥ 2	1^c
UNCERTAINTY	
pos and neg news in same day	1
NEWS BURST INDEX	
news burst index (n. news)	6^d
$(M = 78, 13, 3)$x$(k = 2, 4)$	
news burst index (n. words)	6
$(M = 78, 13, 3)$x$(k = 2, 4)$	
SENTIMENT COND. ON QUANTITY	
pos sent & n. news stories ≥ 2	1
neg sent & n. news stories ≥ 2	1
total for each topic	28
grand total (28 x 31^e)	**868**

Notes: The first column shows the variables grouped by the concepts that originated them. The second column shows the number of transformations, with the total number of measures obtained at the end of the column. We obtain 868 measures.
[a]: When the original measure can only be positive, such as *number of news stories*, the two transformations *flag if* $x > 0$ and *flag if* $x < 0$ are omitted, leaving five transformations. *flag for number of words* $\neq 0$ is omitted because this measure corresponds to *flag for number of news stories* $\neq 0$.
[b]: The transformations applied for are *original measure*, *flag if* $x > 0$, and *flag if* $x < 0$.
[c]: When the number of transformations equals 1, the measure consists of a flag (1 for the occurrence of the event, and 0 otherwise).
[d]: For *news burst index* we report in the second column the number of combinations of the parameters M and k, and do not apply transformations.
[e]: There are seven topics for StreetAccount news stories and six topics and four levels of importance for Thomson Reuters news stories. 31 stands for the sum of the number of topics of StreetAccount (7) and the number of topics times four levels of importance of Thomson Reuters (24).

Table 1.7: Overnight news stories measures.

Variable	**N. Transf.**
STANDARD	
n. news stories	5
n. words	4
sentiment	3
ABNORMAL QUANTITY	
n. news stories ≥ 2	1
UNCERTAINTY	
pos and neg news in same day	1
SENTIMENT COND. ON QUANTITY	
pos sent & n. news stories ≥ 2	1
neg sent & n. news stories ≥ 2	1
total for each topic	16
grand total (16 x 31)	**496**

Table 1.8: Weekly news stories measures.

Variable	**N. Transf.**
STANDARD	
av. n. news stories[a]	5
av. n. words	4
sentiment[b]	3
ABNORMAL QUANTITY	
av. n. news stories ≥ 1	1
NEWS BURST INDEX	
news burst index (n. news)	2
$(M = 5)$x$(k = 2, 4)$	
news burst index (n. words)	2
$(M = 5)$x$(k = 2, 4)$	
SENTIMENT CONDITIONAL ON QUANTITY	
pos sent & av. n. news stories ≥ 1	1
neg sent & av. n. news stories ≥ 1	1
total for each topic	19
grand total (19 x 31)	**589**

[a]: Quantities result from averaged daily quantities over the last five trading days. Av. refers to average.
[b]: *Sentiment* results from the sign of the averaged *sentiment* over the last five trading days.

Table 1.9: Monthly news stories measures.

Variable	**N. Transf.**
STANDARD	
av. n. news stories	5
av. n. words	4
sentiment	3
ABNORMAL QUANTITY	
av. n. news stories ≥ 1	1
NEWS BURST INDEX	
news burst index (n. news)	2
$(M = 22)$x$(k = 2, 4)$	
news burst index (n. words)	2
$(M = 22)$x$(k = 2, 4)$	
SENTIMENT CONDITIONAL ON QUANTITY	
pos sent & av. n. news stories ≥ 1	1
neg sent & av. n. news stories ≥ 1	1
total for each topic	19
grand total (19 x 31)	**589**

Table 1.10: Multi-period news stories measures 1/2.

Variable	**N. Transf.**
QUANTITY VARIATION[a]	
day-to-day Δ n. news stories	7
week-to-day Δ n. news stories	7
month-to-day Δ n. news stories	7
day-to-day Δ n. words	7
week-to-day Δ n. words	7
month-to-day Δ n. words	7
NEWS PERSISTENCE/INTERACTION[b]	
n. news stories today ≥ 2 & n. news stories day before ≥ 2	1
n. news stories today ≥ 2 & av. n. news stories week before ≥ 1	1
n. news stories today ≥ 2 & av. n. news stories month before ≥ 1	1
SENTIMENT INVERSION[c]	
day-to-day sent inv	1
day-to-day sent inv, neg to pos	1
day-to-day sent inv, pos to neg	1
week-to-day sent inv	1
week-to-day sent inv, neg to pos	1
week-to-day sent inv, pos to neg	1
month-to-day sent inv	1
month-to-day sent inv, neg to pos	1
month-to-day sent inv, pos to neg	1

Notes: Measures created from the aggregation or comparison of information across different periods, which can differ from one another.
[a]: *day-to-day Δ n. news stories* is equal to the number of news stories on day t minus the number of news stories on day *t-1*; *week-to-day Δ n. news stories* and *month-to-day Δ n. news stories* are equal to the number of news stories on day t minus the average number of news stories in the week before (from *t-5* to *t-1*) and in the month before (from *t-22* to *t-1*), respectively.
[b]: *news persistence/interaction* describes the event in which the amount of news is above a certain threshold in each of two consecutive periods.
[c]: *day-to-day sent inv* describes the event in which *sentiment* on day t is the opposite of *sentiment* on day *t-1*; *day-to-day sent inv, neg to pos* and *day-to-day sent inv, pos to neg* describe events in which *sentiment* is negative on day *t-1* and positive on day t and the reverse, respectively.

Table 1.11: Multi-period news stories measures 2/2.

Variable	**N. Transf.**
QUANTITY VARIATION COND. ON SENTIMENT[a]	
day-to-day Δ n. news stories > 0 & pos sent today	1
day-to-day Δ n. news stories > 0 & neg sent today	1
day-to-day Δ n. news stories < 0 & pos sent day before	1
day-to-day Δ n. news stories < 0 & neg sent day before	1
week-to-day Δ n. news stories > 0 & pos sent today	1
week-to-day Δ n. news stories > 0 & neg sent today	1
week-to-day Δ n. news stories < 0 & pos sent week before	1
week-to-day Δ n. news stories < 0 & neg sent week before	1
month-to-day Δ n. news stories > 0 & pos sent today	1
month-to-day Δ n. news stories > 0 & neg sent today	1
month-to-day Δ n. news stories < 0 & pos sent month before	1
month-to-day Δ n. news stories < 0 & neg sent month before	1
SENTIMENT COND. ON PAST QUANTITY[b]	
pos sent today & n. news stories day before ≥ 2	1
neg sent today & n. news stories day before ≥ 2	1
pos sent today & av. n. news stories week before ≥ 1	1
neg sent today & av. n. news stories week before ≥ 1	1
pos sent today & av. n. news stories month before ≥ 1	1
neg sent today & av. n. news stories month before ≥ 1	1
total for each topic	72
grand total (72 x 31)	**2232**

[a]: *day-to-day Δ n. news stories > 0 & pos sent today* describes the event in which the number of news stories on day t is greater than the number of news stories on day *t-1* and the *sentiment* n day t is positive; the remaining variables in the group *quantity variation conditional on sentiment* are straightforward. The sentiment conditioning the occurrence of the event is that of the period with a greater amount of news; therefore, we look at the sentiment of the period before day t for negative variations.
[b]: The variables that belong to the group *sentiment conditional on past quantity* describe the events in which the sentiment on day t is positive or negative and when, in the period before, the quantity of news is above a threshold that equals 2 for the number of news stories on the day before and 1 for the average number of news stories in the week and the month before.

Table 1.12: EPS measures.

Variable	**N. Transf.**
daily SUE	8[a]
overnight SUE	8
weekly SUE	8
monthly SUE	8
grand total	**32**

Notes: EPS measures result from the EPS released in the corresponding period. For example, *weekly SUE* is equal to the SUE if there was an EPS release in the last week.
[a]: In addition to the seven transformations of Table 1.5, we add a flag variable for the occurrence of an EPS release (1 for occurrence, and 0 otherwise).

Table 1.13: Macro-measures.

Variable	**N. Transf.**
daily Std_CCONF	8[a]
daily Std_CPI	8
daily Std_FOMC	8
daily Std_GDP	8
daily Std_INDPROD	8
daily Std_BOP	8
daily Std_JOB	8
daily Std_NFP	8
daily Std_PPI	8
daily Std_RSALES	8
overnight Std_Macro[b]	8 x 10
weekly Std_Macro	8 x 10
monthly Std_Macro	8 x 10
grand total	**320**

Notes: Macro-measures result from the macro-announcement released in the corresponding period, as for EPS measures.
[a]: In addition to the seven transformations of Table 1.5, we add a flag variable for the occurrence of a macro-release (1 for occurrence, and 0 otherwise), as for EPS measures.
[b]: *Std_Macro* refers to the standardized surprise of any of the macro-indicators, which are reported only in the *daily* group for reasons of brevity. In the second column, the number of transformations is multiplied by the number of macro-indicators.

Table 1.14: Google Trends measures.

Variable	**N. Transf.**
daily GSI	4[a]
weekly av. GSI	4
monthly av. GSI	4
day-to-day Δ GSI	7
week-to-day Δ GSI	7
month-to-day Δ GSI	7
grand total	**33**

Notes: weekly av. GSI and *monthly av. GSI* correspond to the average of *daily GSI* over the last five and twenty-two trading days, respectively. *day-to-day* Δ *GSI* is equal to GSI on day t minus GSI on day *t-1*; *week-to-day* Δ *GSI* and *month-to-day* Δ *GSI* are equal to GSI on day t minus the average GSI in the week before (from *t-5* to *t-1*) and in the month before (from *t-22* to *t-1*), respectively.
[a]: We use four transformations: *original measure*, *signed square root*, *signed log*, and *signed square*.

1.6 Application: News Measures and Volatility Forecasting

In the previous sections we extracted news stories' sentiment, identified a set of concepts/events to be used for the development of related variables, extracted surprises from expectations of earnings and macro-announcements, and reconstructed the daily series of the Google Search Index. Then we aggregated variables from overnight to monthly time horizons and applied monotonic transformations in order to obtain a large set of news indicators with the aim of reconstructing the portions of information on which heterogeneous market players are likely to base their decisions.

We want to verify the validity of the MDH and, more generally, to shed light on the link between news and volatility. We focus on the relative importance of the five main types of information in our database—that is, news stories from the two providers, earnings announcements, macro-announcements, and Google Trends—as well as on the relative importance of the volume and variations of news stories and their topics, and on announcements *per se* versus surprises from expectations of earnings and macro-announcements. We also compare the two providers with regard to the relevance of the news stories they release to explaining price movements.

We model daily realized volatility with the HAR-TCJ linear model from Corsi et al. (2010), which is based on the HAR-CJ model from Andersen et al. (2007a), and add the news measures as explanatory variables. We face a dimensionality problem and use the LASSO estimation method to solve it and to select the measures that are most useful in explaining volatility. Finally, we employ the news indicators to forecast volatility.

1.6.1 Methodology: Realized Volatility Modelling with News

We compute daily realized volatility from five-minute returns using the preceding or concurrent price nearest to each five-minute mark. Then we decompose the daily realized volatility into its continuous and jump components using the jump test from Corsi et al. (2010). (See Appendix A.2 for a detailed description of realized volatility measurement and jump testing.) Realized volatility is a process characterized by a well-known strong temporal dependence. Andersen et al. (2007a) model realized volatility using the HAR-CJ model, which consists of an extension of the linear HAR model from Corsi (2009). The HAR-CJ model separates the quadratic variation into its continuous part and jumps, and uses them to capture its autoregressive properties. Corsi et al. (2010) use the corrected threshold multi-power variation measures in the HAR-CJ model, referring to it as the HAR-TCJ model.

Let $t = 1, \dots$ be the day index and $RV_t = RV_\delta(X)_t$. For two days t_1 and $t_2 \geq t_1$, define

$$RV_{t_1:t_2} = \frac{1}{t_2 - t_1 + 1} \sum_{t=t_1}^{t_2} RV_t \tag{1.10}$$

According to the HAR-TCJ model:

$$RV_t = \beta_0 + \beta_d \widehat{C}_d + \beta_w \widehat{C}_w + \beta_m \widehat{C}_m + \beta_j \widehat{J}_d + \epsilon_t \tag{1.11}$$

where $\widehat{C}_d = \widehat{C}_{t-1}, \widehat{C}_w = \widehat{C}_{t-5:t-1}, \widehat{C}_m = \widehat{C}_{t-22:t-1}, \widehat{J}_d = \widehat{J}_{t-1}$.

We add the news measures to the explanatory variables of the HAR-TCJ model, and refer to it as the HAR-TCJN (news-augmented HAR-TCJ) model:

$$RV_t = \beta_0 + \beta_d \widehat{C}_d + \beta_w \widehat{C}_w + \beta_m \widehat{C}_m + \beta_j \widehat{J}_d + \beta_{News}^T News_{t-1} + \epsilon_t \tag{1.12}$$

where β_{News} is the k x 1 vector of coefficients, ^{T} denotes transposition, and $News_{t-1}$ is the k x 1 vector of news measures available before the market opens on day t.

We employ the logarithmic counterparts of the models[13], which read:

$$\log RV_t = \beta_0 + \beta_d \log \widehat{C}_d + \beta_w \log \widehat{C}_w + \beta_m \log \widehat{C}_m + \beta_j \log (1 + \widehat{J}_d) + \epsilon_t \tag{1.13}$$

$$\log RV_t = \beta_0 + \beta_d \log \widehat{C}_d + \beta_w \log \widehat{C}_w + \beta_m \log \widehat{C}_m + \beta_j \log (1 + \widehat{J}_d) + \beta_{News}^T News_{t-1} + \epsilon_t \tag{1.14}$$

[13] As Andersen et al. (2003) point out, while the distributions of realized volatilities are clearly right-skewed, the distributions of realized volatilities' logarithms are approximately Gaussian. Andersen et al. (2003) also use the logarithmic transformation to model and forecast the realized volatilities. Corsi et al. (2010), in the part of their empirical analysis related to individual stocks, also report results for the logarithmic model only. Results for the original models and their square root counterparts are available upon request.

Using all the measures we created in section 1.5, k is equal to 5,159. We face a dimensionality issue since the number of regressors is higher than the number of observations, the latter being smaller than 3,000. We use LASSO to address the issue and to select the measures that are the most useful in explaining volatility.

LASSO (Tibshirani (1996)) is an estimation method for linear models that performs variable selection and shrinks coefficients. By minimizing the residual sum of squares, subject to the sum of the absolute value of the coefficients' being less than a constant, LASSO shrinks some coefficients and sets others to 0, thereby providing interpretable models. In addition, the coefficients it produces have potentially lower predictive errors than ordinary least squares do. Audrino and Knaus (2016) also use LASSO to model realized volatility[14].

Suppose we have data $(x^i, y_i), \quad i = 1, ..., N$, where $x^i = (x_{i1}, \ldots, x_{ip})^T$ are the predictor variables, y_i is the response, and N is the number of observations. It is assumed that either the observations are independent or that y_i is conditionally independent given x_{ij} and that x_{ij} is standardized so that $\sum_i x_{ij}/N = 0, \quad \sum_i x_{ij}^2/N = 1$.

Letting $\hat{\beta} = (\hat{\beta_1}, \ldots, \hat{\beta_p})^T$, the LASSO estimate $(\hat{\alpha}, \hat{\beta})$ is:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \left(\sum_{i=1}^{N} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \right) \qquad \text{subject to} \quad \sum_j |\beta_j| \leq t \tag{1.15}$$

where $t \geq 0$ is a tuning parameter. Following James et al. (2013), the tuning parameter t is selected using cross-validation. We use the package *Glmnet* for its software *R*, which computes the cross-validation error for each t among a set of values, and we select from a grid of 100 values the t for which the ten-fold cross-validation mean squared error is smallest.

We estimate the parameters of the HAR-TCJN model using LASSO and apply the restriction to all coefficients except the constant β_0. Therefore, the restriction is applied to β_d, β_w, β_m, β_j and β_{News}, the latter consisting in a vector of 5,159 coefficients.

1.6.2 Uncovering News Impact on RV

For each asset in our set of eighty-nine stocks selected[15] from the S&P 100, we compute the five-minute realized volatility (RV) and build the news measures database for the period from February 2005 to February 2015. As for the past components of realized volatility employed by the HAR-TCJ model, news measures for day t realized volatility are built on the basis of the information available until the market opening time of day t, including overnight news. All measures are centered and standardized by subtracting the mean and then dividing by the standard deviation, as prescribed by Tibshirani (1996).

We use LASSO to select the news-based measures that are most relevant to explaining volatility. Table 1.15 reports the ranking of the thirty most frequently selected indicators, using as the selection criterion the number of assets for which its estimated β is different from zero. The table reports the percentages of positive and negative estimated coefficients, and includes the coefficients associated to the past volatility components. The past continuous volatility components are always selected, and the coefficient is always positive, while past jumps are selected for more than half of the assets, and the coefficient is always positive. Compared to past continuous volatility components, past jumps appear to be much less relevant drivers of volatility, but being selected for more than half of the assets among thousands of measures is a strong indicator of their relevance. The types of news that are most relevant to explaining volatility are macroeconomic indicators and EPS, which fill the first ten positions, while news stories and Google Trends follow.

Eight of the ten indicators (NFP, FOMC, JOBLESS, CPI, CCONF, GDP, INDPROD, RSALES) belong to the thirty most frequently selected measures. NFP is the leading indicator. The monthly surprise (79%)[16] and the monthly log surprise (35%) have both a negative sign, suggesting that lower wages scare investors, who trade more actively as a consequence. With regard to this indicator, investors look at the information released during the most recent weeks. The weekly flag for announcement of an FOMC rate decision (57%) has a positive sign, reaffirming the well-documented market reaction to FOMC rate decisions. FOMC monthly surprise (25%), with a positive sign, indicates that, if rates increased with respect to expectations during the last month, market activity is likely to be higher than if they decreased with respect to expectations, in which case market activity cools down. Rate decisions are always dependent on reactions to positive and negative surprises from expectations and on the state of the economy, so we look also at subsample results reported in tables A.4 – A.6 in Appendix A.3. These tables confirm that FOMC announcements on the most recent day and week increase volatility. With regard to the aggregation of this information during the most recent month, lower-than-expected rate decisions during contractionary periods seem to calm market activity, while higher-than-expected rate decisions during expansionary periods seem to boost markets. Remembering that FOMC

[14] Audrino and Knaus (2016) find that the HAR model's lags structure is not fully in agreement with the one identified from a model-selection perspective using LASSO on real data.

[15] See subsection 1.3.1 for the selection criteria.

[16] Hereafter, "surprise" refers to standardized surprise, and "log," "square root," and "square" refer to sign-preserving transformations. In brackets, the percentage of assets for which the measure is selected.

announcements increase volatility, the results of contractions and expansions may be reversals and persistence effects, respectively. (A complete analysis of market reactions to rate decisions is outside the scope of this study.)

The weekly flag for a positive surprise of JOBLESS (52%), the flag for an overnight surprise different from zero (46%), and the flag for a daily positive surprise (40%) all have positive coefficients, suggesting that jobless claims announcements increase volatility, especially when the indicator is higher than expected (negative news), and information released during the most recent day and week is an influential driver of volatility.

The monthly flag for a CPI announcement (44%) has a positive sign, but since CPI is released monthly, this result, taken alone, is not meaningful. The analysis by subsample shows that only during contractions does the CPI-related information aggregated at daily and weekly horizons count: the flag for a lower-than-expected release during the most recent day (44.94%) and the flag for a surprise that is different from zero during the most recent week (17.98%) both have a positive coefficient, suggesting that investors tend to look at this indicator especially during bad times and are more concerned when the news is worse than expected than when it is better than expected.

The overnight flag for a CCONF announcement (44%) has a negative sign, indicating that the fresh release of a lower-than-expected index of consumer confidence scares markets.

The weekly square surprise of GDP (40%) has a positive sign, while its monthly log surprise (39%) has a negative sign. The coefficient of the monthly indicator suggests that negative news about gross domestic product during the last month scares and moves markets; while the coefficient of the weekly indicator is difficult to interpret, it seems to be at odds with the traditional stronger reaction to negative news reported in the literature. It is possible that positive news about GDP during the most recent week moves markets as well.

With regard to INDPROD, the daily square surprise (36%) and the monthly surprise (29%) both have a negative sign, suggesting that investors become concerned as a consequence of negative news on industrial production and increase their activity. They look mainly at recent announcements but also take into account information from the most recent month.

The weekly flag for a negative RSALES surprise (26%) has a positive sign, and the weekly square surprise (21%) a negative one, both suggesting a rise in volatility as a consequence of negative news about retail sales during the most recent week.

The daily flag for an EPS announcement (72%) and the daily flag for a surprise different from zero (31%) both have a positive sign, indicating that a market reaction takes place the day after EPS announcements are released. Both announcements *per se* and surprises count, and asymmetric effects between positive and negative surprises are not evident.

Both StreetAccount and Thomson Reuters news stories appear to be useful determinants of volatility, although the measures for StreetAccount news stories are selected more often. Analysis of the news stories' indicators is performed below.

Finally, the weekly log Google Search Index (37%) has a positive sign, suggesting that retail investors' attention during the most recent week is positively related to market activity.

Table 1.16 collects the forty most frequently selected variables related to news stories. Remarkably, almost all coefficients are positive. Since they are related to proxies for the quantity of information and to flags for variations in this quantity, we interpret the positive signs as important evidence that news releases, on average, increase volatility.

Both providers are relevant drivers of volatility, although StreetAccount news variables are selected more often.

As expected, earnings is the most important topic, followed by upgrades/downgrades. However, many measures are based on news that is not filtered by topic (all), indicating that news that is related to earnings and upgrades/downgrades does not exhaust the interest of market players. The level of importance assigned by Thomson Reuters does not appear to be as relevant as the topic. Indeed, the first three indicators of Thomson Reuters news that are selected belong to the earnings topic and are not filtered by importance. Nevertheless, for news with the earnings topic, all levels of importance appear between the most frequently selected measures, and the coefficient is positive in all cases. Remembering that a higher level of importance corresponds to a tighter filter and that, as a consequence, news tagged with higher importance also appears among news tagged with lower importance, the positiveness of the coefficients associated with all levels of importance (among news stories on the earnings topic) suggests that filtering by importance implies additional increasing effects on volatility. Therefore, classification by importance may correspond with the news stories' relevance to explaining volatility. With regard to news released by Thomson Reuters, only news on the earnings topic appear among the forty most frequent topics.

With regard to the time horizon, variables based on day-to-day variations and daily aggregation of news dominate. Flags for variations of the quantity of information, both when the daily aggregation is proxied by the number of news stories and when the number of words is used, are all associated with a positive coefficient, suggesting that flows of information are more important than levels of information, probably because investors become accustomed to the rate of information arrival and perceive only variations as informative. Only variations

Table 1.15: Most selected regressors in the log HAR-TCJN model. Sample: 2005 – 2015.

		Past Volatility Components					**% Selected**	**% Pos**	**% Neg**
		$\log C_d$					100.00	100.00	0.00
		$\log C_w$					100.00	100.00	0.00
		$\log C_m$					100.00	100.00	0.00
		$\log(1 + J_d)$					59.55	59.55	0.00
Macro	**Firm-Specific**	**News**	**Importance**	**Topic**	**Time Aggregation**	**Measure**	**% Selected**	**% Pos**	**% Neg**
X		NFP			month	surp	78.65	0.00	78.65
	X	EPS			day	flag for announcement	71.91	71.91	0.00
X		FOMC			week	flag for announcement	57.30	57.30	0.00
X		JOBLESS			week	flag if surp > 0	51.69	51.69	0.00
X		JOBLESS			overnight	flag if surp $\neq 0$	46.07	46.07	0.00
X		CPI			month	flag for announcement	43.82	43.82	0.00
X		CCONF			overnight	flag for announcement	43.82	0.00	43.82
X		JOBLESS			day	flag if surp > 0	40.45	40.45	0.00
X		GDP			week	square surp	40.45	40.45	0.00
X		GDP			month	log surp	39.33	0.00	39.33
	X	SA news		earnings	flow: day-to-day	flag if Δ n. words < 0	38.20	38.20	0.00
		GSI			week	log GSI	37.08	35.96	1.12
X		INDPROD			day	square surp	35.96	0.00	35.96
	X	SA news		all	flow: day-to-day	persistence/interaction	35.96	35.96	0.00
X		NFP			month	log surp	34.83	0.00	34.83
	X	TR news	low	earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	34.83	34.83	0.00
	X	SA news		earnings	flow: day-to-day	flag if Δ n. news stories < 0	34.83	34.83	0.00
	X	SA news		all	flow: day-to-day	flag if Δ n. words < 0	34.83	34.83	0.00
	X	SA news		earnings	flow: day-to-day	flag if Δ n. words $\neq 0$	33.71	33.71	0.00
	X	EPS			day	flag if surp $\neq 0$	31.46	31.46	0.00
X		INDPROD			month	surp	29.21	0.00	29.21
	X	SA news		earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	29.21	29.21	0.00
	X	SA news		all	flow: day-to-day	flag if Δ n. news stories $\neq 0$	29.21	29.21	0.00
X		RSALES			week	flag if surp < 0	25.84	25.84	0.00
	X	TR news	low	earnings	flow: day-to-day	flag if Δ n. words $\neq 0$	25.84	25.84	0.00
	X	SA news		up/downgrades	flow: day-to-day	flag if Δ n. words $\neq 0$	25.84	25.84	0.00
X		FOMC			month	surp	24.72	24.72	0.00
	X	SA news		all	day	flag if n. news stories ≥ 2	23.60	23.60	0.00
	X	SA news		all	flow: day-to-day	flag if Δ n. news stories < 0	22.47	22.47	0.00
X		RSALES			week	sqrt surp	21.35	0.00	21.35

Notes: Ranking of regressors (past volatility components plus the thirty most frequently selected news measures) by percentage of stocks for which they are selected by LASSO in the log HAR-TCJN model, percentage of positive and percentage of negative coefficients. Sample: Feb. 2005 – Feb. 2015.

that differ from zero and negative variations were selected, suggesting that, in many cases, investors wait for the rate of information to decrease to make decisions. Day-to-day news persistence/interaction appears in the second position in the ranking. The release of at least two news stories in each of two consecutive days highlights the importance of the persistence/interaction concept—an observation that, as far as we know, other studies do not highlight. Measures based on daily information levels play a minor role in the explanation of volatility, but they still count. Among them, the flag for the release of at least two news stories in a day is the most important variable.

Only two sentiment-related measures appear among the top forty. They belong to the group *sentiment conditional on past quantity* and consist in the *flag for negative sentiment on day t and number of news stories* ≥ 2 *on day t-1*, and in the *flag for positive sentiment on day t and number of news stories* ≥ 2 *in day t-1*. The most frequently selected sentiment-related measure is that associated with a negative sentiment. It is possible to interpret this result as support, even if faint, for the literature's notion that negative news moves markets more than positive news does. Sentiment-based measures are much less significant determinants of volatility than quantity-based measures are, suggesting either that sentiment is less important than the amount of information or that the sentiment detection procedure should be improved.

Summarizing the results from tables 1.15 and 1.16, macro-announcements and EPS are the most important drivers of volatility, but news stories and Google Trends also play a role. While both macro-announcements and surprises from expectations affect market reactions, markets tend to react more strongly to negative surprises, and they consider the information released during several time horizons, from overnight to the most recent month. EPS announcements and surprises are both important as well, and there is no evident asymmetric effect between positive and negative surprises. Only EPS information released during the most recent day seems relevant to explaining volatility. News stories from StreetAccount are slightly more useful than Thomson Reuters' news stories are in explaining market reactions, especially in the form of variables based on day-to-day variations in the rate of information arrival. In addition, earnings is the most important news topic in affecting volatility. Retail investors' attention during the most recent week, as revealed by Google Trends, is positively linked to volatility. We leave a high frequency analysis for future work.

The MDH states that the rate of information arrival explains "the GARCH effects in asset returns," so it also explains the autoregressive behavior of volatility. In order to test this idea, we perform two OLS regressions with HAC standard errors—one for the HAR-TCJ model and one for the HAR-TCJN model—employing as news variables only those that were selected, and comparing the estimated autoregressive coefficients between the two models. Table 1.17 presents the estimation results for the autoregressive coefficients (cross-sectional average) β_0, β_d, β_w, β_m and β_J for both models and their variation after the inclusion of news[17]. Coefficients are all positive and, with the exception of β_m, their value is lower for the HAR-TCJN model, while the intercept β_0 is much higher for the HAR-TCJN model. These variations highlight the relevance of news as a driver of additional information, which involves effects on the estimated autoregressive coefficients. These results are consistent with the MDH. We also performed an F-test for the joint significance of the news variables' coefficients in the HAR-TCJN model, and the F-test rejects (at the 5% level) for all stocks the null hypothesis that the news regressors have no effect on realized volatility.

[17]News' estimated coefficients are not reported here, as the focus is on the variation of the estimated autoregressive coefficients of volatility after the inclusion of news. News coefficients estimated with LASSO are described in tables 1.15 and 1.16 above.

Table 1.16: Most selected news stories measures in the log HAR-TCJN model. Sample: 2005 – 2015.

Provider	Importance	Topic	Time Aggregation	Measure	% Selected	% Pos	% Neg
StreetAccount		earnings	flow: day-to-day	flag if Δ n. words < 0	38.20	38.20	0.00
StreetAccount		all	flow: day-to-day	persistence/interaction	35.96	35.96	0.00
Thomson Reuters	low	earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	34.83	34.83	0.00
StreetAccount		earnings	flow: day-to-day	flag if Δ n. news stories < 0	34.83	34.83	0.00
StreetAccount		all	flow: day-to-day	flag if Δ n. words < 0	34.83	34.83	0.00
StreetAccount		earnings	flow: day-to-day	flag if Δ n. words $\neq 0$	33.71	33.71	0.00
StreetAccount		earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	29.21	29.21	0.00
StreetAccount		all	flow: day-to-day	flag if Δ n. words $\neq 0$	29.21	29.21	0.00
Thomson Reuters	low	earnings	flow: day-to-day	flag if Δ n. words $\neq 0$	25.84	25.84	0.00
StreetAccount		up/downgrades	flow: day-to-day	flag if Δ n. words $\neq 0$	25.84	25.84	0.00
StreetAccount		all	day	flag if n. news stories ≥ 2	23.60	23.60	0.00
StreetAccount		all	flow: day-to-day	flag if Δ n. news stories < 0	22.47	22.47	0.00
Thomson Reuters	low	earnings	flow: day-to-day	flag if Δ n. news stories < 0	21.35	21.35	0.00
StreetAccount		earnings	day	sqrt n. words	21.35	21.35	0.00
StreetAccount		all	day	n. words	21.35	21.35	0.00
Thomson Reuters	medium	earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	20.22	20.22	0.00
Thomson Reuters	low	earnings	flow: day-to-day	flag if Δ n. words < 0	20.22	20.22	0.00
StreetAccount		earnings	day	flag if n. news stories ≥ 2	20.22	20.22	0.00
StreetAccount		all	flow: day-to-day	neg sent day t, n. news stories ≥ 2 day $t-1$	20.22	20.22	0.00
StreetAccount		all	day	square n. news stories	20.22	20.22	0.00
StreetAccount		up/downgrades	flow: day-to-day	Δ n. news stories < 0, neg sent day $t-1$	19.10	19.10	0.00
StreetAccount		all	flow: day-to-day	square Δ n. news stories	19.10	0.00	19.10
Thomson Reuters	high	earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	17.98	17.98	0.00
StreetAccount		all	flow: day-to-day	pos sent day t, n. news stories ≥ 2 day $t-1$	17.98	15.73	2.25
Thomson Reuters	high	earnings	flow: day-to-day	flag if Δ n. news stories < 0	16.85	16.85	0.00
Thomson Reuters	medium	earnings	flow: day-to-day	flag if Δ n. words < 0	16.85	16.85	0.00
StreetAccount		up/downgrades	flow: day-to-day	flag if Δ n. words < 0	16.85	16.85	0.00
StreetAccount		earnings	day	n. words	16.85	16.85	0.00
StreetAccount		earnings	flow: month-to-day	log Δ n. words	15.73	15.73	0.00
StreetAccount		earnings	day	n. news stories $\neq 0$	15.73	15.73	0.00
Thomson Reuters	high	earnings	flow: day-to-day	flag if Δ n. words < 0	14.61	14.61	0.00
Thomson Reuters	medium	earnings	flow: day-to-day	flag if Δ n. news stories < 0	14.61	14.61	0.00
StreetAccount		up/downgrades	flow: day-to-day	flag if Δ n. news stories $\neq 0$	14.61	14.61	0.00
StreetAccount		all	day	square n. words	14.61	14.61	0.00
StreetAccount		up/downgrades	flow: day-to-day	flag if Δ n. news stories < 0	13.48	13.48	0.00
StreetAccount		all	flow: day-to-day	flag if Δ n. news stories < 0	13.48	13.48	0.00
StreetAccount		all	month	n. words	13.48	0.00	13.48
StreetAccount		all	day	n. news stories	13.48	13.48	0.00
Thomson Reuters	medium	earnings	flow: day-to-day	flag if Δ n. words $\neq 0$	12.36	12.36	0.00
StreetAccount		earnings	day	n. news stories	12.36	12.36	0.00

Notes: Ranking of news stories' measures (the forty most frequently selected) by percentage of stocks for which they are selected by LASSO in the log HAR-TCJN model; percentage of positive and negative coefficients. Sample: Feb. 2005 – Feb. 2015.

Table 1.17: Estimated β_0, β_d, β_w, β_m and β_J for the log HAR-TCJ and the log HAR-TCJN models.

	(a) 2005-2015			**(b) 2005-2007**		
	log HAR-TCJ	**log HAR-TCJN**	$\Delta\beta$	**log HAR-TCJ**	**log HAR-TCJN**	$\Delta\beta$
β_0	0.34	0.65	0.31	0.46	0.86	0.40
	(2.50)	(3.35)		(3.02)	(5.08)	
β_d	0.26	0.23	-0.03	0.25	0.17	-0.08
	(2.64)	(2.52)		(3.70)	(3.06)	
β_w	0.46	0.38	-0.08	0.40	0.24	-0.16
	(2.88)	(2.86)		(3.44)	(2.32)	
β_m	0.20	0.22	0.02	0.16	0.18	0.02
	(2.13)	(2.26)		(1.30)	(1.73)	
β_J	0.18	0.14	-0.04	0.16	0.08	-0.08
	(0.83)	(0.52)		(0.86)	(0.27)	
R^2	0.51	0.59		0.28	0.51	
F-test % rejection hyp. news not significant (sign. level = 5 %)		98.88%			100.00%	
	(c) 2007-2009			**(d) 2009-2015**		
	log HAR-TCJ	**log HAR-TCJN**	$\Delta\beta$	**log HAR-TCJ**	**log HAR-TCJN**	$\Delta\beta$
β_0	1.84	5.19	3.34	0.30	0.49	0.19
	(2.30)	(3.76)		(4.37)	(4.92)	
β_d	0.24	0.20	-0.04	0.45	0.41	-0.04
	(1.98)	(2.17)		(4.63)	(4.49)	
β_w	0.50	0.32	-0.18	0.13	0.10	-0.03
	(2.88)	(2.56)		(1.52)	(1.32)	
β_m	0.12	0.07	-0.05	0.25	0.25	0.00
	(0.99)	(0.67)		(3.14)	(3.37)	
β_J	0.17	-0.02	-0.19	0.22	0.19	-0.03
	(0.12)	(-0.27)		(1.61)	(1.29)	
R^2	0.44	0.63		0.36	0.51	
F-test % rejection hyp. news not significant (sign. level = 5 %)		97.75%			100.00%	

Notes: Estimated (cross-sectional average) β_0, β_d, β_w, β_m and β_J (t-statistics are in brackets), R^2 for the log HAR-TCJ and the log HAR-TCJN models, variation of the coefficients between the two models, and percentage of assets for which the F-test is rejected (null hypothesis: the coefficients of the news variables selected by LASSO are jointly not significant). OLS regression with HAC standard errors, using as explanatory variables for the HAR-TCJN model the regressors of the log HAR-TCJ model plus the news variables selected by LASSO. Samples: (**a**) Feb. 2005 – Feb. 2015 (whole sample); (**b**) Feb. 2005 – Dec. 2007 (expansion); (**c**) Dec. 2007 – Jun. 2009 (contraction); (**d**) June 2009 – Feb. 2015 (expansion).

1.6.3 Evaluating the Improvement in Forecasting Performance

Using a rolling window of 1000 observations, we iteratively estimate the HAR-TCJ and the HAR-TCJN models and apply the estimated coefficients to the information available the day following the most recent day used for estimation, obtaining the one-step-ahead forecast of realized volatility[18].

The forecasting performance of the two models is compared with five metrics, the last three of which were also used in Corsi et al. (2010):

1. The MAE mean absolute error:

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}\left|RV_t - \widehat{RV}_t\right| \tag{1.16}$$

where RV_t is the ex-post value of realized variance, and $\widehat{RV}_t$ is the forecast.

2. The MSE mean square error:

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}\left(RV_t - \widehat{RV}_t\right)^2 \tag{1.17}$$

3. The HRMSE heteroskedasticity adjusted mean square error suggested in Bollerslev and Ghysels (1996):

$$\text{HRMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{RV_t}{\widehat{RV}_t} - 1\right)^2} \tag{1.18}$$

4. The QLIKE loss function:

$$\text{QLIKE} = \frac{1}{T}\sum_{t=1}^{T}\left(\log\widehat{RV}_t + \frac{RV_t}{\widehat{RV}_t}\right) \tag{1.19}$$

5. the R^2 of Mincer-Zarnowitz forecasting regressions.

Results show that the inclusion of news-based measures substantially improves volatility forecasting.

Table 1.18 reports the cross-sectional mean over all assets of the metrics. It also includes in brackets for all metrics except the R^2 MZ the percentage of assets for which the Diebold and Mariano (1995) test rejects at a 5% significance level the null hypothesis of equal predictive accuracy in favor of each model, and in brackets for the R^2 MZ, the percentage of assets for which the metric is higher (meaning a superior predictive accuracy) for each model. The HAR-TCJN model yields, on average, lower MAE, HRMSE, and QLIKE and a higher R^2 MZ. The average MSE is instead lower for the HAR-TCJ model. As the Diebold-Mariano test reveals, the HAR-TCJN model's superior forecasting power is statistically significant for a percentage of stocks ranging (depending on the metrics MAE, MSE, HRMSE, and QLIKE) from 11.24 percent to 82.02 percent. The test does not signal a statistically significant superior predictive accuracy of the HAR-TCJ model for any metric or any stock.

[18] In a few cases, the HAR-TCJN model provides extremely low or extremely high forecasts of realized volatility, which are not reliable. We apply an adjustment procedure, detailed in Appendix A.4.

Table 1.18: One-step-ahead prediction accuracy of log HAR-TCJ and log HAR-TCJN models.

	log HAR-TCJ	**log HAR-TCJN**
MAE	0.96 (0.00%)	0.95 (26.97%)
MSE	33.82 (0.00%)	34.30 (11.24%)
HRMSE	0.92 (0.00%)	0.82 (59.55%)
QLIKE	1.45 (0.00%)	1.44 (82.02%)
R^2 MZ	0.50 (26.97%)	0.51 (73.03)%

Notes: One-step-ahead MAE, MSE, HRMSE, QLIKE, and R^2 MZ of the log HAR-TCJ and the log HAR-TCJN models (cross-sectional average). In brackets, for each model and for each metric except R^2 MZ: percentage of assets for which the Diebold-Mariano test rejects with a 5% level the null hypothesis of equal predictive accuracy in favor of that model; for R^2 MZ: percentage of assets for which the metric is higher for that model.

Figure 1.1 illustrates a dynamic analysis of the metrics. Using a rolling window of 250 days, the graphs report the percentage of assets for which the Diebold-Mariano test rejects (at a 5% significance level) the null hypothesis of equal predictive accuracy of the two models, distinguishing when the best model is the HAR-TCJN and when it is the HAR-TCJ. The dynamic analysis confirms the superior predictive accuracy obtained by including the news-related variables. The HAR-TCJ model is never superior to the HAR-TCJN model, while the HAR-TCJN model obtains a volatility forecast that is statistically never inferior and, especially in the second half of the sample, superior for a consistent percentage of assets.[19]

[19]Results for the original models and their square root counterparts are less convincing. By using news in the original models we are not able to improve volatility forecasting, while by using news in the square root counterparts we improve volatility forecasting only in the second half of the sample.

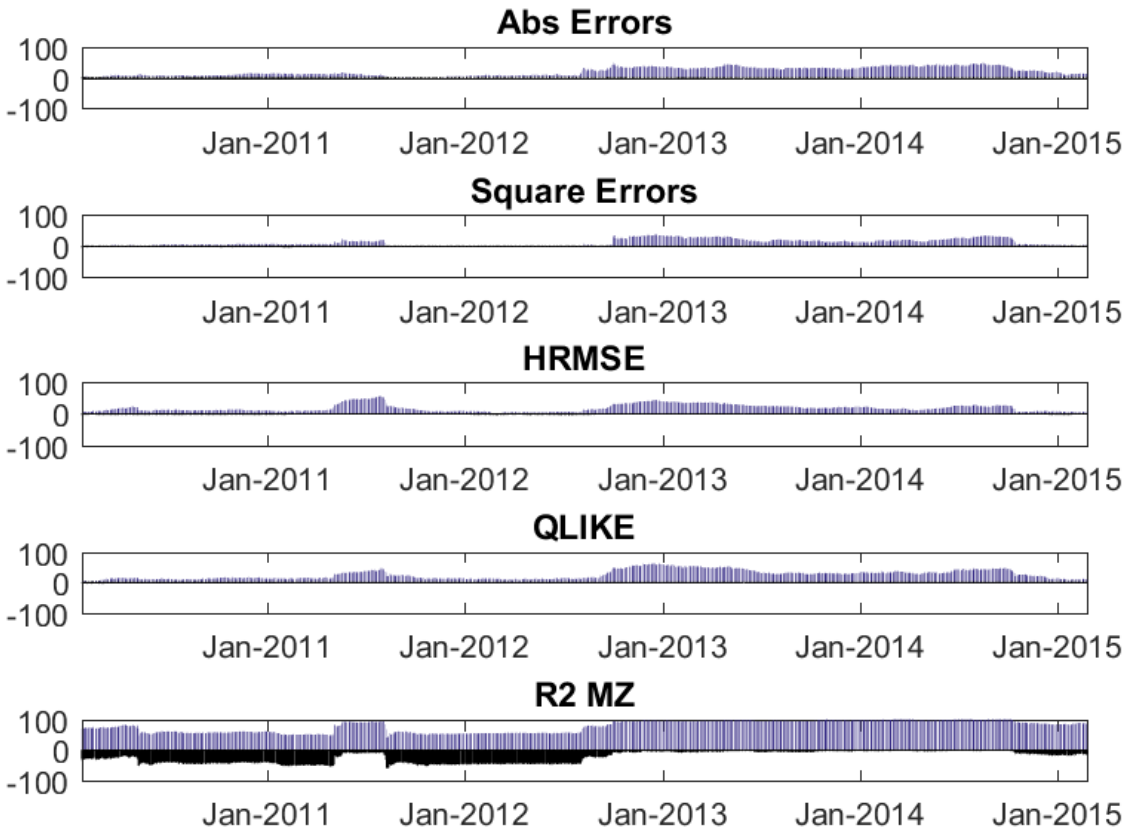

Figure 1.1: Rolling analysis of the one-step-ahead MAE, MSE, HRMSE, QLIKE, and R^2 MZ of the log HAR-TCJ and the log HAR-TCJN models. With a window size of 250 observations, the graphs report for each metric the percentage of assets for which the Diebold-Mariano test rejects at the 5% level the null hypothesis of equal predictive accuracy of the two models, distinguishing when the best model is the log HAR-TCJN (blue bars above the horizontal axis), and when the best model is the log HAR-TCJ (black bars below the horizontal axis). For the R^2 MZ, it reports the percentage of assets for which the metric is higher for the log HAR-TCJN (blue bars above the horizontal axis) and for the log HAR-TCJ (black bars below the horizontal axis).

1.7 Concluding Remarks

We created an extensive and innovative database that contains macroeconomic announcements, earnings announcements, firm-specific news stories from two professional news providers, and Google Trends, all of which are useful in analyzing the asset price dynamics of the S&P 100 companies. We applied a bag-of-words approach to detect the sentiment of news stories and introduced a set of negations with the aim of generalizing the method in order to extract the sentiment of any type of financial text. Then we built a set of news measures that provide natural proxies for the information used by heterogeneous market players and for retail investors' attention.

Our empirical results validate the MDH, showing the relevance of news in explaining volatility. Macro-news and EPS are the most important drivers of volatility, followed by news stories and Google Trends. The topics of news stories that are most relevant in affecting volatility are earnings and upgrades/downgrades, but the rest of the news is also influential. Aggregating information over various time horizons and looking at variations of the volume of information across time helps to explain volatility. By including news-based information, we substantially improve volatility forecasting.

Future research should develop a more refined sentiment-detection technique and study the relationship between news and intraday asset price dynamics.

Chapter 2

News Indicators and Intraday Jumps

Francesco Poli

2.1 Introduction

Given the strong evidence for the presence of jumps in financial markets and the reactions in financial markets to firm-specific and macroeconomic news announcements, the literature recently started to investigate the extent to which news coincide with statistically significant intraday jumps. Early papers have conjectured that jumps are caused by the arrival of important new information, most often specific to the firm, and occasionally more general economic or market news. Examining the possible explanations for jumps helps in better explaining market phenomena and improving pricing models. We study the relation between news and jumps, and concentrate on the following questions: what is the likelihood that a news release causes a jump and what proportion of jumps are associated with each type of news? Does information such as sentiment of news stories and surprises from expectations of earnings and macro-announcements have an impact on the probability of jump occurrence? Does the economic importance of jumps, in terms of returns predictability and volatility persistence, change on the basis of the type of news provoking jumps?

We try to answer to the previous questions by using a unique database containing firm-specific news stories of the S&P 100 components released by two news providers – Factset-StreetAccount and Thomson Reuters-Thomson One –, the companies' earnings per share (EPS) announcements, and 23 macroeconomic announcements. News stories are assigned a topic by the provider and their sentiment is extracted with a text-analysis technique based on the method of Loughran and McDonald (2011), see Caporin and Poli (2017). With regard to EPS and macro-announcements, we employ the standardized surprises. All news report date with minute-precision time.

In Section 2.4 we identify the precise intraday intervals at which price jumps of the stocks occur relying on the method of Andersen et al. (2007c), and apply the following modification: when requested by the procedure, instead of the realized bipower variation of Barndorff-Nielsen and Shephard (2004, 2006) we use the corrected threshold bipower variation of Corsi, Pirino and Renò (2010), that was shown to be more accurate for the estimation of jumps. Data about both news and stock prices ranges from February 2005 to February 2015.

In Section 2.5 we perform a matching analysis based on the coincidences of news and jumps, separately for all topics of news stories, EPS, and all macro-announcements. From this analysis it is possible to say that EPS, FOMC rate decisions and news stories classified as *top* by Thomson Reuters represent potentially very useful information to determine the causes of jumps.

News contain a lot of information, from sentiment of news stories to surprises from expectations of EPS and macro-announcements, and this information may be crucial in determining jumps. In addition, heterogeneous market players can react with differing speeds to news. With the aim of reconstructing the different portions of information assimilated by heterogeneous market players, in Section 2.6 we build more than 1,500 news-related variables, by considering: 1) various time horizons; 2) a series of concepts for news stories; 3) the standardized surprises obtained from EPS and macro-announcements. We apply Elastic Net as penalized maximum likelihood estimation method to a logistic regression linking the probability of intraday jumps occurrence to the news indicators. To our knowledge, no previous study employs this method to investigate the relation between news and jumps. This method confirms the results of the matching analysis and detects further elements: in addition to FOMC rate decisions, three other macro-announcements emerge as potential determinants of jumps: federal budget, natural gas stocks and ECRI. For them, both announcements per se and surprises (both above and below expectations) count. News stories released by StreetAccount are also a potential cause of jumps. News releases per se as well as positive and negative news count, and all topics are relevant, especially *M&A* and *Earnings Related*. The event in which news stories with opposite sentiment are sequentially released – that we

call *sentiment inversion* – is also relevant. Both macro-announcements and news stories are likely to increase the probability of jumps in the following hour, but when news are negative jumps occur within at most five minutes.

In Section 2.7, by separating jumps on the basis of the main types of news to which they can be associated – news stories, EPS, macro, and absence of news – we investigate: 1) the relative importance of jumps to asset return dynamics in terms of return predictability and volatility persistence, at both high frequency and daily level, and 2) the exposure of future returns to various measures of jump risk. We find effects on returns and volatility at both high frequency and daily level, and that these effects vary on the basis of the type of news to which jumps are associated. Finally, we find that future quarterly and yearly returns seem to be negatively exposed to three jump risk measures called Jump Intensity, Jump Mean and Jump Volatility, but only when these measures are built using jumps related to macro-announcements. Future returns seem also to be negatively exposed to Jump Volatility when it is built using only positive jumps related to news stories.

2.2 Literature Review

Many recent studies show that stochastic volatility models that include a discontinuous jump term in addition to a diffusive component are better able to capture the empirical characteristics of equity returns. The literature has then begun studying how financial prices respond to public news announcements. For price jumps in stocks and indexes, the literature supports two main explanations: first, price discontinuities are likely the result of uncertainty resolution associated with the release of new and relevant information, or news, as argued by Maheu and McCurdy (2004); the second explanation for the cause of jumps is a local lack of liquidity on the market.

Andersen et al. (2007b) characterize the high frequency response of US, German and British stock, bond and foreign exchange (FX) markets to real-time US macroeconomic news and find that asset price dynamics are linked to fundamentals. Lee and Mykland (2008) find that individual stock jumps are associated with prescheduled earnings announcements and other company-specific news events, while S&P 500 jumps are instead associated with general market news announcements. Bollerslev, Law and Tauchen (2008) examine the relationship between jumps in individual stocks and jumps in an aggregate market index. They show that firm-specific news events are the dominant cause in terms of immediate price impact at the individual stock level, and find a strong tendency for the stocks to move sharply together, i.e. cojump, around 10 am Eastern time, corresponding to the regularly scheduled release-time for many macroeconomic announcements. Beine et al. (2007) explore the role of central bank interventions as a potential source for jumps in the FX market. Larkin and Ryan (2008) employ news sentiment generated from the Dow Jones network with news stories being classified as either positive, negative or neutral in relation to a particular market or sector of interest, and use genetic programming to predict large intraday price jumps on the S&P 500. Rangel (2011) examine the effects of macroeconomic releases on stock market volatility through a Poisson-Gaussian-Garch process with time-varying jump intensity developed by Maheu and McCurdy (2004), which is allowed to respond to such information, and find evidence on macroeconomic variables relevance in explaining jump dynamics and improving volatility forecasts on event days. Lahaye, Laurent and Neely (2011) identify jumps and cojumps from stock index futures, US bond futures, and exchange rates, and relate the dynamics of these discontinuities, in terms of jumps frequency and timing, to US macroeconomic releases. Evans (2011) investigates the association of US macroeconomic news announcements with intraday jumps in US equity, bond and FX markets, and find that approximately one third of jumps corresponds to news and that news-related jumps are larger, on average, in absolute terms than jumps not related to news. Gloß-Klußmann and Hautsch (2011) analyze to which extent high frequency movements in returns, volatility and liquidity of 39 stocks traded at the London Stock Exchange can be explained by the underlying nonscheduled news arrivals, employing the trading signals from the *Reuters NewsScope Sentiment Engine*[1]. They find that high frequency trading activity significantly reacts to news items which are identified as relevant, and that the strongest effects are in terms of volatility and cumulative trading volumes. Miao et al. (2014) examine the influence of macroeconomic news on intraday jumps of the S&P 500 futures and document a strong association: over 75% of the jumps between 8:30 am and 8:35 am and over 60% of the jumps between 10:00 am and 10:05 am are related to news released at 8:30 am and 10:00 am, respectively. Huang (2015) separates US equity and bond market responses into continuous volatility effects and jumps, and finds a larger proportion of days with jumps within macroeconomic announcement days. Caporin, Kolokolov and Renò (2016) develop a test for multivariate jumps (multi-jumps) and detect them in a panel of US stocks. They interpret multi-jumps as systemic events affecting the market on a whole, and associate these rare but statistically and economically important events to relevant market-wide financial, political and (mainly) economic news.

With regard to the second explanation for jumps, i.e. the lack of liquidity, Madhavan (2000) argue that an inefficient provision of liquidity caused by an imbalanced market microstructure can cause extreme price

[1]This service automatically classifies firm-specific news according to positive, neutral and negative author sentiments based on linguistic pattern recognition techniques and provides numeric indicators classifying the relevance of news as well as their novelty.

movements. Bouchaud et al. (2006) and Joulin et al. (2008) call such an event "relative liquidity". Bajgrowicz et al. (2016) relate jumps of the 30 DJIA stocks to macroeconomic news, prescheduled company-specific announcements and stories from news agencies which include a variety of unscheduled and uncategorized events, and find that the majority of news do not cause jumps but may generate bursts of volatility. For them, the main reason for the small impact of news is that managers strategically shift important announcements outside market hours, and liquidity pressures are probably an important factor of jumps.

Both interpretations are provided by Boudt and Petitjean (2014), who study the dynamics of liquidity and news releases around jumps for the 30 DJIA constituents. They retrieve all macroeconomic news announcements, prescheduled or not, and all firm-specific news provided by the Dow Jones and Reuters News Service, and match 1/3 of jumps with macroeconomic news, 5% with firm-specific news and more than 50% with liquidity variations. Firm-specific news events have a higher effect on jump magnitude with respect to macroeconomic news and are identified as the dominant factor in terms of their impact on the occurrence of jumps at the individual stock level. Interestingly, jumps are mostly driven by variations in the demand for immediacy, which is amplified by news, and to market's inability to absorb them without moving the price significantly.

The relation between jumps and news has been studied for other types of assets, as well. With regard to Treasury markets, Dungey et al. (2009) relate jumps and cojumps in the term structure of the US Treasury market and macroeconomic news, Jiang et al. (2011) investigate the causes of jumps in the US Treasury bond prices, examining the relative importance of macroeconomic announcements versus liquidity shocks, and Cui and Zhao (2015) investigate the impact of macro-news announcements on intraday jumps of the China's Treasury bond market. For interest rates, León and Sebestyén (2012) propose surprise measures for the ECB monetary policy, and relate them to variation and jumps of interest rates in the euro area, Bjursell et al. (2013) study the response of jumps and trading activity in interest rates futures markets to macroeconomic announcements, and Meurer et al. (2015) assess the extent to which monetary policy surprises drive jumps in interest rates in the Brazilian interbank market. About foreign exchange, Neely (2011) provides an extensive review on research in FX volatility reaction to macro-announcements, Dewatcher et al. (2014) examine the intra-day effects of verbal statements and comments from monetary officials and policy-makers on euro-dollar exchange rate volatility and jumps, Chatrah et al. (2014) investigate the impact of macro-news on currency jumps and cojumps across Europe, Japan and the US, and Frömmel et al. (2015) examine the link between jumps in the HUF/EUR FX market, macroeconomic news and other news potentially relevant to the exchange rate determination. Mizrach (2012) analyze jumps and cojumps in subprime home equity derivatives and their link with news about Federal Reserve actions, news on firms with securities in an index of CDS on subprime mortgage-backed securities, and macroeconomic news. Elder et al. (2013) relate economic news and jumps in oil prices. Borovkova and Mahakena (2015) investigate the impact of news sentiment on returns, jumps and volatility of natural gas futures.

2.3 Dataset

The dataset includes 1-min price data of the S&P 100 stocks. The work is based on continuously compounded 5-min returns, which are calculated as $\log(p_j/p_{j-1}) \cdot 100$, where p_j denotes the price at the end of the jth 5-min interval. Data ranges from February 4, 2005 to February 25, 2015.

We use the news database of Caporin and Poli (2017), where from two news providers – Factset-StreetAccount and Thomson Reuters-Thomson One – we collect firm-specific news stories and earnings per share (EPS) announcements of eighty-nine stocks among the S&P 100 constituents. In this study, we include 23 macroeconomic announcements.[2] As for prices, data ranges from February 4, 2005 to February 25, 2015 and all news report date with minute-precision time. News stories are assigned a topic by the providers – Thomson Reuters assigns also a level of importance – and their sentiment is extracted with a text-analysis technique which consists in the method of Loughran and McDonald (2011) modified in relation to negation scopes, see Caporin and Poli (2017). With regard to EPS and macro-announcements, we employ the standardized surprises. In the following, the news database is described:

- **News Stories**

  StreetAccount news stories report a topic, while Thomson Reuters news stories report both a topic and a level of importance. In addition, the database contains the sentiment of news.

  We use seven topics from StreetAccount and six topics from Thomson Reuters. Thomson Reuters news stories' level of importance is assigned by the provider on the basis of the expected effect that the event will have on the company's operational and/or financial performance. The levels of importance are four: *low*, *medium*, *high* and *top*, and each level consists in a filter which eliminates all news with a lower level, e.g. *low* gives all news and *medium* gives all news tagged with *medium*, *high* and *top*. For ease of illustration,

[2] In Caporin and Poli (2017) we use 10 macroeconomic announcements, most of which are released when the stock market is closed. Here, we substitute them with 23 macro-announcements released only when the stock market is open.

in the following we report the levels of importance as if they were topics.[3] Table 2.1 lists the topics for each provider.

Table 2.1: Topics list by provider.

StreetAccount	**Thomson Reuters**
all	all
earnings related	earnings pre-announcements
M&A	M&A
litigation	litigation
regulatory	regulatory/company investigation
newspapers	financial
up/downgrades	medium
	high
	top

The so-called sentiment is an indicator of whether the content of a document is good, bad or neutral in relation to the issue it talks about. In Caporin and Poli (2017), we develop a sentiment extraction technique based on the work of Loughran and McDonald (2011) but improved with respect to negation scopes. Loughran and McDonald (2011) procedure consists in counting the words belonging to two lists, which are suited for financial texts, and which are associated to the categories positive and negative respectively. Dealing with US companies 10-Ks filings, they account for negation only for six words and only if these words precede a word classified as positive. We generalize their procedure with the aim of making it appropriate for news created by news providers, that are shorter and less formal. We invert the sentiment each time a word, irrespective of whether it is positive or negative, is preceded by a negation and, as a further improvement, we use 28 single words, 24 sequences of two words and 6 sequences of three words. This modification allows to extract the sentiment of a text with more confidence and independently of the type, length and audience of a financial text. See Caporin and Poli (2017) for details.

- **Earnings Announcements**

  StreetAccount News Stories with topic *earnings related* report the quarterly EPS announcements of each company along with their consensus forecast, given by the mean of a set of surveys at the time of reporting. It is possible to compare the two figures and to determine whether the company has met, exceeded or fallen short of the street's expectations.

- **Macroeconomic Announcements**

  From Thomson Reuters, a series of 23 US macroeconomic announcements released during market trading time is available. They are listed in Table 2.2. As for EPS, both released figure and their consensus forecast are available.

2.4 Intraday Jumps Estimation

Diffusive stochastic volatility models have problems in explaining behaviour of asset prices, especially during market crashes and in general during turbulent periods, since they would require sometimes a volatility level too high for their formulation. As a solution, the total daily return variability has been decomposed into its continuous and discontinuous components based on the bipower variation measures developed by Barndorff-Nielsen and Shephard (2004, 2006). The empirical results in Andersen et al. (2007a) suggest that most of the predictable variation in the volatility stems from the strong own dynamic dependencies in the continuous price path variability, while the predictability of jumps is typically minor.

After filtering the periodic component of intraday volatility through the technique of Boudt et al. (2011), we rely on the method of Andersen et al. (2007c) to identify the precise intraday intervals at which jumps occur, and apply the following modification: when requested by the procedure, instead of the realized bipower variation we use the corrected threshold bipower variation of Corsi, Pirino and Renò (2010), that was shown to be more accurate for the estimation of jumps.

[3]While it is possible to filter Thomson Reuters news stories by both topic and level of importance, we only apply the filter by topic or by importance, and obtain 6 topics plus 4 levels of importance. We avoid to combine filters, which would yield 6 x 4 = 24 combinations.

Table 2.2: Macroeconomic announcements list.

Announcement	**Release Time**
Business Inventories	10:00
Chicago PMI	09:45/10:00
Construction Spending	10:00
Consumer Confidence	10:00
Consumer Credit	15:00
Michigan Consumer Sentiment Index	09:45/09:55/10:00
EIA Crude Oil Stocks	10:30
ECRI Weekly	10:30
IBD Economic Optimism	10:00
Employment Trends Index	10:00
Existing Home Sales	10:00
Factory Orders	10:00
Federal Budget	14:00
FOMC Rate Decisions	12:30/14:00/14:15
NAHB Housing Market	10:00/13:00
Leading Index	10:00
ISM Manufacturing Index	10:00
EIA Natural Gas Stocks	10:30
New Home Sales	10:00
New York NAPM Index	09:45
Pending Home Sales	10:00
Philadelphia Fed Business Index	10:00/12:00
Wholesale Inventories	10:00

We assume that the scalar logarithmic asset price follows a standard jump-diffusion process

$$dX_t = \mu_t dt + \sigma_t dW_t + dJ_t \tag{2.1}$$

where μ_t is predictable, σ_t is cadlag, $dJ_t = c_t dN_t$ where N_t is a non-explosive Poisson process whose intensity is an adapted stochastic process λ_t, the times of the jumps are $(\tau_j)_{j=1,\dots,N_t}$ and c_j are i.i.d. adapted random variables measuring the size, which is always positive, of the jump at time τ_j.

Quadratic variation of the process over a time window T, e.g. one day, is defined as

$$[X]_t^{t+T} = X_{[t+T]}^2 - X_t^2 - 2\int_t^{t+T} X_{s-} dX_s \tag{2.2}$$

where t indexes the day. It can be decomposed into its continuous and discontinuous component

$$[X]_t^{t+T} = [X^c]_t^{t+T} + [X^d]_t^{t+T} \tag{2.3}$$

where $[X^c]_t^{t+T} = \int_t^{t+T} \sigma_s^2 ds$ and $[X^d]_t^{t+T} = \sum_{j=N_t}^{N_t+T} c_j^2$. To estimate these quantities, the time interval $[t, t+T]$ is divided into n subintervals of length $\Delta = T/n$ and the evenly sampled returns are defined as

$$\Delta_{j,t} X = X_{j\Delta+t} - X_{(j-1)\Delta+t}, \qquad j = 1, \dots, n \tag{2.4}$$

The quadratic variation process and its separate components are, of course, not directly observable. Instead, we resort to popular model-free non-parametric consistent measures, including the familiar realized variance

$$RV_\Delta(X)_t = \sum_{j=1}^{n} (\Delta_j X)^2 \tag{2.5}$$

which converges in probability to $[X]_t^{t+T}$ as $\Delta \to 0$

The theory discussed above hinges on the notion of increasingly finer sampled high frequency returns but, in practice, the sampling frequency is limited by the actual quotation or transaction frequency and the observed prices are contaminated by market microstructure frictions, including price discreteness and bid-ask spreads, which render the assumption of a semimartingale price process invalid at the tick-by-tick level. In response to

this, we follow a relevant strand of the literature and compute our daily realized variance and jump measures from five-minute returns, using the nearest preceding or concurrent price to each five-minute mark.

In order to separately measure the jump part, we rely on the corrected threshold bipower variation (*C-TBPV*) measure, a version of the corrected threshold multipower variation (*C-TMPV*) developed by Corsi, Pirino and Renò (2010), which consists in turn in a modification of the realized bipower variation (*BPV*) of Barndorff-Nielsen and Shephard (2004, 2006)

$$\begin{aligned} C\text{-}TBPV_{\Delta}(X)_t &= \mu_1^{-2}\, C\text{-}TMPV_{\Delta}(X)_t^{1,1} \\ &= \mu_1^{-2} \sum_{j=2}^{[T/\Delta]} Z_1(\Delta X_j, \vartheta_j) Z_1(\Delta X_{j-1}, \vartheta_{j-1}) \end{aligned} \tag{2.6}$$

where $\mu_\alpha = E(|Z|^\alpha)$ for $Z \sim N(0,1)$.

The corrected threshold multipower variation is defined as

$$C\text{-}TMPV_{\Delta}(X)_t^{[\gamma_1,\ldots,\gamma_M]} = \Delta^{1-\frac{1}{2}(\gamma_1+\ldots+\gamma_M)} \sum_{j=M}^{[T/\Delta]} \prod_{k=1}^{M} Z_{\gamma k}(\Delta_{j-k+1}X, \vartheta_{j-k+1}) \tag{2.7}$$

the function $Z_\gamma(x,y)$ is

$$Z_\gamma(x,y) = \begin{cases} |x|^\gamma & \text{if} \quad x^2 \le y \\ \frac{1}{2N(-c_\vartheta)\sqrt{\pi}} (\frac{2}{c_\vartheta^2} y)^{\frac{\gamma}{2}} \Gamma(\frac{\gamma+1}{2}, \frac{c_\vartheta^2}{2}) & \text{if} \quad x^2 \ge y \end{cases} \tag{2.8}$$

where $N(x)$ is the standard normal cumulative function, $\Gamma(\alpha, x)$ is the upper incomplete gamma function, $\vartheta = c_\vartheta^2 \sigma^2$ and σ^2 is the variance of $\Delta_j X$ under the assumption that $\Delta_j X \sim N(0, \sigma^2)$. Following Corsi, Pirino and Renò (2010), we set $c_\vartheta = 3$.

As $\Delta \to 0$, C-TBPV converges to $\int_t^{t+T} \sigma^2(s) ds$

The difference between the realized variance and the corrected threshold bipower variation consistently estimates the part of the quadratic variation due to jumps

$$RV_{\Delta}(X)_T - C\text{-}TBPV_{\Delta}(X)_T \xrightarrow[\Delta \to 0]{P} [X^d]_t^{t+T} \tag{2.9}$$

As $\Delta \to 0$, the test statistic

$$C\text{-}T_Z = \Delta^{\frac{1}{2}} \frac{(RV_{\Delta}(X)_T - C\text{-}TBPV_{\Delta}(X)_T) \cdot RV_{\Delta}(X)_T^{-1}}{\sqrt{\left(\frac{\pi^2}{4} + \pi - 5\right) \max\left(1, \frac{C\text{-}TTriPV_{\Delta}(X)_T}{(C\text{-}TBPV_{\Delta}(X)_T)^2}\right)}} \tag{2.10}$$

where C-TTriPV$_\Delta(X)_T$ is a quarticity estimator, see Corsi, Pirino and Renò (2010), and is asymptotically standard normally distributed under the null hypothesis of no jumps.

Based on the above jump detection test statistic, the realized measure of the jump contribution to the quadratic variation of the logarithmic price process is then measured by

$$\widehat{J}_t = I_{(\text{C-T}_Z > \Phi_\alpha)} \cdot (RV_t - BPV_t)^+ \tag{2.11}$$

where $I_{(\cdot)}$ denotes the indicator function and Φ_α refers to the appropriate critical value from the standard normal distribution.

Consequently, the realized measure for the integrated variance is

$$\widehat{C}_t = RV_t - \widehat{J}_t \tag{2.12}$$

The method described above is useful to isolate days containing at least one jump. We want, instead, to identify the precise intraday intervals at which jumps occur, and rely on the procedure of Andersen et al. (2007c).

They define a randomly selected intraday return as $\Delta_\xi X = \sum_{j=1}^{T/\Delta} \Delta_j X \cdot I(\xi = j)$, where ξ is an independently drawn index, uniformly distributed, from the set $\{1, 2, \ldots, T/\Delta\}$. It is identified as jump if its absolute value is higher than an appropriately scaled realized bipower variation. Assuming that intraday scaled returns are distributed as $\Delta^{-1/2} \cdot \Delta_\xi X \sim N(0, IV_t)$, where $IV_t = \int_t^{t+T} \sigma^2(s) ds$ is the daily integrated variance of day t, Andersen et al. (2007c) use the realized bipower variation for its empirical counterpart, such that randomly drawn intraday diffusive returns are distributed approximately as $N(0, \Delta \cdot BV_t)$. Multiple intraday jumps are detected by:

$$c_j = \Delta_j X \cdot I\left[|\Delta_j X| > \Phi_{1-\beta/2} \cdot \sqrt{\Delta \cdot BV_t}\right], \qquad j = 1, 2, \ldots, \frac{T}{\Delta} \tag{2.13}$$

where $\Phi_{1-\beta/2}$ is the appropriate critical value from the standard normal distribution and $\Delta = 1/78$, corresponding to a partition of the length of the market trading day, which is open from 9:30 to 16:00, into seventy-eight 5-min intervals. Following Andersen et al. (2007c), we choose the size of the jump test at the daily level $\alpha = 10^{-5}$ in eq. (2.11) and define $\beta = 1 - (1-\alpha)^{\Delta} = 1.28 \cdot 10^{-7}$.

Instead of the realized bipower variation of Barndorff-Nielsen and Shephard (2004, 2006) we use the corrected threshold bipower variation (*C-TBPV*) of Corsi, Pirino and Renò (2010), which was shown to be more accurate for the estimation of jumps. We identify, therefore, intraday jumps by:

$$c_j^* = \Delta_j X \cdot I\left[|\Delta_j X| > \Phi_{1-\beta/2} \cdot \sqrt{\Delta \cdot C\text{-}TBPV_t}\right], \qquad j = 1, 2, \ldots, \frac{T}{\Delta} \tag{2.14}$$

following Corsi, Pirino and Renò (2010), we set $c_\vartheta = 3$ when requested, see their paper for details.

2.5 Matching Analysis

Figure 2.1 presents the time series over the whole sample of the number of jumps, the median of the absolute jump size, the number of StreetAccount news stories, and the number of Thomson Reuters news stories.[4] All sums are taken over all assets.

We can notice a number of jumps higher than usual from the end of 2007 to the beginning of 2009. The absolute size of jumps is higher than usual during the same period, peaking at the beginning of 2009. It is not surprising to see an intense activity during the global financial crisis (GFC), which extends to the period December 2007 – June 2009. The number of news stories released by both StreetAccount and Thomson Reuters is slightly decreasing over the sample, and we interpret this as a result of an increasing selection endeavour by the providers, which are interested in the release of relevant news only. In addition, from the end of 2008 to the beginning of 2010, that is from the beginning of the GFC until one year after its end, the number of news stories released by Thomson Reuters is higher than in the rest of the sample, possibly as a consequence of an unusual high attention raised by the crises. It is not possible to infer any relation between jumps and news stories from this graph.

[4] Earnings and macro-announcements not are reported since they are released periodically and their frequency is constant.

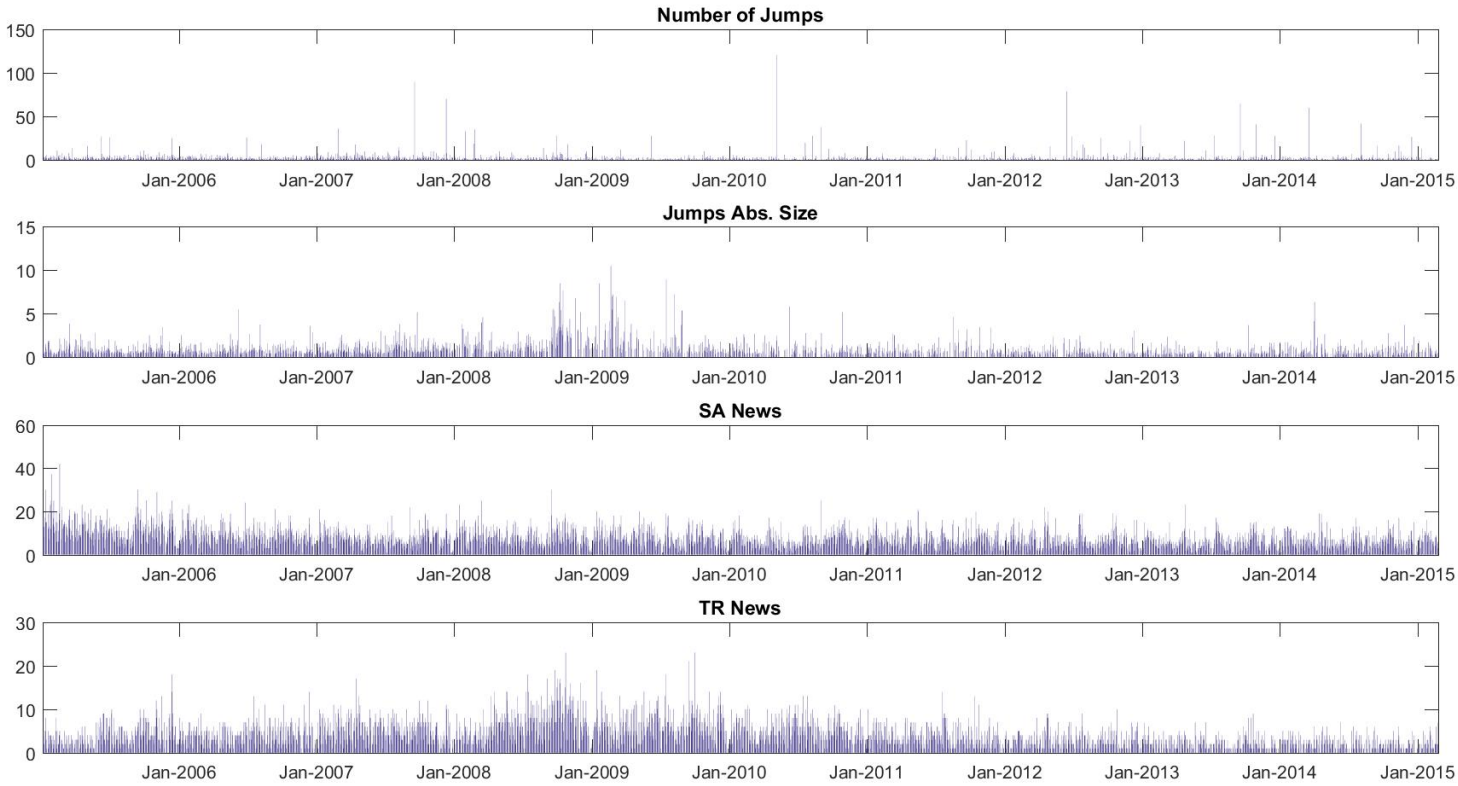

Figure 2.1: N. of jumps, median of abs. jumps size and n. of news stories over the sample
Notes: for each day, the figure shows the sum of the number of jumps, the median of the absolute jump size, the sum of StreetAccount news stories, and the sum of Thomson Reuters news stories. The number of jumps and the number of news stories are summed over all assets, while the median of the absolute jump size is computed across the jumps of all assets.

Figure 2.2 reports the distribution of the jump frequency and the jump absolute size by intraday interval. The jump frequency is unusually high during the opening and closing times of the market, and it peaks at the following times: 10:00-10:05, 14:00-14:05, 14:15-14:20, 14:20-14:25, 15:05-15:10. The absolute jump size is characterized by a U-shaped distribution over the day, and shows peaks at the times 14:10-14:15 and 14:20-14:25.

It is not possible to define a clear relation between jumps and macro-announcements at this stage, but we can notice that some intervals characterized by an unusual high jump frequency or size follow the release of macro-news: at 10:00 sixteen out of twenty-three announcements are released, and at 14:00 and 14:15 FOMC rate decisions are communicated to the market.

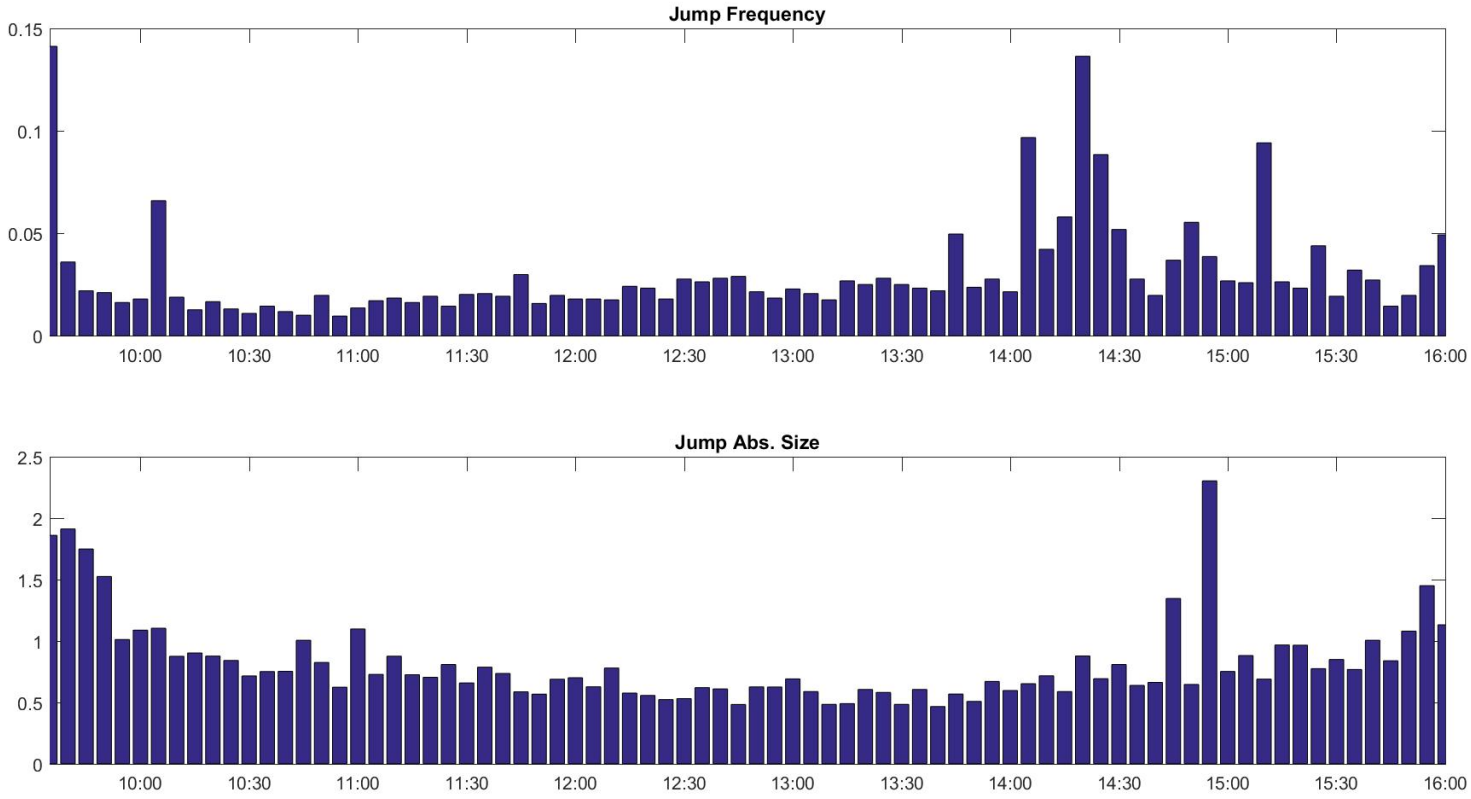

Figure 2.2: Frequency of jumps (in %) and median of abs. jump size for each intraday interval.

We define a news-jump coincidence as the occurrence of a jump in a 5-min interval and the release of a news in the same or in the preceding interval. This definition of coincidence takes into account the possibility that news cause jumps after few minutes, as well as almost instantaneously.

We use three metrics to analyze the matching of news and jumps:

- $P(J|N)$: (number of news-jump coincidences)/number of news
- median$(J|N)$: median of absolute size of jumps coincident with news
- $P(N|J)$: (number of news-jump coincidences)/number of jumps

These metrics are widely used, and give different types of information. $P(J|N)$ describes the likelihood that a news release causes a jump, while $P(N|J)$ measures what proportion of jumps are associated with a particular type of news.

Tables 2.3–2.7 report the cross-sectional average of the metrics. In Table 2.3 we compare StreetAccount news stories (topic *all*, corresponding to all news), Thomson Reuters news stories (topic *all*), EPS, macro-announcements (all), and *No News*, which consists in the lack of news of any kind. In tables 2.4–2.7 we compare StreetAccount news stories topics, Thomson Reuters news stories topics, and macro-announcements.

From Table 2.3 we see that EPS is the type of news, with a $P(J|N)$ of 9.85%, which causes a jump with the highest probability. With a median$(J|N)$ of 2.24, EPS also cause the jumps with the highest absolute value, which is greater for negative jumps. In terms of $P(J|N)$, EPS are followed in descending order by StreetAccount news stories, Thomson Reuters news stories, macro-announcements and *No News*. By the way, EPS are announced rarely with respect than the other news and, as a consequence, their $P(N|J)$ (the proportion of jumps associated with them) is the lowest one, equal to 0.17%. StreetAccount news stories have both a higher $P(J|N)$ and a higher $P(N|J)$ with respect to Thomson Reuters news stories. It seems that StreetAccount news stories are more relevant than Thomson Reuters ones in causing jumps, but we are going to present a more detailed analysis in tables 2.4 and 2.5. For *No News*, $P(N|J)$ is 79.75%, revealing that the majority of jumps is not associated with news. Jumps which cannot be related to news may be caused by a lack of liquidity in presence of an excessive demand for trading. For news-related jumps, it is interesting to extend the analysis by furtherly

classifying news. As a first step, we now look at news stories' topics and at single macro-announcements. In section 2.6 we will consider additional information.

Table 2.3: Main sources of news and jumps matching.

News		**All**	**Pos**	**Neg**
EPS	$P(J\|N)$	9.85	5.40	4.45
	median$(J\|N)$	2.24	1.24	2.36
	$P(N\|J)$	0.17	0.07	0.09
StreetAccount	$P(J\|N)$	0.91	0.55	0.36
News Stories	median$(J\|N)$	1.36	1.41	1.55
	$P(N\|J)$	5.86	3.46	2.39
Thomson Reuters	$P(J\|N)$	0.44	0.29	0.14
News Stories	median$(J\|N)$	1.50	1.37	1.78
	$P(N\|J)$	1.38	1.02	0.35
Macro	$P(J\|N)$	0.11	0.08	0.04
Announcements	median$(J\|N)$	0.93	0.91	1.01
	$P(N\|J)$	14.03	9.49	4.54
No News	$P(J\|N)$	0.02	0.01	0.01
	median$(J\|N)$	0.72	0.75	0.70
	$P(N\|J)$	79.75	41.36	38.40

Notes: $P(J|N)$, median$(J|N)$ and $P(N|J)$ for (all) StreetAccount news stories, (all) Thomson Reuters news stories, (all) macroeconomic announcements, and EPS. News are sorted in descending order by $P(J|N)$. The last three columns show the metrics distinguishing: all jumps, positive jumps, negative jumps. Numbers are expressed in percentages.

With regard to StreetAccount news stories, Table 2.4 shows that the topic *Newspapers* is the one characterized by the highest $P(J|N)$, equal to 1.55%, and also by the highest median$(J|N)$, equal to 2.67. The topic *M&A* is the second in terms of $P(J|N)$, equal to 0.93%, and *all* follows, therefore it seems that for StreetAccount news stories the topics *Newspapers* and *M&A* help to filter information potentially causing jumps. *Earnings Related*, with a $P(N|J)$ equal to 1.20% which is second only to the $P(N|J)$ of *all* news stories, even with a probability to cause a jump lower than *Newspapers* and *M&A*, is the topic mostly associated to jumps.

With regard to Thomson Reuters news stories, Table 2.5 shows an interesting very high $P(J|N)$ for the topic *Top*, which is equal to 7.79%. Thomson Reuters *Top* news stories' $P(J|N)$ is also higher than the $P(J|N)$ of any topic of both providers and of any macro-announcement, and is only lower than EPS' $P(J|N)$. The other topics of Thomson Reuters news stories with a $P(J|N)$ higher than the topic *all* are, in descending order: *Earnings Pre-Announcements*, *Financial*, *High*, and *Medium*. *High* and *Medium* news stories are also the ones mostly associated to jumps, with a $P(N|J)$ equal to 1.15% and 1.17%, respectively. In terms of median$(J|N)$, instead, jumps associated to *Litigation* and *M&A* have the highest absolute size. It seems that Thomson Reuters is able to effectively classify news stories by importance and that (in addition, as expected, to *Earnings Pre-Announcements*), events that have an impact on the balance sheet of a company – identified by the topic *Financial* – are also relevant in causing jumps.

Finally, with regard to macro-announcements, from Tables 2.6 and 2.7 it is clear that FOMC Rate Decision is the macro-announcement most likely causing jumps with a $P(J|N)$ equal to 3.64%, while the other macro-announcements have a $P(J|N)$ lower than the majority of news stories' topics. FOMC Rate Decision's $P(N|J)$, equal to 9.63%, is remarkably higher than the $P(N|J)$ of any other macro-announcement and of any other type of news, highlighting that it is the type of news associated with the highest proportion of jumps.

It is possible to conclude that, although most of the jumps seem not to be related to news, EPS, FOMC rate decisions and *Top* Thomson Reuters news stories represent potentially very useful information to determine the causes of jumps.

Table 2.4: StreetAccount news stories and jumps matching.

Topic		**All**	**Pos**	**Neg**
Newspapers	$P(J\|N)$	1.55	1.14	0.40
	median$(J\|N)$	2.67	2.06	2.68
	$P(N\|J)$	0.26	0.12	0.14
M&A	$P(J\|N)$	0.93	0.65	0.28
	median$(J\|N)$	1.60	1.86	1.27
	$P(N\|J)$	0.38	0.29	0.09
All	$P(J\|N)$	0.91	0.55	0.36
	median$(J\|N)$	1.36	1.41	1.55
	$P(N\|J)$	5.86	3.46	2.39
Earnings	$P(J\|N)$	0.90	0.35	0.55
Related	median$(J\|N)$	1.41	1.59	1.62
	$P(N\|J)$	1.20	0.54	0.66
Up/Downgrades	$P(J\|N)$	0.74	0.49	0.25
	median$(J\|N)$	1.17	1.23	1.04
	$P(N\|J)$	0.68	0.48	0.19
Litigation	$P(J\|N)$	0.67	0.36	0.31
	median$(J\|N)$	1.78	2.41	1.36
	$P(N\|J)$	0.35	0.21	0.14
Regulatory	$P(J\|N)$	0.65	0.13	0.52
	median$(J\|N)$	1.51	1.79	1.18
	$P(N\|J)$	0.22	0.12	0.10

Notes: $P(J|N)$, median$(J|N)$ and $P(N|J)$ for StreetAccount news stories. The topics are sorted in descending order by $P(J|N)$. The last three columns show the metrics distinguishing: all jumps, positive jumps, negative jumps. Numbers are expressed in percentages.

Table 2.5: Thomson Reuters news stories and jumps matching.

Topic		**All**	**Pos**	**Neg**
Top	$P(J\|N)$	7.79	4.83	2.96
	median$(J\|N)$	1.56	1.14	2.18
	$P(N\|J)$	0.40	0.29	0.11
Earnings	$P(J\|N)$	3.34	1.73	1.61
Pre-Announcements	median$(J\|N)$	1.93	1.89	1.93
	$P(N\|J)$	0.28	0.09	0.18
Financial	$P(J\|N)$	3.03	2.97	0.06
	median$(J\|N)$	0.95	0.96	1.02
	$P(N\|J)$	0.73	0.69	0.04
High	$P(J\|N)$	0.98	0.69	0.29
	median$(J\|N)$	1.59	1.45	1.88
	$P(N\|J)$	1.15	0.83	0.32
Medium	$P(J\|N)$	0.86	0.62	0.24
	median$(J\|N)$	1.59	1.46	1.88
	$P(N\|J)$	1.17	0.85	0.32
All	$P(J\|N)$	0.44	0.29	0.14
	median$(J\|N)$	1.50	1.37	1.78
	$P(N\|J)$	1.38	1.02	0.35
Litigation	$P(J\|N)$	0.26	0.26	0.00
	median$(J\|N)$	3.56	3.56	-
	$P(N\|J)$	0.05	0.05	0.00
M&A	$P(J\|N)$	0.20	0.08	0.12
	median$(J\|N)$	2.62	3.33	1.54
	$P(N\|J)$	0.11	0.06	0.05
Regulatory	$P(J\|N)$	0.00	0.00	0.00
	median$(J\|N)$	-	-	-
	$P(N\|J)$	0.00	0.00	0.00

Notes: $P(J|N)$, median$(J|N)$ and $P(N|J)$ for Thomson Reuters news stories. The topics are sorted in descending order by $P(J|N)$. The last three columns show the metrics distinguishing: all jumps, positive jumps, negative jumps. Numbers are expressed in percentages.

Table 2.6: Macro-announcements and jumps matching 1/2.

Announcement		**All**	**Pos**	**Neg**
FOMC	$P(J\|N)$	3.64	2.48	1.15
Rate Decision	median$(J\|N)$	0.90	0.88	1.01
	$P(N\|J)$	9.63	6.54	3.10
Construction	$P(J\|N)$	0.43	0.41	0.02
Spending	median$(J\|N)$	1.16	1.19	0.83
	$P(N\|J)$	1.84	1.77	0.08
ISM	$P(J\|N)$	0.43	0.41	0.02
Manufacturing	median$(J\|N)$	1.17	1.20	0.83
PMI	$P(N\|J)$	1.85	1.77	0.08
Consumer	$P(J\|N)$	0.06	0.03	0.03
Confidence	median$(J\|N)$	1.14	1.06	1.21
	$P(N\|J)$	0.25	0.11	0.14
Consumer	$P(J\|N)$	0.06	0.02	0.03
Credit	median$(J\|N)$	1.37	1.24	1.44
	$P(N\|J)$	0.21	0.09	0.12
Leading	$P(J\|N)$	0.06	0.00	0.05
Index	median$(J\|N)$	1.18	0.96	1.20
	$P(N\|J)$	0.22	0.02	0.20
Michigan	$P(J\|N)$	0.05	0.04	0.02
Sentiment	median$(J\|N)$	1.08	0.96	1.31
	$P(N\|J)$	0.42	0.30	0.12
Federal	$P(J\|N)$	0.04	0.03	0.01
Budget	median$(J\|N)$	0.57	0.62	0.42
	$P(N\|J)$	0.17	0.14	0.03
Philly Fed	$P(J\|N)$	0.04	0.00	0.04
Business	median$(J\|N)$	1.17	0.96	1.20
Index	$P(N\|J)$	0.18	0.02	0.16
Business	$P(J\|N)$	0.03	0.03	0.00
Inventories	median$(J\|N)$	0.97	0.97	-
	$P(N\|J)$	0.11	0.11	0.00
NAHB	$P(J\|N)$	0.03	0.02	0.01
	median$(J\|N)$	0.70	0.57	0.83
	$P(N\|J)$	0.14	0.07	0.07

Table 2.7: Macro-announcements and jumps matching 2/2.

Announcement		**All**	**Pos**	**Neg**
New	$P(J\|N)$	0.03	0.01	0.02
Home Sales	median$(J\|N)$	0.96	1.12	0.88
	$P(N\|J)$	0.13	0.04	0.09
Existing	$P(J\|N)$	0.02	0.02	0.00
Home Sales	median$(J\|N)$	0.74	0.74	-
	$P(N\|J)$	0.10	0.10	0.00
Natural Gas	$P(J\|N)$	0.02	0.01	0.02
Stocks	median$(J\|N)$	1.14	1.55	0.98
	$P(N\|J)$	0.47	0.11	0.36
Oil	$P(J\|N)$	0.02	0.01	0.01
Stocks	median$(J\|N)$	1.12	1.12	1.09
	$P(N\|J)$	0.39	0.14	0.26
Chicago PMI	$P(J\|N)$	0.01	0.01	0.00
	median$(J\|N)$	0.92	1.05	0.66
	$P(N\|J)$	0.05	0.03	0.02
ECRI Weekly	$P(J\|N)$	0.01	0.00	0.00
Index	median$(J\|N)$	0.78	0.63	0.92
	$P(N\|J)$	0.18	0.09	0.09
Employment	$P(J\|N)$	0.01	0.01	0.00
Trends	median$(J\|N)$	0.37	0.37	-
	$P(N\|J)$	0.03	0.03	0.00
Factory	$P(J\|N)$	0.01	0.00	0.00
Orders	median$(J\|N)$	1.69	1.52	1.86
	$P(N\|J)$	0.03	0.02	0.02
IBD	$P(J\|N)$	0.01	0.01	0.00
Economic	median$(J\|N)$	0.95	1.18	0.50
Optimism	$P(N\|J)$	0.07	0.04	0.03
Pending	$P(J\|N)$	0.01	0.00	0.01
Home Sales	median$(J\|N)$	0.67	-	0.67
	$P(N\|J)$	0.02	0.00	0.02

Notes: $P(J|N)$, median$(J|N)$ and $P(N|J)$ for macro-announcements. The announcements are sorted in descending order by $P(J|N)$. The last three columns show the metrics distinguishing: all jumps, positive jumps, negative jumps. Numbers are expressed in percentages. The two announcements New York NAPM Index and Wholesale Inventories are not shown because there are no news-jump coincidences for them.

2.6 News Indicators and Jumps

News contain a lot of information, from sentiment of news stories to surprises from expectations of EPS and macro-announcements, and this information may be crucial in determining the relation between news and jumps. In addition, heterogeneous market players can react with differing speeds to news. We try to reconstruct the different portions of information assimilated by heterogeneous market players and build more than 1,500 news-related variables. Then, we apply Elastic Net as penalized maximum likelihood estimation method to a logistic regression linking the probability of intraday jumps occurrence to the news indicators, with the aim of identifying when news cause jumps with more detail with respect to the previous section.

2.6.1 News Indicators Creation

In the spirit of Caporin and Poli (2017), we build news-related variables from the dataset. In the following we detail, in order: 1) the time horizons over which information is aggregated; 2) the concepts to be used to build variables from news stories; 3) the standardized surprises obtained from EPS and macro-announcements; 4) the indicators we build for a high frequency analysis of asset price dynamics.

Time Horizons

Starting from the reasoning that market players assimilate and react to news disclosure at differing speeds, we look at how news are released over time with the aim to reconstruct the different portions of information on which the different market players base their decisions. For each 5-min interval from t_0 to t_1 during which a jump may occur, news indicators are built by looking at the information released during five lead-and-lag intervals: three lagged intervals, the contemporaneous interval $[t_0, t_1]$ and a lead interval. The leaking of information may indeed cause capital market participants to act in advance of news announcements and, consequently, it becomes necessary to also consider the effect of lead responses by the market. The intervals are illustrated in Table 2.8.

Table 2.8: Lead-and-lag intervals for news indicators.

Lead-and-Lag	**Start**	**End**
-3	t_0 -60 min	t_0 -30 min
-2	t_0 -30 min	t_0 -10 min
-1	t_0 -10 min	t_0
0	t_0	t_1
+1	t_1	t_1 +10 min

Notes: time intervals on which information is aggregated to build news indicators. t_0 and t_1 are the beginning and the end of each 5-min interval during which a jump may occur. News indicators based on these time horizons cannot be related to jumps occurring during the first 60 minutes and during the last 10 minutes of the trading day.

Concepts for News Stories

We go beyond the standard techniques used to assign numbers to textual information, and identify a set of concepts/events which are based on how news are released over time. Concepts can refer to one or more lags, and each one is peculiar in the reaction it potentially causes in the market:[5]

- **standard**: news occurrence, sentiment;
- **uncertainty**: occurrence of news with opposite sentiment within the same interval;
- **abnormal quantity**: quantity of news over a threshold;
- **news persistence/interaction**: release of news in each of two consecutive intervals;
- **sentiment inversion**: event in which the sentiment of an interval equals the opposite of the sentiment in a previous interval;
- **sentiment conditional on quantity**: sentiment of the reference interval conditional on the occurrence of news during previous intervals. We consider the possibility that investors base their decisions on the sentiment of the reference interval but their attention is raised during previous intervals.

[5]We follow Caporin and Poli (2017), to which we refer for a detailed description of news concepts.

Standardized Surprises of EPS and Macro-Announcements

With regard to earnings announcements, from actual figures and consensus forecasts we compute the SUE (Standardized Unexpected Earnings) score, which measures the number of standard deviations the reported EPS differs from the mean estimates.

$$SUE_t = \frac{EPS_t^{actual} - EPS_t^{forecast}}{\sigma(EPS_t^{actual} - EPS_t^{forecast})} \tag{2.15}$$

where $\sigma(EPS_t^{actual} - EPS_t^{forecast})$ is the standard deviation of $(EPS_t^{actual} - EPS_t^{forecast})$.

With regard to macro-announcements, from reported announcements and consensus forecasts we compute the standardized surprise as for EPS, and we call it Std_Macro.

$$Std\_Macro_t = \frac{Macro_t^{actual} - Macro_t^{forecast}}{\sigma(Macro_t^{actual} - Macro_t^{forecast})} \tag{2.16}$$

where $\sigma(Macro_t^{actual} - Macro_t^{forecast})$ is the standard deviation of $(Macro_t^{actual} - Macro_t^{forecast})$, and $Macro$ stands for any of the macro-indicators.[6]

High Frequency News Indicators

We propose a set of news indicators, suitable to be linked to the occurrence of intraday jumps. They are listed in tables 2.9–2.12. For each stock, we end up with a total of 1,696 news indicators.

Table 2.9: News stories standard indicators.

Variable	**N. Transf.**
STANDARD	
news occurrence flag	1[a]
sentiment	1
positive sent flag	1
negative sent flag	1
UNCERTAINTY	
pos and neg news in same interval	1
ABNORMAL QUANTITY	
n. news ≥ 2	1
total for each topic and lead-lag index	6
grand total (6 x 16 x 5)[b]	**480**

Notes: The first column shows the variables grouped by the concepts that originated them. The second column shows the number of transformations, with the total number of measures obtained at the end of the column.
[a]: When the number of transformations equals 1, the measure consists of a flag (1 for the occurrence of the event, and 0 otherwise).
[b]: There are 16 topics (6 topics for StreetAccount news stories plus 10 topics for Thomson Reuters news stories) and 5 lead-and-lag indexes.

[6]For the three announcements ECRI Weekly, Employment Trends and New York NAPM Index the consensus forecast is not available. For them, we compute the standardized change with respect to the previous release.

Table 2.10: News stories flow indicators.

Variable	**N. Transf.**
NEWS PERSISTENCE/INTERACTION	
news occurrence in 2 consecutive intervals	1
SENTIMENT INVERSION	
sentiment inversion	1
SENTIMENT COND. ON PAST QUANTITY	
pos sent and news occurrence in previous interval	1
neg sent and news occurrence in previous interval	1
total for each topic and lead-lag index	4
grand total (4 x 16 x 4^a)	**256**

^{a}: Differently from news stories' standard indicators, news stories' flow indicators are based on the aggregation of information over consecutive intervals. From 5 lead-and-lag intervals we obtain 4 couples of consecutive intervals.

Table 2.11: EPS indicators.

Variable	**N. Transf.**
SUE	8
total for each lead-lag index	8^a
grand total (8 x 5)	**40**

^{a}: we apply the following 8 transformations: flag for announcement, x, flag if $x \neq 0$, flag if $x > 0$, flag if $x < 0$, $\text{sign}(x) \cdot \sqrt{x}$, $\text{sign}(x) \cdot \log(1 + |x|)$, $\text{sign}(x) \cdot x^2$, where x stands for standardized surprise.

Table 2.12: Macro-indicators.

Variable	**N. Transf.**
Std_Macro	8
total for each macro-announcement and lead-lag index	8
grand total (8 x 23 x 5)	**920**

2.6.2 Indicators Selection from Penalized Logistic Regression

Each indicator is constructed in order to be potentially linked to a market reaction causing jumps and we want to see which indicators are the most useful ones.

We want to describe a dependent binary variable that takes value one for the occurrence of jumps and zero otherwise using as explanatory variables the 1,696 news indicators described above.[7] If we had a smaller number of indicators, we would estimate a standard logistic regression: with a binary random variable y_i that can take the values one and zero with probabilities π_i and $1-\pi_i$, respectively, y_i has a Bernoulli distribution with parameter π_i. The logistic function relates π_i with the explanatory variables x_i and can be written as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta^T x_i \tag{2.17}$$

where the odds of an event happening $\pi/(1-\pi)$ is defined as the probability that the event occurs divided by the probability that the event does not occur. (2.17) corresponds to the maximum likelihood

$$L(\beta_0, \beta) = \sum_{i=1}^{n} \left[y_i \cdot (\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0 + \beta^T x_i}) \right] \tag{2.18}$$

The regression coefficients β represent the amount of change expected in the natural logarithm of the odds $\log(\pi/(1-\pi))$ for a one unit change in each predictor variable x with all the other variables in the model held constant. Note that since the logit is based on natural logs, there is a clear advantage to using the natural logarithm: the coefficient β represents an elasticity of the odds. So, for example, a coefficient $\beta = 2$ means that a 1% increase in x is associated with roughly a 2% increase in the odds of success (in our case, in the odds of jump occurrence).

Dealing with a huge number of regressors, we apply the Elastic Net of Zou and Hastie (2005) as penalized maximum likelihood estimation method (PMLE) to the logistic regression illustrated above. Elastic Net is an estimation method which shrinks and selects parameters, preventing overfitting. Friedman, Hastie and Tibshirani (2010) point out that logistic regression is often plagued with degeneracies when the number of covariates p is greater than the number of observations n and exhibits wild behavior even when n is close to p; Elastic Net alleviates these issues, and regularizes and selects variables as well.

Pavlou et al. (2016) review and evaluate the predictive performance of the main penalized regression methods using real and simulated data, focusing on regression models with low-dimensional data, binary outcome and few events. Their simulation study shows that maximum likelihood estimation tends to produce overfitted models with poor predictive performance in scenarios with few events, and penalized methods can offer improvement.

The objective function for the Elastic Net is

$$(\beta_0, \beta) = \text{argmin} - \left[\frac{1}{N} \sum_{i=1}^{N} y_i \cdot (\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_o + \beta^T x_i}) \right] + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1 \right] \tag{2.19}$$

where $\lambda \geq 0$ is a complexity parameter and $0 \leq \alpha \leq 1$ is a compromise between Ridge ($\alpha = 0$) and LASSO ($\alpha = 1$). λ is automatically selected with 5-fold cross-validation, in order to avoid overfitting, and is set equal to the one minimizing the area under the ROC curve. We set $\alpha = 0.5$.[8] In addition, predictors are scaled by standard deviation as prescribed by Tibshirani (1996), in order to avoid distortion of the shrinkage correction and to allow coefficients comparison.

We apply the estimation procedure to each asset. Tables 2.13–2.15 report the ranking of the indicators which are *selected* – their estimated β is different from zero – for at least 30% of the stocks, and include the percentages of positive and negative estimated coefficients. The tables report the results for the full sample (Feb 2005 – Feb 2015), a contractionary period (Dec 2007 – Jun 2009) and an expansionary period (Jun 2009 – Feb 2015), respectively.[9] From Table 2.13, FOMC rate decisions is the news which gives rise to the two mostly selected indicators: announcements in lags 0 and 1 (first and sixth row) and a lower than expected rate announced in lag 0 (eleventh row) increase the probability of jump. The announcement of a rate different from expectations (third row) has a positive coefficient for one third of the assets and a negative coefficient for two thirds of them, so this indicator has a more uncertain impact on jump probability. Looking at tables 2.14 and 2.15, we see that the probability of jumps is increased during contractions by a higher than expected rate (Table 2.14, fourth

[7] We do not include regressors based on past jumps since jumps are rare events and there are no studies, to our knowledge, documenting jumps probability persistence. Bajgrowicz et al. (2016) use the runs test developed by Mood (1940) to detect clustering of jumps in time for the 30 Dow Jones stocks over the period from January 2006 to December 2008, and their results do not detect time clustering phenomena of jumps arrivals.

[8] Pavlou et al. (2016) show that Ridge regression performs well, except in scenarios with many noise predictors, LASSO performs better than Ridge in scenarios with many noise predictors and worse in the presence of correlated predictors, and Elastic Net performs well in all scenarios.

[9] In Table 2.14 (contractionary period), indicators are reported if they are selected for at least 15% of the stocks.

row), and during expansions by a lower than expected rate (in Table 2.15, from third to sixth row, various sign-preserving transformations of the surprise from expectations have a negative coefficient). The announcement of a federal budget below expectaions in lag 0 (second row) also increases the jump probability. Natural gas stocks announcements in lag -3 (fourth row) has a positive sign for 25.84% of the assets, so this macro-announcement is likely to cause jumps with a delay ranging from half an hour to one hour. ECRI surprises in lag -3 (from seventh to ninth row) have positive and negative coefficients, depending on the transformation, therefore surprises both above and below expectations seem to increase the probability of jumps, with a delay ranging, again, from half an hour to one hour.

The release of StreetAccount news stories with topic *All* in lag 3 (fifth row) has a positive sign for 25.84% of the assets, as well as the release of a negative news in lag 0 (tenth row) for 33.71% of the assets. It seems that StreetAccount news stories, independently of the topic, are likely to cause jumps in the following hour, but markets react much more quickly when news are negative. Finally, the sentiment of *M&A* and *Earnings Related* StreetAccount news stories in lag 0 (last two rows) has also a positive coefficient, and our interpretation is that also news with positive sentiment, for these topics, cause jumps. Indeed, *All* news stories include *M&A* and *Earnings Related* ones, and the positive impact on jump probability of negative news belonging to these two topics were already caught by the regressor associated to negative *All* news stories (tenth row). Subsample analyses tell us something more: news stories released by Thomson Reuters are also likely to increase the probability of jumps, especially the topics *M&A* and *Litigation* (Table 2.13, from fifth to ninth row). In addition, the event called *sentiment inversion* seem also to increase the probability of jumps (Table 2.14, seventh row and Table 2.15, last row).

No indicators based on lead 1 compare on the tables. It is possible that market participants are not able to act in advance of news, or that it happens but not systematically. It is also possible that this event increases volatility instead of the probability of jumps. We leave the study of leaking of information and insider trading detection for future research.

These results tell us a different, however not contradicting, story with respect to the matching analysis conducted in Section 2.5. First, EPS and *Top* Thomson Reuters news stories are not associated with the increase of jump probability. They represent, though, rare events, and the estimation of their β can be biased toward zero when there are no jumps or no news in some of the so-called test sets of k-fold cross-validation. Second, the use of news indicators built on the basis of much more information allows to discover additional potential sources of jumps. Summarizing, FOMC rate decisions are confirmed to be a very important determinant of jumps, along with three other macro-announcements: federal budget, natural gas stocks and ECRI, for which both announcements per se and surprises (both above and below expectations) count. With regard to FOMC rate decisions, jumps probability is increased by higher than expected rates during contractions, and by lower than expected rates during expansions. News stories released by StreetAccount are also a potential cause of jumps. News releases per se as well as positive and negative news count, and all topics are relevant, especially *M&A* and *Earnings Related*. The event of *sentiment inversion* is also relevant. Both macro-announcements and news stories are likely to increase the probability of jumps in the following hour, but when news are negative jumps occur within at most five minutes.

Table 2.13: News indicators selected by penalized logistic regression (full sample).

News Type	**Topic/Macro Ann.**	**Lead-Lag**	**Measure**	**% Selected**	**% Pos.**	**% Neg.**
Macro	FOMC Rate Dec.	0	flag for announcement	98.88	98.88	0.00
Macro	Federal Budget	0	square surp	98.88	0.00	98.88
Macro	FOMC Rate Dec.	0	flag for surp $\neq 0$	95.51	30.34	65.17
Macro	Natural Gas Stocks	-3	flag for announcement	68.54	25.84	42.70
StreetAcc.	All	-3	flag for news release	65.17	25.84	39.33
Macro	FOMC Rate Dec.	-1	flag for announcement	59.55	59.55	0.00
Macro	ECRI	-3	square surp	39.33	0.00	39.33
Macro	ECRI	-3	log surp	39.33	39.33	0.00
Macro	ECRI	-3	sqrt surp	39.33	39.33	0.00
StreetAcc.	All	0	flag for negative sent	33.71	33.71	0.00
Macro	FOMC Rate Dec.	0	flag for surp < 0	30.34	30.34	0.00
StreetAcc.	M&A	0	sentiment	30.34	30.34	0.00
StreetAcc.	Earnings Related	0	sentiment	30.34	30.34	0.00

Notes: Regressors sorted in descending order by the number of assets for which their coefficient estimated with Penalized Logistic Regression is different from zero. The table reports the regressors for which the percentage of assets with estimated coefficients different from zero is higher than 30%, and shows in the last three columns the percentage of assets with: estimated coefficients different from zero, positive estimated coefficients, and negative estimated coefficients. Surp stands for standardized surprise. Flags whose coefficients are negative for most of the assets are discarded, since in this case they do not increase the probability of jumps. Sample = Feb 2005 – Feb 2015 (all sample).

Table 2.14: News indicators selected by penalized logistic regression (contraction).

News Type	Topic/Macro Ann.	Lead-Lag	Measure	% Selected	% Pos.	% Neg.
Macro	FOMC Rate Dec.	0	flag for announcement	61.80	49.44	12.36
Macro	FOMC Rate Dec.	-2	flag for announcement	61.80	49.44	12.36
Macro	FOMC Rate Dec.	-1	flag for announcement	46.07	33.71	12.36
Macro	FOMC Rate Dec.	-3	flag for surp > 0	44.94	44.94	0.00
T. Reuters	M&A	-1	flag for news release	29.21	29.21	0.00
T. Reuters	M&A	-2	flag for news release	29.21	29.21	0.00
T. Reuters	Litigation	-2 and -1	flag for sent inversion	29.21	29.21	0.00
T. Reuters	Litigation	-1	flag for news release	29.21	29.21	0.00
T. Reuters	Litigation	-2	flag for news release	29.21	29.21	0.00
Macro	ECRI	-3	square surp	16.85	16.85	0.00
Macro	ECRI	-3	sqrt surp	16.85	12.36	4.49

Notes: Sample = Dec 2007 – Jun 2009 (contraction).

Table 2.15: News indicators selected by penalized logistic regression (expansion).

News Type	Topic/Macro Ann.	Lead-Lag	Measure	% Selected	% Pos.	% Neg.
Macro	FOMC Rate Dec.	0	flag for announcement	70.79	65.17	5.62
StreetAcc.	All	-3	flag for news release	66.29	30.34	35.96
Macro	FOMC Rate Dec.	0	square surp	59.55	1.12	58.43
Macro	FOMC Rate Dec.	0	log surp	59.55	1.12	58.43
Macro	FOMC Rate Dec.	0	sqrt surp	59.55	1.12	58.43
Macro	FOMC Rate Dec.	0	surprise	59.55	1.12	58.43
Macro	Federal Budget	0	square surp	51.69	0.00	51.69
Macro	Federal Budget	-2	square surp	43.82	40.45	3.37
StreetAcc.	All	-3 and -2	flag for sent inversion	37.08	37.08	0.00

Notes: Sample = Jun 2009 – Feb 2015 (expansion).

2.7 Relative Economic Importance of News-Related Jumps

This section investigates the impact of jumps on asset return dynamics. We first separate jumps on the basis of the type of news to which they are associated and then, following Evans (2011), in turn based on Ederington and Lee (1993), we investigate: 1) the relative importance of jumps to asset return dynamics in terms of return predictability and volatility persistence; 2) the exposure of future returns to various measures of jump risk.

In sections 2.5 and 2.6 we analyzed the relation between news and jumps on the basis of a finer and finer distinction of the information that results from news. The analyses that follow hinge on the separation of jumps in subsamples, each of which includes only jumps – and related concurrent and subsequent returns and volatilities – related to one type of news. In order to perform these analyses each subsample cannot be too small, so we separate jumps by looking at four main news types:

- News Stories: all news stories released by StreetAccount and Thomson Reuters
- EPS
- Macro: all 23 macroeconomic announcements
- No News: absence of any news

and define an association between a jump and a news type as the occurrence of a jump in a 5-min interval and the release of a news belonging to a news type in the same on in the preceding two intervals (10 minutes).[10]

2.7.1 Returns Predictability and Volatility Persistence

By separating jumps on the basis of the types of news to which they are associated, the analysis attempts to understand the relative importance of jumps to asset return dynamics in terms of return predictability and volatility persistence, at both high frequency and daily level.

High Frequency Returns and Squared Returns

Using a subsample of the dataset that includes only intraday jumps, returns and squared returns in subsequent intraday intervals are analysed as follows:

$$r_j = \beta_{NS} JD_{NS,j} + \beta_{EPS} JD_{EPS,j} + \beta_{Macro} JD_{Macro,j} + \beta_{NoNews} JD_{NoNews,j} + \varepsilon_j \tag{2.20}$$

$$r_j^2 = \beta_{NS} JD_{NS,j} + \beta_{EPS} JD_{EPS,j} + \beta_{Macro} JD_{Macro,j} + \beta_{NoNews} JD_{NoNews,j} + \varepsilon_j \tag{2.21}$$

where r_j represents the 5-min return in interval j, which is equivalent to the intraday jump if a jump was observed at interval j, $JD_{k,j}$ is a dummy variable equal to one if the intraday jump is associated to the news type k, and k represents NS (news stories), EPS, Macro, and NoNews. Eqs. 2.20 and 2.21 are estimated first using only the intraday jumps, and then for returns and squared returns for each of the 12 intraday intervals (1 hour) following the jumps.[11]

Figures 2.3–2.6 illustrate the estimation results of eq. 2.20, separately for positive and negative jumps. It does not make sense to perform this analysis including all jumps independently of their sign. EPS-related jumps have no significant effects, so we look here at only the other three news types. News stories-related jumps absolute sizes are higher for both positive and negative jumps, as can be noticed by looking at the values of β at the post-jump interval 0, and obviously the sign of β coincides with the sign of jumps. For all jumps, independently of the news associated to them, there is a reversal effect in the following interval, while there is no evident effect in the post-jump intervals from 2 to 12.

[10] Results from penalized logistic regression show that jumps could be related to news released until one hour before, but in this Section we choose to relate jumps to news occurring until two preceding intervals only, with the aim of reducing the risk of spurious associations.

[11] We discard the first and the last intraday intervals from the analysis because they are influenced by market microstructure. In addition, jumps in the first intraday interval may be caused by news released overnight.

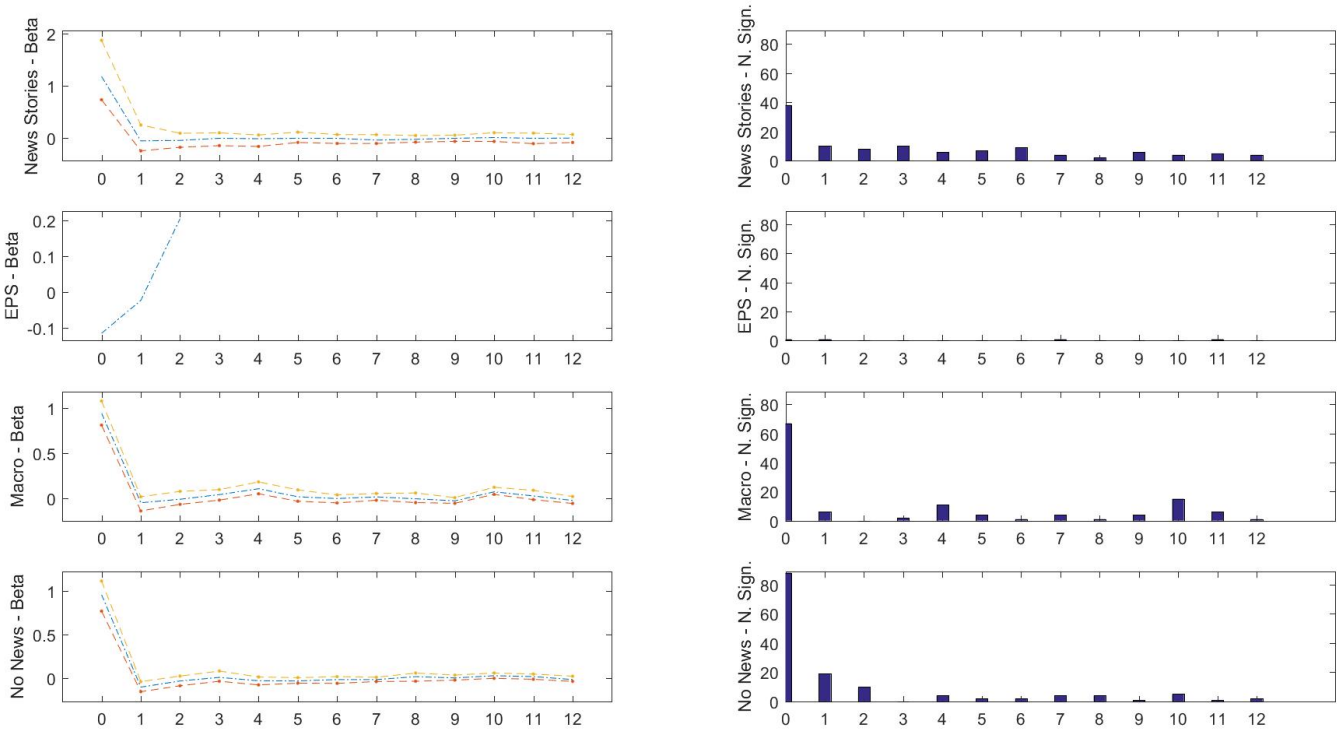

Figure 2.3: Jumps impact on high freq. returns (positive jumps)

Notes: on the left, 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right, number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the number of 5-min intervals after the jump. Only positive jumps.

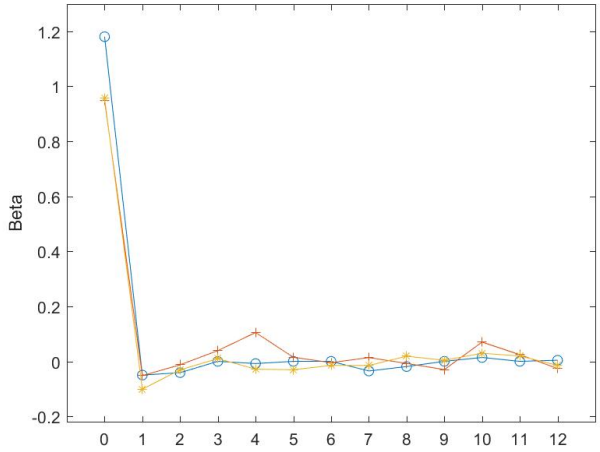

Figure 2.4: Jumps impact on high freq. returns - news type comparison (positive jumps)

Notes: median of the estimated β across all assets, for each news type. Circle: news stories-related jumps; plus: macro-related jumps; star: no news-related jumps. The horizontal axis represents the number of 5-min intervals after the jump. Only positive jumps.

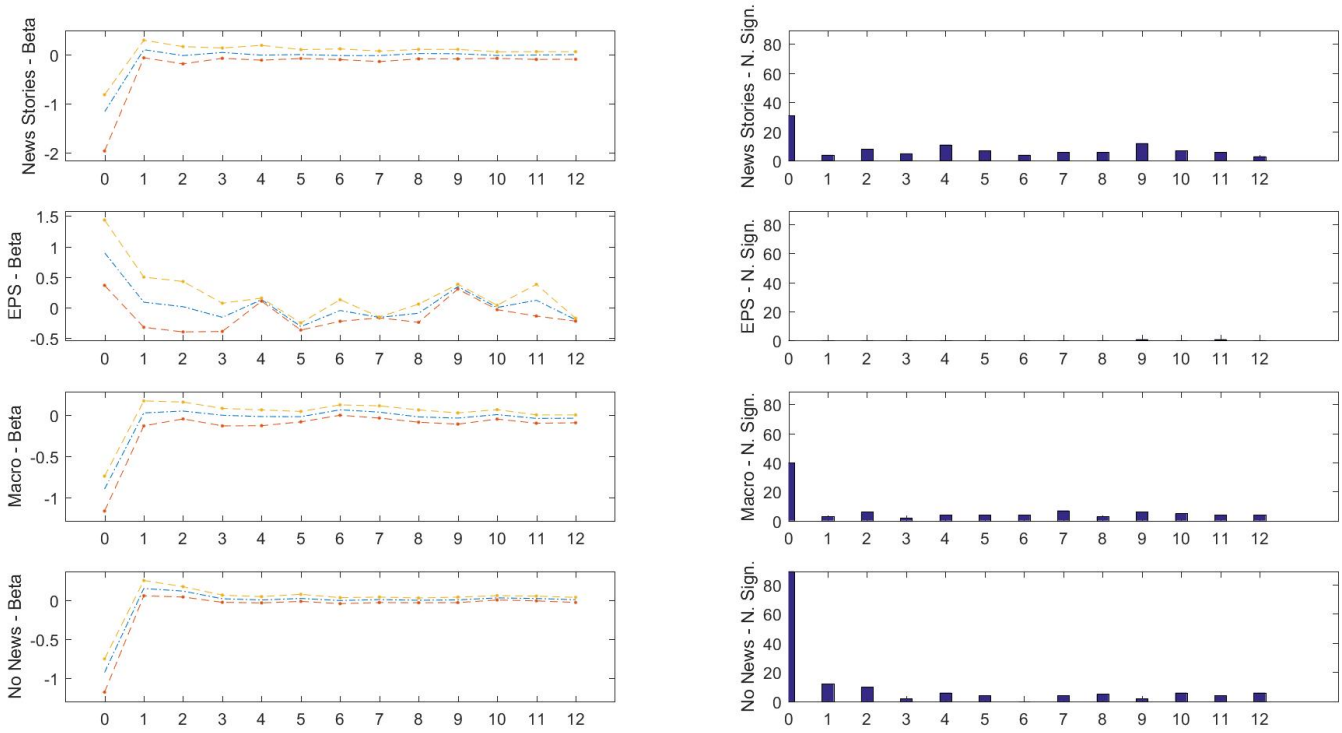

Figure 2.5: Jumps impact on high freq. returns (negative jumps)

Notes: on the left, 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right, number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the number of 5-min intervals after the jump. Only negative jumps.

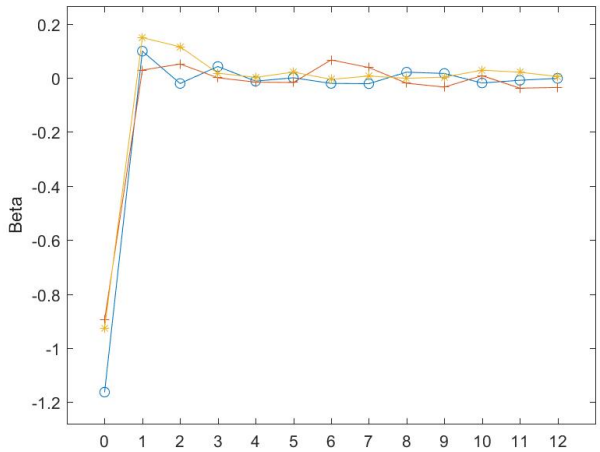

Figure 2.6: Jumps impact on high freq. returns - median comparison (negative jumps)

Notes: median of the estimated β across all assets, for each news type. Circle: news stories-related jumps; plus: macro-related jumps; star: no news-related jumps. The horizontal axis represents the number of 5-min intervals after the jump. Only negative jumps.

Figures 2.7–2.8 illustrate the estimation results of eq. 2.21. EPS-related jumps have no significant effects, so we look here at only the other three news types. All types of jumps have a positive effect on high frequency squared returns, with persistence up to at least one hour. In the post-jump intervals, no news-related jumps have the highest effect. Distinguishing positive and negative jumps leads to qualitatively and quantitatively

similar results, which are available on request.

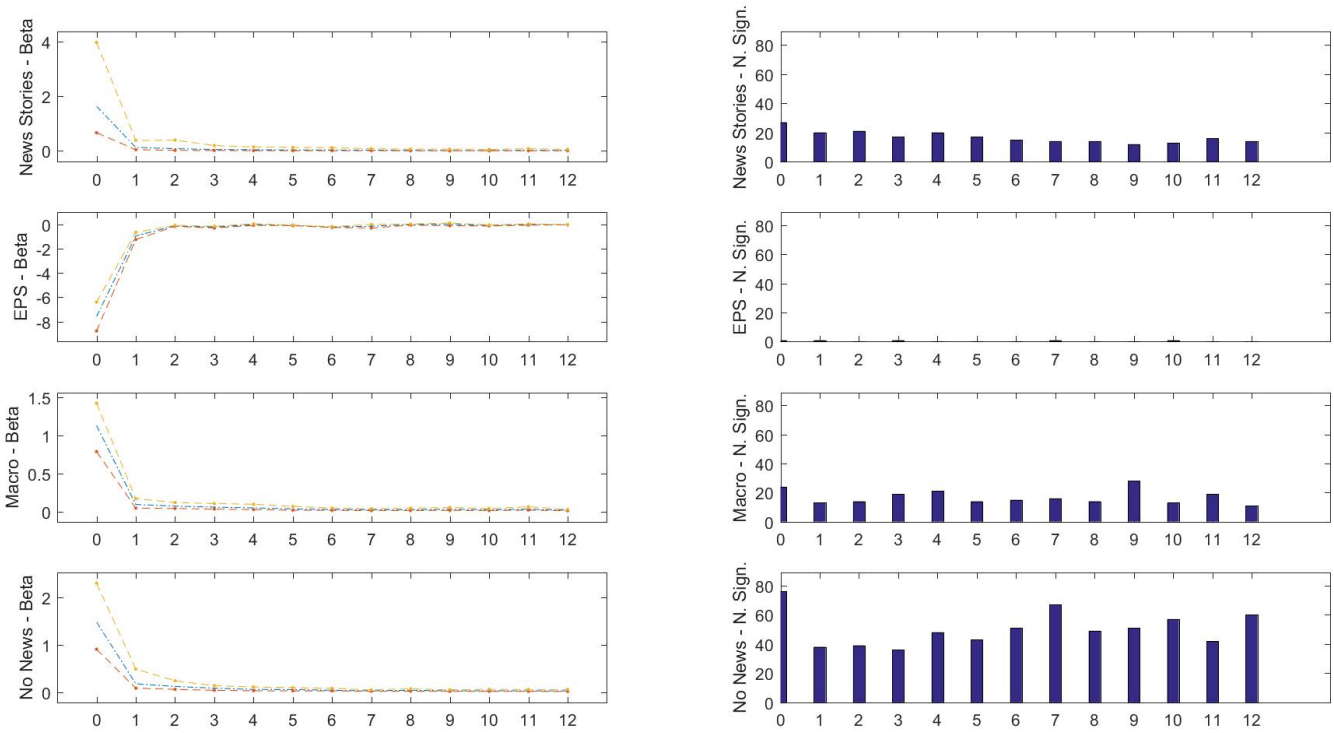

Figure 2.7: Jumps impact on high freq. squared returns (all jumps)

Notes: on the left, 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right, number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the number of 5-min intervals after the jump.

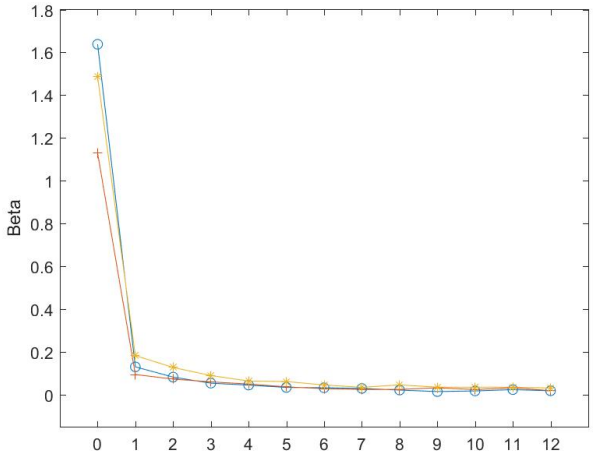

Figure 2.8: Jumps impact on high freq. squared returns - median comparison (all jumps)

Notes: median of the estimated β across all assets, for each news type. Circle: news stories-related jumps; plus: macro-related jumps; star: no news-related jumps. The horizontal axis represents the number of 5-min intervals after the jump.

Daily Returns and Realized Volatility

In order to test the dynamics of returns and volatility following jumps at the daily level, a similar procedure is followed:

$$r_t = \beta_{NS} JD_{NS,t} + \beta_{EPS} JD_{EPS,t} + \beta_{Macro} JD_{Macro,t} + \beta_{NoNews} JD_{NoNews,t} + \varepsilon_t \tag{2.22}$$

$$RV_t = \beta_{NS} JD_{NS,t} + \beta_{EPS} JD_{EPS,t} + \beta_{Macro} JD_{Macro,t} + \beta_{NoNews} JD_{NoNews,t} + \varepsilon_t \tag{2.23}$$

where r_t and RV_t represent, respectively, the daily return and realized volatility in day t, $JD_{k,t}$ is a dummy variable equal to one if the intraday jump in day t is associated to the news type k, and k represents NS (news stories), EPS, Macro, and NoNews. Eqs. 2.22 and 2.23 are estimated first using only daily returns and RV of the jump days, and then for returns and RV for each of the 10 days following the jump day.[12]

Figures 2.9–2.12 illustrate the estimation results of eq. 2.22, separately for positive and negative jumps. As for high frequency returns, it does not make sense to perform this analysis including all jumps independently of their sign. EPS-related jumps have no significant effects, so we look here at only the other three news types. News stories and macro-related jumps β at post-jump day 0 (same day of jump) is higher than no news-related jumps β for both positive and negative jumps. As expected, for all news types the sign of β at post-jump day 0 coincides with the sign of the jump, indicating that the sign of jump dominates the sign of the daily return. News stories-related positive jumps positive effect on daily return show persistence up to the day after the jump and a reversal in the second day after the jump, while news stories-related negative jumps have a negative impact on daily return which persists up to three days after the jump and reverses in the fourth day after the jump. The other types of jumps do not show any evident effect on the post-jump days.

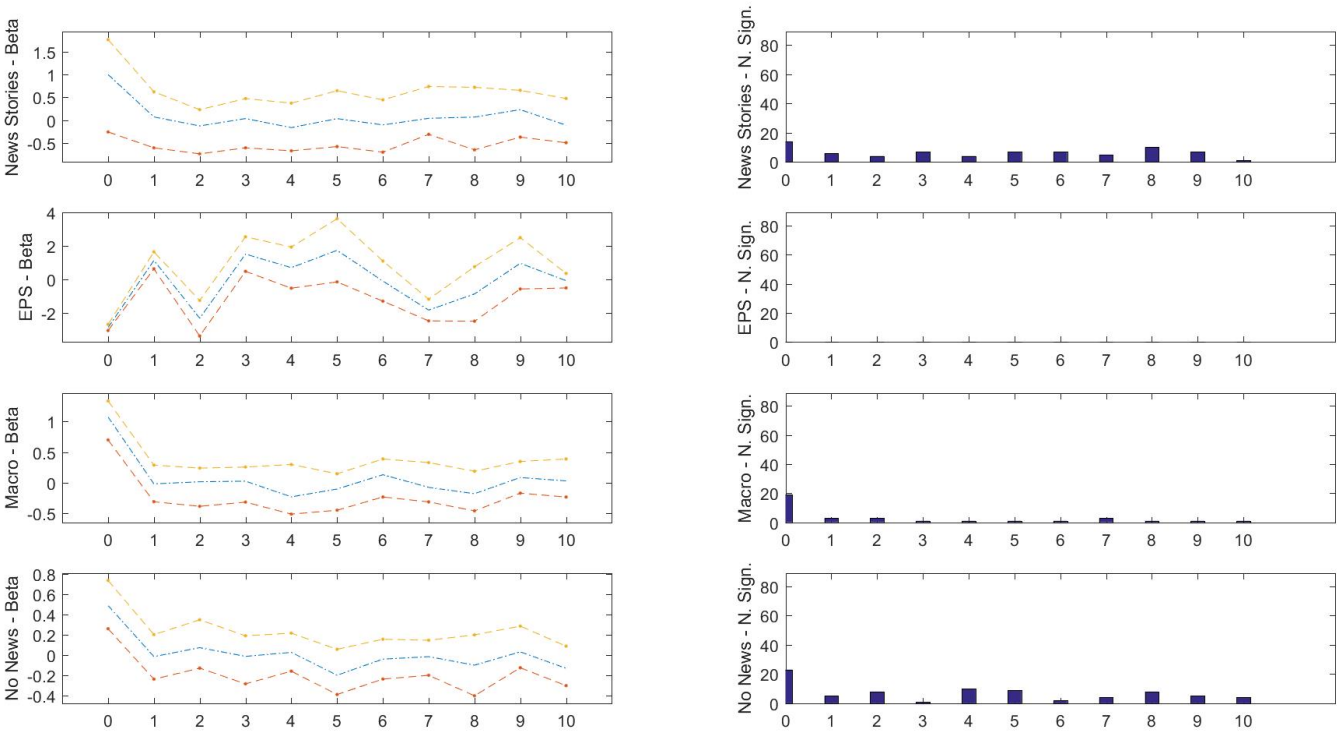

Figure 2.9: Jumps impact on daily returns (positive jumps)

Notes: on the left, 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right, number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the number of days after the jump. Only positive jumps.

[12] As for the high frequency analysis above, we discard jumps occurring in the first and in the last intraday intervals. Daily returns and realized volatility are instead computed over the entire day.

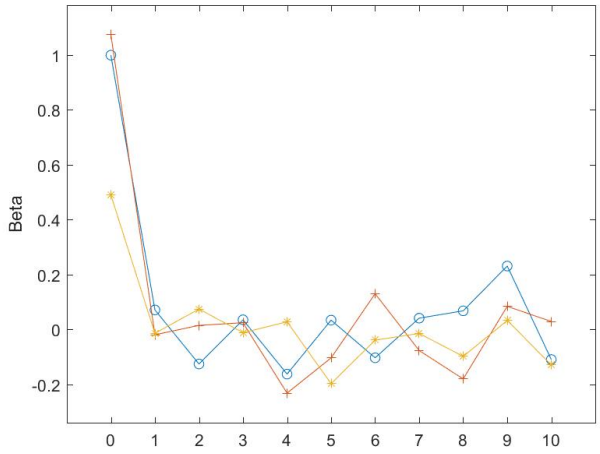

Figure 2.10: Jumps impact on daily returns - median comparison (positive jumps)
Notes: median of the estimated β across all assets, for each news type. Circle: news stories-related jumps; plus: macro-related jumps; star: no news-related jumps. The horizontal axis represents the number of days after the jump. Only positive jumps.

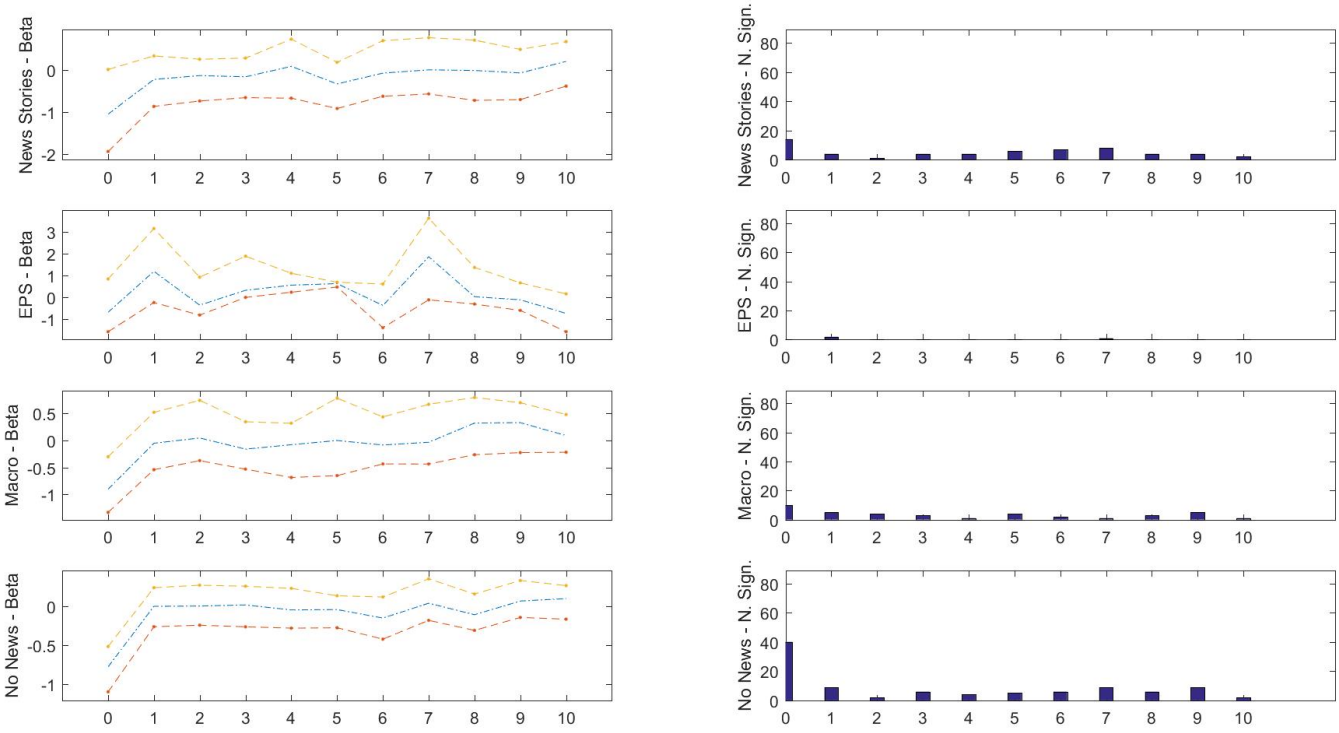

Figure 2.11: Jumps impact on daily returns (negative jumps)
Notes: on the left, 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right, number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the number of days after the jump. Only negative jumps.

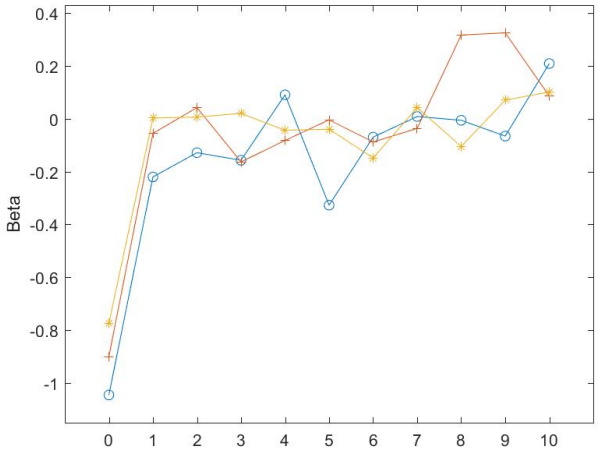

Figure 2.12: Jumps impact on daily returns - median comparison (negative jumps)
Notes: median of the estimated β across all assets, for each news type. Circle: news stories-related jumps; plus: macro-related jumps; star: no news-related jumps. The horizontal axis represents the number of days after the jump. Only negative jumps.

Figures 2.13–2.14 illustrate the estimation results of eq. 2.23. EPS-related jumps have no significant effects, so we look here at only the other three news types. All jumps, independently of the news associated to them, have a positive effect on the daily RV of the jump day, with news stories and no news-related jumps showing a higher effect than macro-related jumps. All jumps show a persistence on RV that lasts at least until the tenth post-jump day, but we can notice a much higher persistence of no news-related jumps with respect to the rest of jumps. Results from distinguishing positive and negative jumps are qualitatively and quantitatively similar, and available on request.

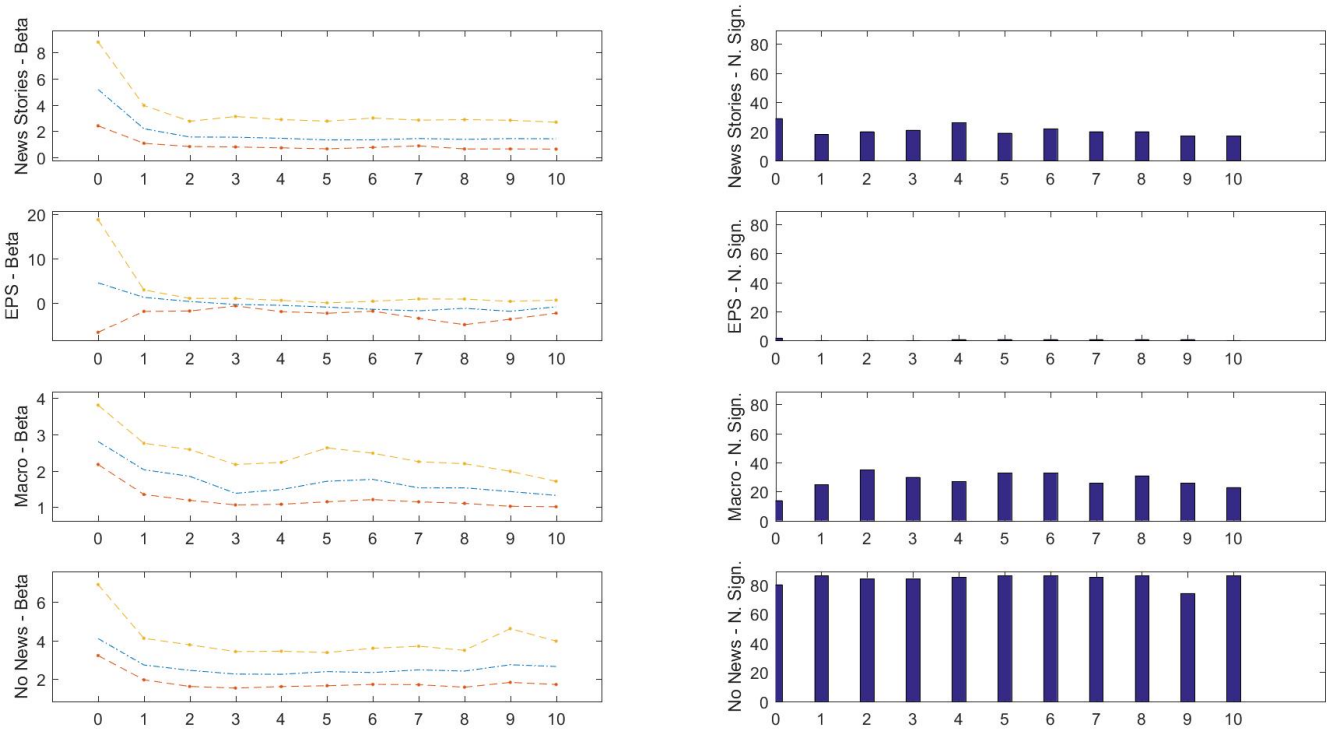

Figure 2.13: Jumps impact on daily RV (all jumps)

Notes: on the left, 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right, number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the number of days after the jump.

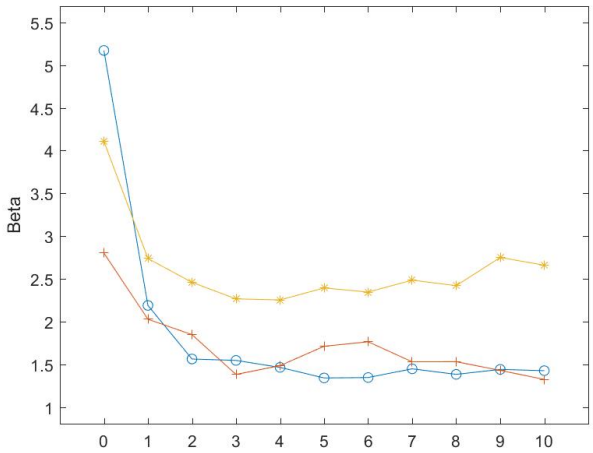

Figure 2.14: Jumps impact on daily RV - median comparison (all jumps)

Notes: median of the estimated β across all assets, for each news type. Circle: news stories-related jumps; plus: macro-related jumps; star: no news-related jumps. The horizontal axis represents the number of days after the jump.

Summarizing, jumps show effects on returns and volatility at both high frequency and daily level. From high frequency analysis, we can notice that news stories-related jumps absolute sizes are higher than the jumps associated to the other types of news. In addition all jumps, independently of the type of news to which they are associated, show a reversal effect in terms of a return with an opposite sign in the following 5-min intraday interval. Furtherly, still independently of the news to which they are related, they increase future intraday

squared returns for at least one hour.

From daily analysis, we see that all jumps dominate, in terms of sign, the daily return of the day on which they occur, and that news stories and macro-related jumps have a greater effect. Only news stories-related jumps show persistence and reversal effects on the returns of the following days, and the effects vary depending on the jumps sign: positive jumps increase the return of the following day and the effect reverses the day after (second post-jump day), while negative jumps negatively impact the daily return of the three following days and the effect reverses the day after (fourth post-jump day). With regard to daily realized volatility all jumps have a positive effect on the volatility of the same day on which they occur, and the effect persists until at least the tenth day after the jump. Jumps not related to any news increase the volatility of the post-jump days much more than jumps associated to news.

2.7.2 Returns Exposure to Jump Risk Measures

Evans (2011) extends the analysis of return predictability to incorporate the more recent, sophisticated and relevant techniques of Wright and Zhou (2009) and Tauchen and Zhou (2011) who, given the relatively infrequent occurrence of jumps in asset prices, define more appropriate rolling measures of jump risk and then investigate the exposure of future returns to these smoothed measures, as a way to analyze their association with risk premia.

The rolling measure of realized variation is defined as the average daily measure over a 22-day month:

$$RV_t^h = \frac{1}{h \cdot 22} \sum_{l=0}^{h \cdot 22-1} RV_{t-l} \tag{2.24}$$

where h measures the length of the rolling window. The jump risk measures of Jump Intensity (JI), Jump Mean (JM) and Jump Volatility (JV) are defined respectively as

$$JI_t^h = \frac{1}{h \cdot 22} \sum_{l=0}^{h \cdot 22-1} JD_{t-l}^{DLY} \tag{2.25}$$

$$JM_t^h = \frac{\sum_{l=0}^{h \cdot 22-1} J_{t-l}^{DLY} \cdot JD_{t-l}^{DLY}}{\sum_{l=0}^{h \cdot 22-1} JD_{t-l}^{DLY}} \tag{2.26}$$

$$JV_t^h = \sqrt{\frac{\sum_{l=0}^{h \cdot 22-1} \left(J_{t-l}^{DLY} - JM_t^h \right)^2 \cdot JD_{t-1}^{DLY}}{\sum_{l=0}^{h \cdot 22-1} JD_{t-l}^{DLY}}} \tag{2.27}$$

where h measures the rolling window length (number of months), J_t represents a daily measure of the jump size and JD_t is a dummy variable equal to one if a jump occurs on a particular day. In order to assess the relation of future returns to the jump risk measures on the basis of the news types to which jumps can be associated, JI, JM and JV are separately calculated to provide different measures according to each news type.

The exposure of future returns to the jump risk measures can then be investigated by estimating various regressions which are nested in the following expression:

$$r_t^f = \alpha + \beta_{RV} RV_t^1 + \sum_{k=NS}^{NoNews} \beta_{JI,k} JI_{k,t}^h + \sum_{k=NS}^{NoNews} \beta_{JM,k} JM_{k,t}^h + \sum_{k=NS}^{NoNews} \beta_{JV,k} JV_{k,t}^h + \varepsilon_t^f \tag{2.28}$$

where $r_t^f = \log(p_{t+f \cdot 22}/p_t) \cdot 100$ represents the continuously compounded return from day $(t+1)$ to day $(t+f \cdot 22)$, $JI_{k,t}^h$, $JM_{k,t}^h$ and $JV_{k,t}^h$ are the jump risk measures computed using only the jumps associated to news type k, and $k = NS$, EPS, $Macro$, and $NoNews$.[13]

Figures 2.15 and 2.16 illustrate the exposure to JI of future returns cumulated over three months and one year, respectively. It seems that only macro-related JI has an influence on returns. Precisely, macro-related JI computed over the last year ($h = 12$) is negatively related to future quarterly returns ($f = 3$). However, the extremely high magnitude of the estimated coefficients makes us cautious in defining the existence of a strong relationship between these variables. Distinguishing positive and negative jumps leads to similar results, available on request.

[13] As above, we discard jumps occurring in the first and in the last intraday intervals in the computation of the rolling measures of jump risk.

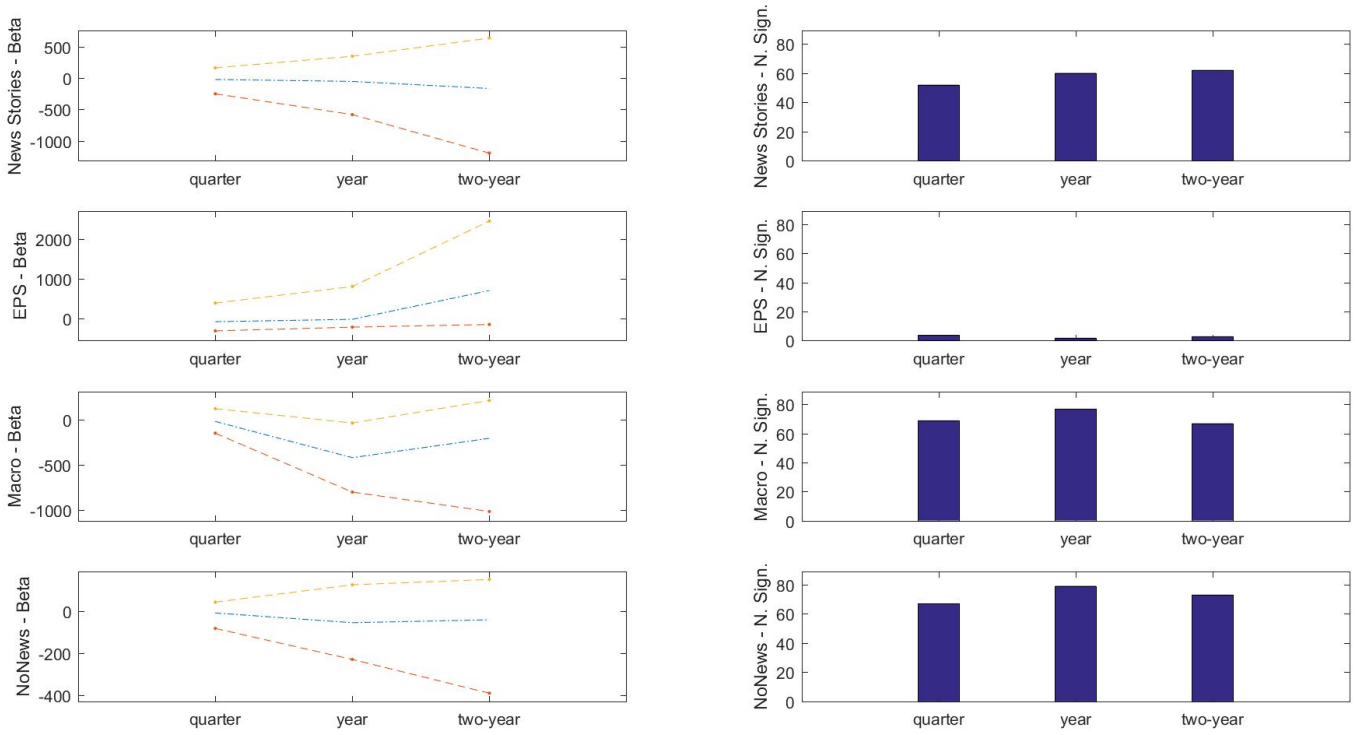

Figure 2.15: JI estimation results, f = 3

Notes: estimation results of the equation $r_t^f = \alpha + \sum_{k=NS}^{NoNews} \beta_{JI,k} JI_{k,t}^h + \varepsilon_t^f$, with $f = 3$ (one quarter). On the left: 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right: number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the time horizon h over which the jump risk measures are computed.

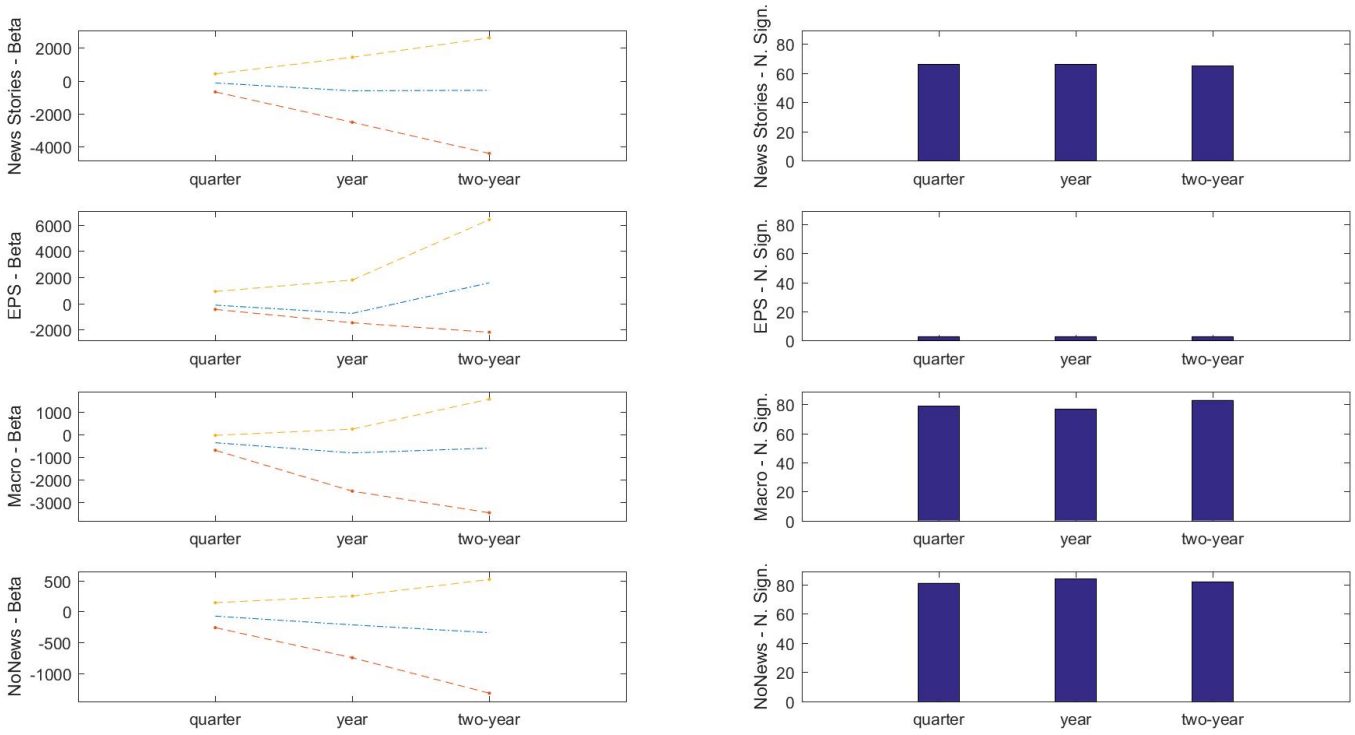

Figure 2.16: JI estimation results, f = 12

Notes: estimation results of the equation $r_t^f = \alpha + \sum_{k=NS}^{NoNews} \beta_{JI,k} JI_{k,t}^h + \varepsilon_t^f$, with $f = 12$ (one year). On the left: 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right: number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the time horizon h over which the jump risk measures are computed.

Figures 2.17 and 2.18 illustrate the exposure of future returns to *JM*. As for *JI*, it seems that only macro-related *JM* has an influence on returns. Macro-related *JM* computed over the last year is negatively related to future quarterly returns, as well as macro-related *JM* computed over the last quarter to future yearly returns. Distinguishing positive and negative jumps leads to similar results, available on request.

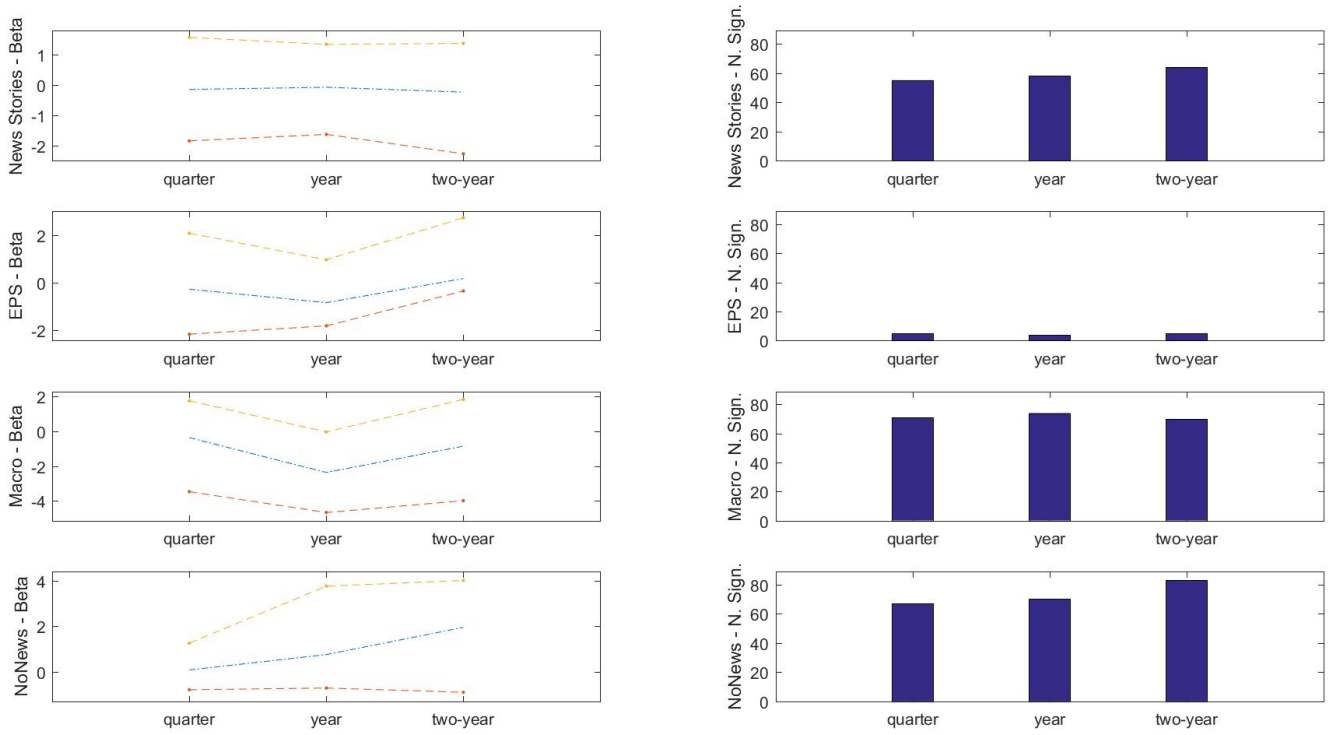

Figure 2.17: JM estimation results, f = 3

Notes: estimation results of the equation $r_t^f = \alpha + \sum_{k=NS}^{NoNews} \beta_{JM,k} JM_{k,t}^h + \varepsilon_t^f$, with $f = 3$ (one quarter). On the left: 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right: number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the time horizon h over which the jump risk measures are computed.

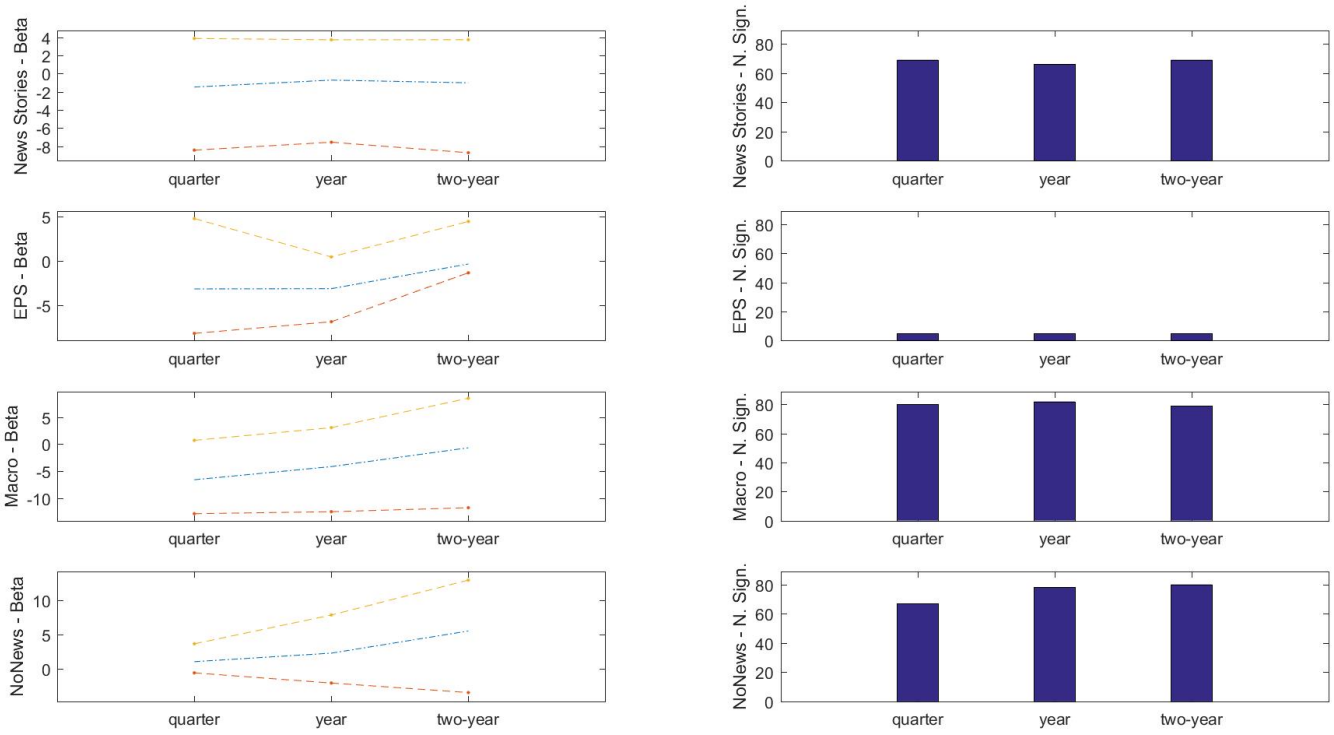

Figure 2.18: JM estimation results, f = 12

Notes: estimation results of the equation $r_t^f = \alpha + \sum_{k=NS}^{NoNews} \beta_{JM,k} JM_{k,t}^h + \varepsilon_t^f$, with $f = 12$ (one year). On the left: 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right: number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the time horizon h over which the jump risk measures are computed.

Figures 2.19 and 2.20 illustrate the exposure of future returns to *JV*. As for *JI* and *JM*, it seems that only macro-related *JV* has an influence on returns. Macro-related *JV* computed over the last year is negatively related to future quarterly returns, as well as macro-related *JV* computed over the last quarter and over the last year to future yearly returns. As for JI, we are cautious on this interpretation because of the high magnitude of coefficients. Distinguishing positive and negative jumps leads to similar results, available on request.

Figure 2.21 illustrates the exposure to *JV* of future returns cumulated over one year, computing *JV* using only positive jumps. Interestingly, news stories-related *JV* computed over the last quarter seems to be negatively related to future yearly returns. Instead, news stories-related *JV* computed using only negative jumps does not seem to impact future returns in any way. Again, coefficients are not reliable.

Summarizing, all three jump risk measures *JI*, *JM* and *JV* show a relation with future returns, but only when they are built using macro-related jumps. Future returns are negatively exposed to these measures. In addition, *JV* built using only news stories-related positive jumps seems also to be negatively related to future returns. Estimated coefficients, however, are not reliable and further research has to be done.

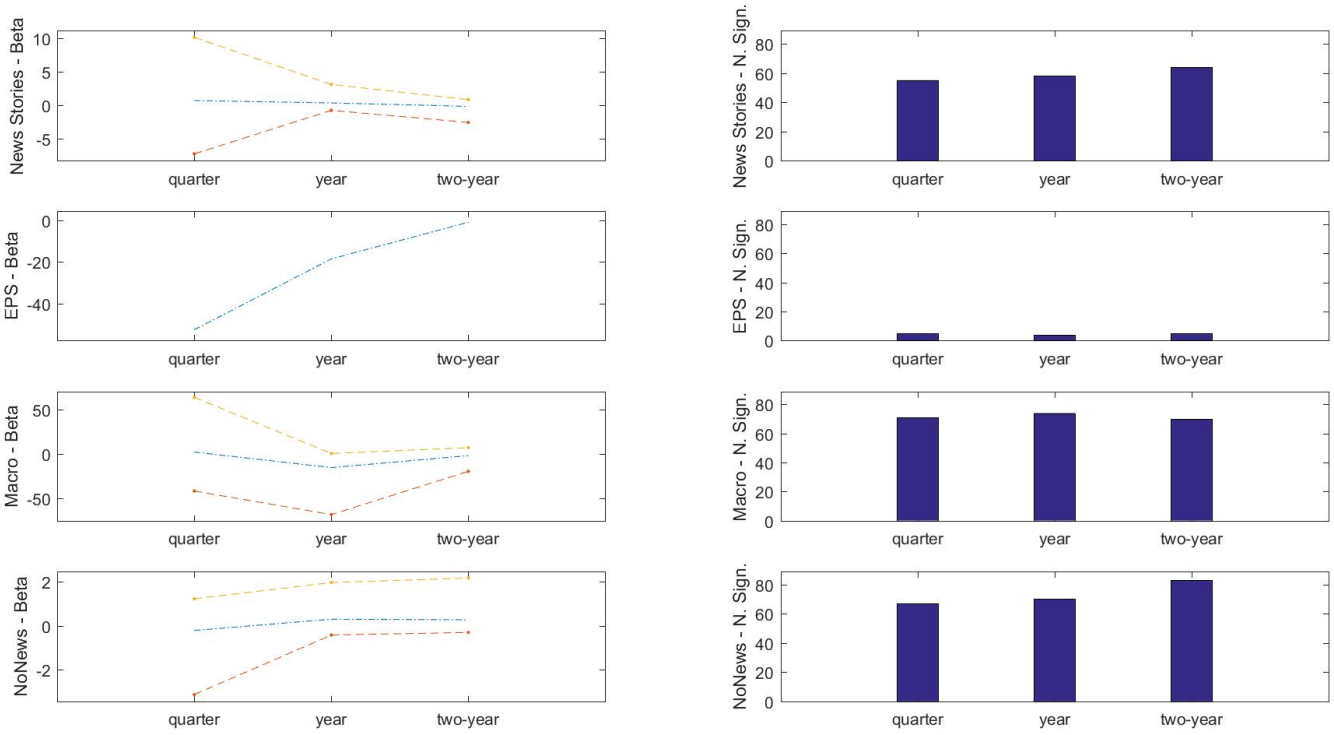

Figure 2.19: JV estimation results, f = 3

Notes: estimation results of the equation $r_t^f = \alpha + \sum_{k=NS}^{NoNews} \beta_{JV,k} JV_{k,t}^h + \varepsilon_t^f$, with $f = 3$ (one quarter). On the left: 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right: number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the time horizon h over which the jump risk measures are computed.

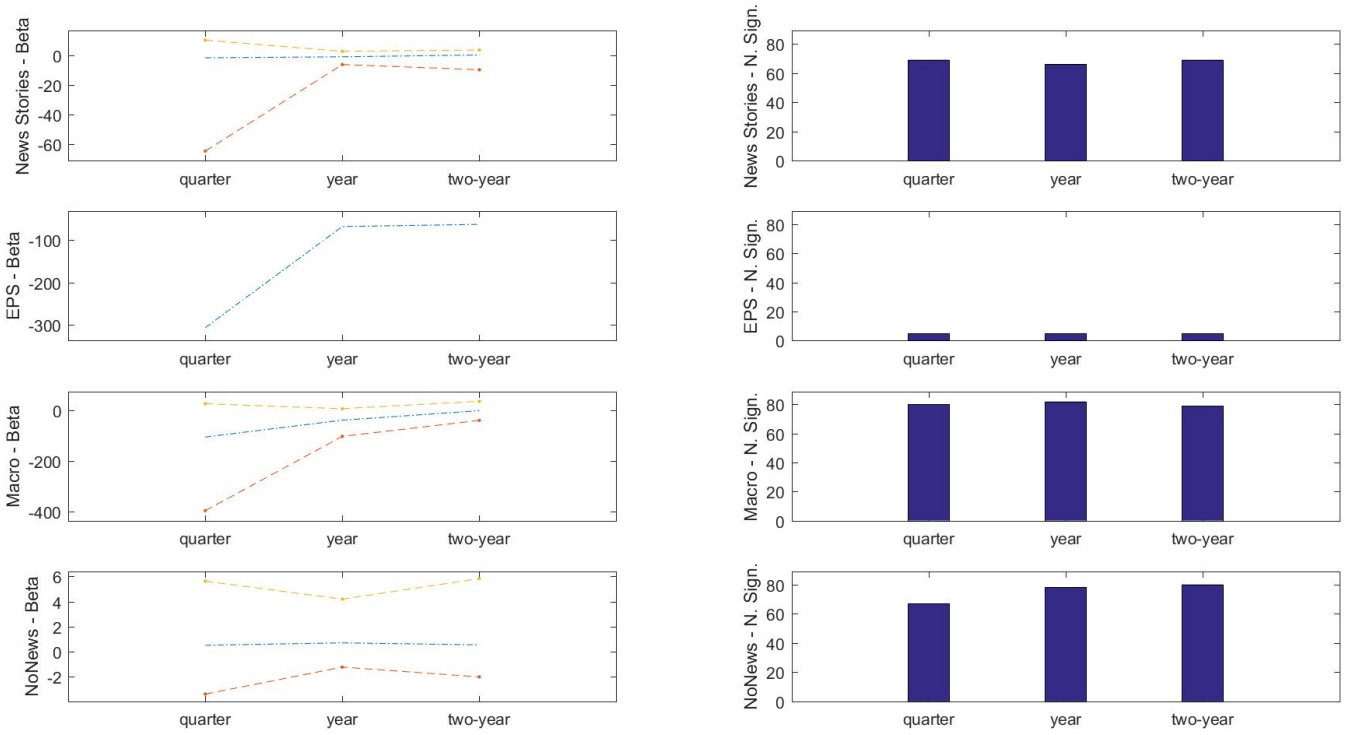

Figure 2.20: JV estimation results, f = 12

Notes: estimation results of the equation $r_t^f = \alpha + \sum_{k=NS}^{NoNews} \beta_{JV,k} JV_{k,t}^h + \varepsilon_t^f$, with $f = 12$ (one year). On the left: 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right: number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the time horizon h over which the jump risk measures are computed.

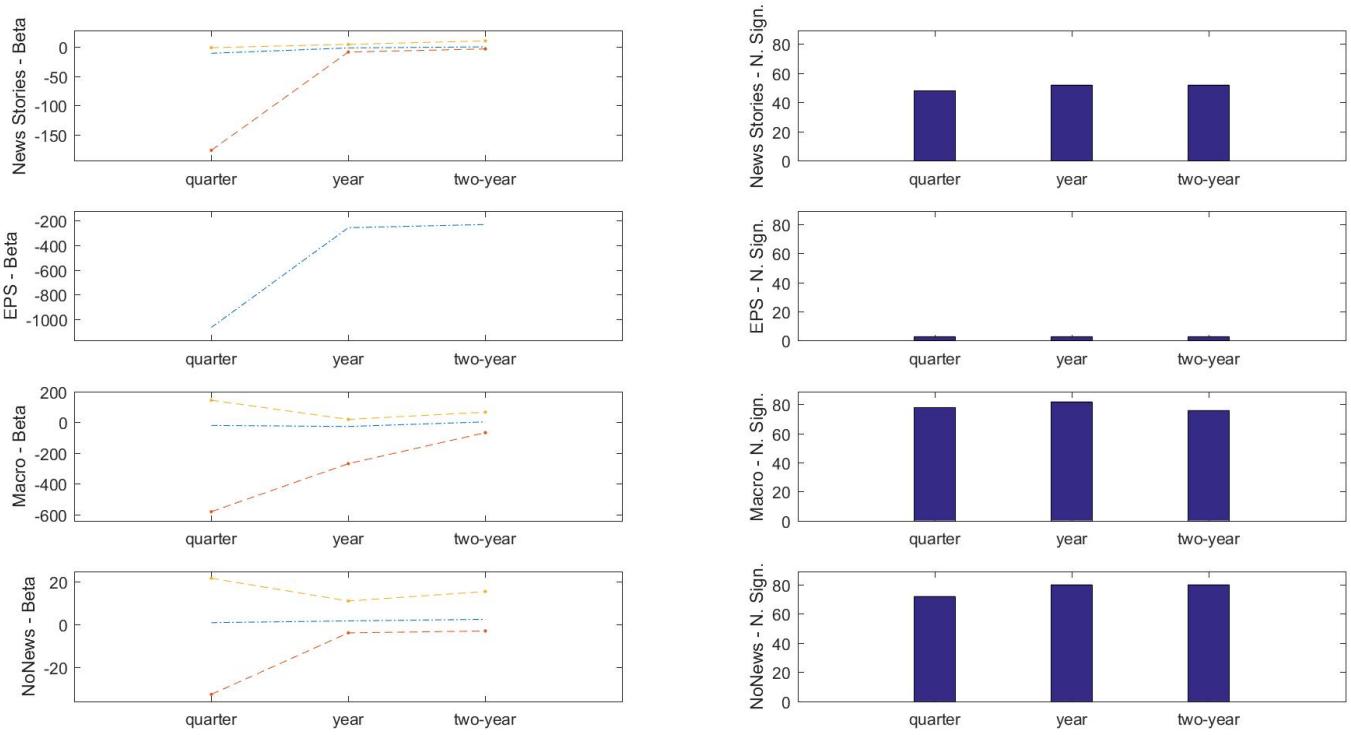

Figure 2.21: JV estimation results, f = 12, positive jumps

Notes: estimation results of the equation $r_t^f = \alpha + \sum_{k=NS}^{NoNews} \beta_{JV,k} JV_{k,t}^h + \varepsilon_t^f$, with $f = 12$ (one year). On the left: 25% quantile, median and 75% quantile of the estimated β across all assets, for each news type; on the right: number of assets with a statistically significant β at a 5% level for a two-tailed test that $\beta = 0$, for each news type. The horizontal axis represents the time horizon h over which the jump risk measures are computed. Jump risk measures computed using only positive jumps.

2.8 Concluding Remarks

We identified intraday jumps of the S&P 100 components' stocks and related them to firm-specific news stories, EPS and macro-announcements. From a matching analysis based on news-jump coincidences it is possible to say that, although the majority of jumps is not associated with news and may be due to market frictions, EPS, FOMC rate decisions and news stories classified as *top* by Thomson Reuters represent potentially very useful information to determine the causes of jumps.

Using sentiment of news stories and surprises from expectations of EPS and macroeconomic announcements, we built more than 1,500 news-related variables with the aim of reconstructing the different portions of information assimilated by heterogeneous market players. Then, we applied penalized logistic regression estimation method to understand which indicators are more likely to cause jumps. The use of much more information for the construction of news indicators allows to detect further potential sources of jumps. With regard to macro-news, FOMC rate decisions are confirmed to be a very important determinant of jumps, along with three other announcements: federal budget, natural gas stocks and ECRI, for which both announcements per se and surprises (both above and below expectations) count. For FOMC rate decisions, jumps probability is increased by higher than expected rates during contractions, and by lower than expected rates during expansions. With regard to news stories, releases by both StreetAccount and Thomson Reuters are important. News releases per se as well as positive and negative news count, and all topics are relevant, especially *M&A* and *Earnings Related*. The event of *sentiment inversion* is also relevant. Both macro-announcements and news stories are likely to increase the probability of jumps in the following hour, but when news are negative jumps occur within at most five minutes.

We then investigated the impact of jumps on asset return dynamics, distinguishing the main types of news to which jumps can be associated (news stories, EPS, macro, and absence of any news). Jumps show effects on returns and volatility at both high frequency and daily level, and effects vary on the basis of the news to which jumps are associated. News stories-related jumps absolute sizes are higher with respect to the other jumps. All jumps, independently of the type of news to which they are associated, are usually followed by an intraday

return with an opposite sign and, in addition, post-jump intraday squared returns remain elevated for at least one hour. Interestingly, only news stories-related jumps show persistence and reversal effects on daily returns of the following days, and the effects vary depending on the jumps sign: positive jumps increase the return of the following day, which is followed by a reversal in the day after, while negative jumps negatively impact the returns of the first three post-jump days, which are also followed by a reversal in the fourth post-jump day. All jumps increase the daily realized volatility until at least the tenth post-jump day, but jumps not related to any news exhibit a much greater effect than jumps related to news.

Finally, we found that future quarterly and yearly returns seem to be negatively exposed to the three jump risk measures Jump Intensity, Jump Mean and Jump Volatility, but only when they are built using jumps related to macro-announcements. Future returns seem also to be negatively exposed to Jump Volatility when it is built using only positive jumps related to news stories. Coefficients, however, are not reliable and further research should investigate these relationships.

Possible future research directions involve the relation between news and cojumps and the impact of information spillovers between assets.

Chapter 3

Bag-of-Rules for Sentiment Detection

Francesco Poli

3.1 Related Literature

Several studies investigate the relationship between news sentiment and changes in asset dynamics. Antweiler and Frank (2004) are the first to develop news sentiment measures to understand stock returns; using a Naive Bayes algorithm based on the number of occurrences of words, they infer trading signals from posts on internet message boards and find that while such signals are able to predict market volatility, their effect on stock returns is small. Zhang et al. (2012) incorporate several methodological improvements and create news sentiment indices that are significant directional indicators. Tetlock (2007) undertakes the bag-of-words approach, that has become more spread in the literature. Classifying words on the basis of categories from the Harvard psychosocial dictionary, he quantifies optimism and pessimism from Wall Street Journal's Abreast of the Market column and reports that high levels of media pessimism predict declining market prices which are followed by price reversals. Using a similar technique, Tetlock et al. (2008) use Harvard IV-4 psychological dictionary and find that the fraction of negative words in Dow Jones News Service and Wall Street Journal stories forecasts firm earnings, reporting that this is due to the linguistic content of news messages capturing hard to quantify aspects of fundamentals which are quickly incorporated into stock prices. Thanks to the recent advances in technology, software packages such as Reuters NewsScope Sentiment Engine (RNSE) and the more recent Thomson Reuters News Analytics (TRNA) [1] , and Ravenpack News Analytics [2] have been developed. They utilize advanced algorithms and assign sentiment indicators to firm-specific newswire releases, enabling investors willing to pay for the service to employ real-time trading signals from textual analysis in quantitative trading strategies. Gloß-Klußmann and Hautsch (2011) employ the trading signals from RNSE and find that high-frequency responses in market activity and volatility are significant especially after the release of intraday company-specific news, and that the classification of news according to indicated relevance is crucial for noise filtering and identification of significant effects. By using sentiment scores at high frequencies generated by RavenPack News Analytics, Ho et al. (2013) investigate the circumstances in which public news sentiment is related to the intraday volatility of stocks and find a significant impact of firm-specific news sentiment on intraday volatility persistence, even after controlling for the potential effects of macroeconomic news. Firm-specific news sentiment apparently accounts for a greater proportion of overall volatility persistence compared with macroeconomic news sentiment, and negative news have a greater impact on volatility than positive ones. Riordan et al. (2013) suggest that negative newswire messages from RNSE are associated with higher adverse selection costs, are more informative, and have a more significant impact on high-frequency asset price discovery and liquidity. Smales (2015) use TRNA sentiment scores and create aggregate daily news sentiment indicators to examine the relationship between news sentiment and stock market returns. They report that positive and negative news result in above and below average returns, respectively, and that neutral news days are indistinguishable from days without news. In the field of bag-of-words methods in financial contexts, Loughran and McDonald (2011) show that word lists developed for other disciplines misclassify common words in financial text, and develop an alternative negative word list, an alternative positive word list, and four other word lists, that better reflect tone in financial text. They show that the proportion of negative words in annual 10-Ks reports is associated with lower returns.

[1] Reuters NewsScope Sentiment Engine and Thomson Reuters News Analytics are tools of the Reuters company which provide for each news a sentiment and linguistic analytics, such as novelty and relevance indicators. The indicators are produced based on an automated linguistic pattern recognition of news texts.

[2] RavenPack News Analytics is a service of RavenPack.com, a provider of news analytics and machine-readable content. RavenPack News Analytics provides event and sentiment information to financial services clients.

3.2 Methodology

The so-called *sentiment* is an indicator of whether the content of a document is good, bad or neutral in relation to the issue it talks about. Our aim consists in the development of a general and robust procedure for the sentiment detection of financial texts. We develop a sentiment extraction technique based on the work of Loughran and McDonald (2011), and extend it by employing:

1. an extended negations list of single words, two-word sequences, and three-word sequences
2. lists of sentiment-related expressions
3. lists of sentiment-related words combinations

Loughran and McDonald (2011) develop six word lists (*negative, positive, uncertainty, litigious, strong modal, weak modal*) and show that the proportion of negative words is associated with lower returns. Their lists are tailored for financial texts, e.g. do not contain words such as *liability*, *earnings* or *tax*, which are expected to appear in both positive and negative contexts. They account for negation but only for six words (*no, not, none, neither, never, nobody*) and only if one of them precedes a word classified as positive, and motivate this choice saying that the methodology is applied to US companies 10-Ks[3] filings and these texts are very unlikely to contain negation for negative words. We claim that their procedure is not adequate to extract the sentiment of general financial texts, for three reasons: 1) differently from 10-Ks that are given to the SEC, texts do not necessarily have a formal tone; 2) companies which fill 10-Ks are interested in giving a positive image of themselves and negating a negative word gives, generally, a less strong positive meaning than plain positive words, but general texts, like news providers' news stories, are not affected by this bias; 3) 10-Ks are long enough such that, if some negated negative words occur, the contribution of their wrongly detected sentiment is negligible for the assignment of sentiment to the whole document, while we want a method suitable also for texts shorter than few dozens of words.

We introduce the following improvements:

- we invert the sentiment each time a word, irrespective of whether it is positive or negative, is preceded by a negation, and in place of their short list of 6 single words, we use 28 single words, 24 sequences of two words and 6 sequences of three words. Negations are the following:
  - single words: *no, not, none, never, nothing, nobody, nowhere, neither, nor, hardly, scarcely, seldom, barely, few, little, rarely, instead, can't, cannot, don't, doesn't, didn't, mustn't, won't, despite, overly, too, less*
  - two words sequences: *can not, do not, did not, short of, not every, not all, not much, not many, not always, not so, instead of, far from, not to, never to, no way, out of, not very, not enough, too few, too little, no big, not big, no significant, not significant*
  - three words sequences: *not at all, by no means, in no way, in place of, in spite of, in lieu of*
- we use lists of sentiment-related expressions. Examples are:
  - positive: *maintains market perform rating*
  - negative: *anti competitive*
- we use lists of sentiment-related words combinations. Examples are:
  - positive: *contract + announce*, *goal + accomplish*, *performance + solid*
  - negative: *contract + terminated*, *goal + out of reach*, *performance + slowing*

The procedure works as follows:

1. positive items (words, expressions and combinations) are given a value of 1, negative items -1
2. the value is inverted in case of negation
3. values of all items are summed up to get the sentiment sum *Sent_Sum*

$$Sent\_Sum = \sum_{i=1}^{N} s_i \tag{3.1}$$

where i is the item index, N is the number of items in a text and s_i is the sentiment of the item indexed by i

[3] A Form 10-K is an annual report required by the U.S. Securities and Exchange Commission (SEC), that gives a comprehensive summary of a company's financial performance.

4. *Sent_Sum* is divided by the number of items, obtaining a standardized quantity that we call relative sentiment *Rel_Sent*, comprised between -1 and 1 by construction

$$Rel\_Sent = \frac{Sent\_Sum}{N} \tag{3.2}$$

5. If *Rel_Sent* is bigger or smaller than 0.05 we associate, respectively, a positive (1) or a negative sentiment (-1) to the text, otherwise a neutral sentiment (0) is given

$$Text\_Sent = \begin{cases} -1 & \text{if} \quad Rel\_Sent < -0.05 \\ 0 & \text{if} \quad -0.05 \leq Rel\_Sent \leq 0.05 \\ 1 & \text{if} \quad Rel\_Sent > 0.05 \end{cases} \tag{3.3}$$

Further refinements of the procedure consist in, but are not limited to, assigning items a weight. The weight is based on the order of appearance of the items, as well as on the relevance of words (e.g., *earnings* and *dividends* are more relevant words than *commitment* and *initiative*).

3.3 Research Steps

This study is based on three main steps:

1. **Dataset.**

   In Caporin and Poli (2017) we collect from two news providers the firm-specific news stories about the S&P 100 constituents. These texts are composed by a headline and a story – long, on average, few dozens of words –, and are distinguished by topic. The release date of each news story, with minute-precision time, is reported.

2. **Application of the Developed Algorithm.**

   Following the procedure outlined in Section 3.2, we apply the algorithm to each news story.

3. **Results Evaluation.**

   In Caporin and Poli (2017) we try to extract the sentiment of news stories with a method less refined than the one illustrated above: with respect to Loughran and McDonald (2011), we introduce the above-mentioned negations (28 single words, 24 sequences of two words and 6 sequences of three words) and invert the sentiment each time a negation precedes a word, irrespective of whether the latter is positive or negative.

   A preliminary evaluation of the results consists in analyzing the discrepancies between the technique of Caporin and Poli (2017) and the more refined technique applied in step 2, by looking at how many news stories result in a different sentiment from the application of the two algorithms. Further assessment of the accuracy of the novel technique involves the manual assignment of sentiment to news stories by one or more researchers, and the comparison of the detected sentiment with the manually assigned one.

Appendices

Appendix A

Appendix

A.1 Assets, news topics and news summary stats

Table A1. Assets list with ticker symbol, company name, and sector.

Ticker	Name	Sector
AAPL	Apple	Consumer Goods
ABT	Abbott Laboratories	Healthcare
ACN	Accenture plc	Technology
AEP	American Electric Power Co., Inc.	Utilities
AIG	American International Group, Inc.	Financial
ALL	The Allstate Corporation	Financial
AMGN	Amgen Inc.	Healthcare
AMZN	Amazon.com, Inc.	Services
APA	Apache Corp.	Basic Materials
APC	Anadarko Petroleum Corporation	Basic Materials
AXP	American Express Company	Financial
BA	The Boeing Company	Industrial Goods
BAX	Baxter International Inc.	Healthcare
BHI	Baker Hughes Incorporated	Basic Materials
BIIB	Biogen Inc.	Healthcare
BK	The Bank of New York Mellon Corporation	Financial
BMY	Bristol-Myers Squibb Company	Healthcare
BRK.B	Berkshire Hathaway Inc.	Financial
C	Citigroup Inc.	Financial
CAT	Caterpillar Inc.	Industrial Goods
CELG	Celgene Corporation	Healthcare
CL	Colgate-Palmolive Co.	Consumer Goods
CMCSA	Comcast Corporation	Services
COF	Capital One Financial Corporation	Financial
COP	ConocoPhillips	Basic Materials
COST	Costco Wholesale Corporation	Services
CSCO	Cisco Systems, Inc.	Technology
CVS	CVS Health Corporation	Healthcare
CVX	Chevron Corporation	Basic Materials
DD	E. I. du Pont de Nemours and Company	Basic Materials
DIS	The Walt Disney Company	Services
DOW	The Dow Chemical Company	Basic Materials
EBAY	eBay Inc.	Services
EMC	EMC Corporation	Technology
EMR	Emerson Electric Co.	Industrial Goods
EXC	Exelon Corporation	Utilities
FCX	Freeport-McMoRan Inc.	Basic Materials
FDX	FedEx Corporation	Services
GD	General Dynamics Corporation	Industrial Goods

GE	General Electric Company	Industrial Goods
GILD	Gilead Sciences Inc.	Healthcare
GS	The Goldman Sachs Group, Inc.	Financial
HAL	Halliburton Company	Basic Materials
HD	The Home Depot, Inc.	Services
HON	Honeywell International Inc.	Industrial Goods
HPQ	HP Inc.	Technology
IBM	International Business Machines Corporation	Technology
INTC	Intel Corporation	Technology
JNJ	Johnson & Johnson	Healthcare
JPM	JPMorgan Chase & Co.	Financial
KO	The Coca-Cola Company	Consumer Goods
LLY	Eli Lilly and Company	Healthcare
LMT	Lockheed Martin Corporation	Industrial Goods
LOW	Lowe's Companies, Inc.	Services
MCD	McDonald's Corp.	Services
MDLZ	Mondelez International, Inc.	Consumer Goods
MDT	Medtronic plc	Healthcare
MET	MetLife, Inc.	Financial
MMM	3M Company	Industrial Goods
MO	Altria Group, Inc.	Consumer Goods
MON	Monsanto Company	Basic Materials
MRK	Merck & Co. Inc.	Healthcare
MSFT	Microsoft Corporation	Technology
NKE	NIKE, Inc.	Consumer Goods
NSC	Norfolk Southern Corporation	Services
ORCL	Oracle Corporation	Technology
OXY	Occidental Petroleum Corporation	Basic Materials
PEP	Pepsico, Inc.	Consumer Goods
PFE	Pfizer Inc.	Healthcare
PG	The Procter & Gamble Company	Consumer Goods
QCOM	QUALCOMM Incorporated	Technology
RTN	Raytheon Company	Industrial Goods
SBUX	Starbucks Corporation	Services
SLB	Schlumberger Limited	Basic Materials
SO	Southern Company	Utilities
SPG	Simon Property Group Inc.	Financial
T	AT&T, Inc.	Technology
TGT	Target Corp.	Services
TXN	Texas Instruments Inc.	Technology
UNH	UnitedHealth Group Incorporated	Healthcare
UNP	Union Pacific Corporation	Services
UPS	United Parcel Service, Inc.	Services
USB	U.S. Bancorp	Financial
UTX	United Technologies Corporation	Industrial Goods
WBA	Walgreens Boots Alliance, Inc.	Services
WFC	Wells Fargo & Company	Financial
WMB	Williams Companies, Inc.	Basic Materials
WMT	Wal-Mart Stores Inc.	Services
XOM	Exxon Mobil Corporation	Basic Materials

Table A.2: Topics available from the news providers.

	StreetAccount		**Thomson Reuters**
1.	Conjecture	1.	General Products
2.	Corporate Actions	2.	Production Guidance
3.	Earnings	3.	Business Deals
4.	Guidance	4.	M & A
5.	Litigation	5.	Officer Changes
6.	M & A	6.	Divestitures
7.	Management Changes	7.	Spin-Offs
8.	News	8.	New Business/Units/Subsidiary
9.	Regulatory	9.	New Markets
10.	Syndicate	10.	Equity Investments
11.	Up/Downgrades	11.	Share Repurchases
		12.	General Reorganization
		13.	Layoffs
		14.	Labor Issues
		15.	Class Action Lawsuit
		16.	Bankruptcy / Related
		17.	Initial Public Offerings
		18.	Equity Financing / Related
		19.	Debt Financing / Related
		20.	Indices Changes
		21.	Exchange Changes
		22.	Name Changes
		23.	Other Accounting
		24.	Restatements
		25.	Delinquent Filings
		26.	Change in Accounting Method/Policy
		27.	Corporate Litigation
		28.	Earnings Announcements
		29.	Negative Earnings Pre-Announcement
		30.	Positive Earnings Pre-Announcement
		31.	Other Pre-Announcement
		32.	Strategic Combinations
		33.	Regulatory/Company Investigation
		34.	Dividends
		35.	Debt Ratings
		36.	Special Events

A.2 Realized volatility measurement and jump testing

A huge literature dealing with modelling and forecasting the dynamic dependencies in financial market volatility has emerged over the past two decades. Until few years ago, most of the empirical results were based on the use of daily or coarser frequency data coupled with formulations within the GARCH or stochastic volatility model class. Then, high-frequency data started to be incorporated into longer-run volatility modelling and forecasting problems through the use of simple reduced-form time series models for non-parametric daily realized volatility measures based on the summation of intraday squared returns, see Andersen et al. (2003). Diffusive stochastic volatility models, however, have problems in explaining behaviour of asset prices, especially during market crashes and in general during turbulent periods, since they would require sometimes a volatility level too high for their formulation. As a solution, the total daily return variability has been decomposed into its continuous and discontinuous components based on the bipower variation measures developed by Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen and Shephard (2004). The empirical results in Andersen et al. (2007a) suggest that most of the predictable variation in the volatility stems from the strong own dynamic dependencies in the continuous price path variability, while the predictability of jumps is typically minor.

We assume that the scalar logarithmic asset price follows a standard jump-diffusion process:

$$dX_t = \mu_t dt + \sigma_t dW_t + dJ_t \tag{A.1}$$

where μ_t is predictable, σ_t is cadlag, $dJ_t = c_t dN_t$ where N_t is a non-explosive Poisson process whose intensity is an adapted stochastic process λ_t, the times of the jumps are $(\tau_j)_{j=1,\ldots,N_t}$ and c_j are i.i.d. adapted random variables measuring the size, which is always positive, of the jump at time τ_j.

Quadratic variation of the process over a time window T, e.g. one day, is defined as:

$$[X]_t^{t+T} = X_{[t+T]}^2 - X_t^2 - 2\int_t^{t+T} X_{s-} dX_s \tag{A.2}$$

where t indexes the day. It can be decomposed into its continuous and discontinuous component:

$$[X]_t^{t+T} = [X^c]_t^{t+T} + [X^d]_t^{t+T} \tag{A.3}$$

where $[X^c]_t^{t+T} = \int_t^{t+T} \sigma_s^2 ds$ and $[X^d]_t^{t+T} = \sum_{j=N_t}^{N_t+T} c_j^2$. To estimate these quantities, the time interval $[t, t+T]$ is divided into n subintervals of length $\delta = T/n$ and the evenly sampled returns are defined as:

$$\Delta_{j,t} X = X_{j\delta+t} - X_{(j-1)\delta+t}, \qquad j = 1, \ldots, n \tag{A.4}$$

The quadratic variation process and its separate components are, of course, not directly observable. Instead, we resort to popular model-free non-parametric consistent measures, including the familiar realized variance:

$$RV_\delta(X)_t = \sum_{j=1}^{n} (\Delta_j X)^2 \tag{A.5}$$

which converges in probability to $[X]_t^{t+T}$ as $\delta \to 0$.

The theory discussed above hinges on the notion of increasingly finer sampled high-frequency returns but, in practice, the sampling frequency is limited by the actual quotation or transaction frequency and the observed prices are contaminated by market microstructure frictions, including price discreteness and bid-ask spreads, which render the assumption of a semimartingale price process invalid at the tick-by-tick level. In response to this, we follow a relevant strand of the literature and compute our daily realized variance and jump measures from five-minute returns and we use the nearest preceding or concurrent price to each five-minute mark.

In order to separately measure the jump part, we rely on the corrected threshold bipower variation C-TBPV measure, a version of the corrected threshold multipower variation C-TMPV developed by Corsi et al. (2010), which consists in turn in a modification of the realized bipower variation of Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen and Shephard (2004):

$$\begin{aligned} \text{C-TBPV}_\delta(X)_t &= \mu_1^{-2} \text{C-TMPV}_\delta(X)_t^{1,1} \\ &= \mu_1^{-2} \textstyle\sum_{j=2}^{[T/\delta]} Z_1(\Delta X_j, \vartheta_j) Z_1(\Delta X_{j-1}, \vartheta_{j-1}) \end{aligned} \tag{A.6}$$

where $\mu_\alpha = E(|Z|^\alpha)$ for $Z \sim N(0,1)$.

The corrected threshold multipower variation is defined as:

$$\text{C-TMPV}_\delta(X)_t^{[\gamma_1,\ldots,\gamma_M]} = \delta^{1-\frac{1}{2}(\gamma_1+\ldots+\gamma_M)} \sum_{j=M}^{[T/\delta]} \prod_{k=1}^{M} Z_{\gamma k}(\Delta_{j-k+1} X, \vartheta_{j-k+1}) \tag{A.7}$$

the function $Z_\gamma(x,y)$ is:

$$Z_\gamma(x,y) = \begin{cases} |x|^\gamma & \text{if} \quad x^2 \le y \\ \frac{1}{2N(-c_\vartheta)\sqrt{\pi}}(\frac{2}{c_\vartheta^2}y)^{\frac{\gamma}{2}}\Gamma(\frac{\gamma+1}{2}, \frac{c_\vartheta^2}{2}) & \text{if} \quad x^2 \ge y \end{cases} \tag{A.8}$$

where $N(x)$ is the standard normal cumulative function, $\Gamma(\alpha,x)$ is the upper incomplete gamma function, $\vartheta = c_\vartheta^2\sigma^2$ and σ^2 is the variance of $\Delta_j X$ under the assumption that $\Delta_j X \sim N(0,\sigma^2)$. Following Corsi et al. (2010), we set $c_\vartheta = 3$.

As $\delta \to 0$, C-TBPV converges to $\int_t^{t+T} \sigma^2(s)ds$

The difference between the realized variance and the corrected threshold bipower variation consistently estimates the part of the quadratic variation due to jumps:

$$RV_\delta(X)_T - \text{C-TBPV}_\delta(X)_T \xrightarrow[\delta\to 0]{P} [X^d]_t^{t+T} \tag{A.9}$$

As $\delta \to 0$, the test statistic

$$\text{C-T}_Z = \delta^{\frac{1}{2}} \cdot \frac{(RV_\delta(X)_T - \text{C-TBPV}_\delta(X)_T) \cdot RV_\delta(X)_T^{-1}}{\sqrt{\left(\frac{\pi^2}{4} + \pi - 5\right)\max\left(1, \frac{\text{C-TTriPV}_\delta(X)_T}{(\text{C-TBPV}_\delta(X)_T)^2}\right)}} \tag{A.10}$$

where $\text{C-TTriPV}_\delta(X)_T$ is a quarticity estimator, see Corsi et al. (2010), is asymptotically standard normally distributed under the null hypothesis of no jumps.

Based on the above jump detection test statistic, the realized measure of the jump contribution to the quadratic variation of the logarithmic price process is then measured by:

$$\widehat{J}_t = I_{(\text{C-T}_Z > \Phi_\alpha)} \cdot (RV_t - BPV_t)^+ \tag{A.11}$$

where $I_{(\cdot)}$ denotes the indicator function and Φ_α refers to the appropriate critical value from the standard normal distribution.

Consequently, the realized measure for the integrated variance is:

$$\widehat{C}_t = RV_t - \widehat{J}_t \tag{A.12}$$

We use a critical value of $\alpha = 99.9\%$, in line with recent studies.

Table A.3 provides the summary statistics of the distribution across all assets of the percentage of jump days, and Fig. A.1 shows the corresponding histogram, grouping the frequencies for graphic clarity. Frequencies range from 1.57% to 4.82%, and are close to the mean of approximately 3% for most of the assets.

Table A.3: Basic summary statistics of assets' percentage of days with at least one jump.

	Min	**Max**	**Mean**	**Median**
Jump days (%)	1.57	4.82	3.02	2.90

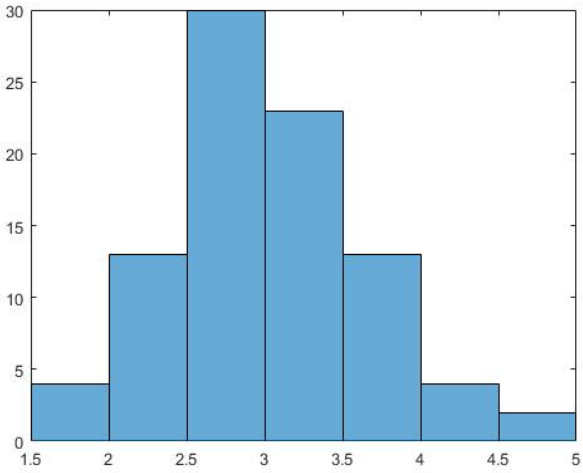

Figure A.1: Distribution of assets' percentage of days with at least one jump.

A.3 Most selected regressors in the log HAR-TCJN model by subsample

Table A.4: Most selected regressors in the log HAR-TCJN model. Sample: 2005 – 2007.

		Past Volatility Components					**% Selected**	**% Pos**	**% Neg**
		$\log C_d$					100.00	100.00	0.00
		$\log C_w$					98.88	98.88	0.00
		$\log C_m$					94.38	94.38	0.00
		$\log (1 + J_d)$					21.35	21.35	0.00
Macro	**Firm-Specific**	**News**	**Importance**	**Topic**	**Time Aggregation**	**Measure**	**% selected**	**% pos**	**% neg**
X		RSALES			month	flag if surp < 0	40.45	40.45	0.00
X		JOBLESS			overnight	flag if surp $\neq 0$	39.33	39.33	0.00
	X	EPS			day	flag for announcement	38.20	38.20	0.00
	X	SA news		earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	33.71	33.71	0.00
	X	SA news		earnings	flow: day-to-day	flag if Δ n. words $\neq 0$	32.58	32.58	0.00
X		RSALES			month	flag if surp $\neq 0$	30.34	30.34	0.00
X		JOBLESS			month	square surp	30.34	30.34	0.00
X		JOBLESS			month	flag if surp > 0	29.21	29.21	0.00
X		BOP			week	flag if surp > 0	28.09	28.09	0.00
X		GDP			month	square surp	28.09	28.09	0.00
X		NFP			month	surp	26.97	0.00	26.97
X		FOMC			overnight	flag for announcement	25.84	0.00	25.84
X		CPI			month	flag if surp < 0	23.60	0.00	23.60
	X	EPS			day	flag if surp $\neq 0$	23.60	23.60	0.00
	X	TR news	low	earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	22.47	22.47	0.00
	X	SA news		earnings	flow: day-to-day	flag if Δ n. news stories < 0	22.47	22.47	0.00
X		JOBLESS			month	log surp	20.22	20.22	0.00
	X	TR news	low	earnings	flow: day-to-day	flag if Δ n. words $\neq 0$	20.22	20.22	0.00
	X	SA news		earnings	flow: day-to-day	flag if Δ n. words < 0	20.22	20.22	0.00
X		JOBLESS			overnight	flag if surp > 0	19.10	19.10	0.00
X		BOP			week	sqrt surp	19.10	19.10	0.00
X		GDP			overnight	flag for announcement	19.10	19.10	0.00
X		CCONF			overnight	flag if surp > 0	19.10	0.00	19.10
X		RSALES			month	surp	17.98	0.00	17.98
	X	EPS			day	flag if surp > 0	17.98	17.98	0.00
	X	TR news	low	earnings	flow: day-to-day	flag if Δ n. news stories < 0	17.98	17.98	0.00
X		NFP			month	square surp	16.85	0.00	16.85
X		JOBLESS			day	flag for announcement	16.85	16.85	0.00
X		GDP			overnight	flag if surp > 0	16.85	16.85	0.00
X		CPI			month	flag if surp $\neq 0$	16.85	0.00	16.85

Notes: Ranking of regressors (past volatility components plus the thirty most frequently selected news measures) by percentage of stocks for which they are selected by LASSO in the log HAR-TCJN model, percentage of positive and percentage of negative coefficients. Sample: Feb 2005 – Dec 2007 (expansion).

Table A.5: Most selected regressors in the log HAR-TCJN model. Sample: 2007 – 2009.

		Past Volatility Components					**% Selected**	**% Pos**	**% Neg**
		$\log C_d$					100.00	100.00	0.00
		$\log C_w$					100.00	100.00	0.00
		$\log C_m$					87.64	87.64	0.00
		$\log (1 + J_d)$					12.36	11.24	1.12
Macro	**Firm-Specific**	**News**	**Importance**	**Topic**	**Time Aggregation**	**Measure**	**% Selected**	**% Pos**	**% Neg**
X		GDP			week	flag for announcement	48.31	0.00	48.31
X		CPI			week	flag if surp < 0	47.19	47.19	0.00
X		GDP			week	flag if surp $\neq 0$	44.94	0.00	44.94
X		CPI			day	flag if surp < 0	44.94	44.94	0.00
X		FOMC			day	flag for announcement	41.57	41.57	0.00
X		GDP			month	log surp	40.45	0.00	40.45
X		JOBLESS			month	flag if surp > 0	35.96	35.96	0.00
X		JOBLESS			month	flag if surp < 0	34.83	0.00	34.83
X		FOMC			month	flag if surp < 0	33.71	0.00	33.71
X		GDP			week	square surp	28.09	28.09	0.00
X		NFP			month	surp	25.84	0.00	25.84
X		BOP			day	flag if surp < 0	25.84	25.84	0.00
X		INDPROD			week	flag if surp > 0	25.84	25.84	0.00
X		NFP			month	log surp	24.72	0.00	24.72
	X	EPS			day	flag for announcement	24.72	24.72	0.00
X		RSALES			week	flag if surp > 0	23.60	0.00	23.60
X		RSALES			month	flag for announcement	22.47	22.47	0.00
X		RSALES			month	flag if surp $\neq 0$	21.35	21.35	0.00
X		NFP			day	flag for announcement	21.35	21.35	0.00
X		JOBLESS			week	flag if surp < 0	21.35	0.00	21.35
X		RSALES			week	sqrt surp	17.98	0.00	17.98
X		NFP			day	flag if surp > 0	17.98	17.98	0.00
X		NFP			day	flag if surp $\neq 0$	17.98	17.98	0.00
X		INDPROD			day	square surp	17.98	0.00	17.98
X		CPI			week	flag if surp $\neq 0$	17.98	17.98	0.00
	X	SA news		earnings	flow: day-to-day	flag if Δ n. news stories < 0	17.98	17.98	0.00
X		BOP			week	flag if surp < 0	16.85	16.85	0.00
X		BOP			day	flag for announcement	16.85	16.85	0.00
X		RSALES			month	flag if surp < 0	15.73	15.73	0.00
X		BOP			day	flag if surp $\neq 0$	15.73	15.73	0.00

Notes: Ranking of regressors (past volatility components plus the thirty most frequently selected news measures) by percentage of stocks for which they are selected by LASSO in the log HAR-TCJN model, percentage of positive and percentage of negative coefficients. Sample: Dec 2007 – Jun 2009 (contraction).

Table A.6: Most selected regressors in the log HAR-TCJN model. Sample: 2009 – 2015.

		Past Volatility Components					**% Selected**	**% Pos**	**% Neg**
		$\log C_d$					100.00	100.00	0.00
		$\log C_w$					100.00	100.00	0.00
		$\log C_m$					100.00	100.00	0.00
		$\log (1 + J_d)$					43.82	43.82	0.00
Macro	**Firm-Specific**	**News**	**Importance**	**Topic**	**Time Aggregation**	**Measure**	**% Selected**	**% Pos**	**% Neg**
	X	EPS			day	flag for announcement	53.93	53.93	0.00
X		FOMC			week	flag for announcement	49.44	49.44	0.00
X		CCONF			overnight	flag for announcement	44.94	0.00	44.94
	X	SA news		earnings	flow: day-to-day	flag if Δ n. words < 0	33.71	33.71	0.00
	X	SA news		all	flow: day-to-day	flag if Δ n. words < 0	31.46	31.46	0.00
X		FOMC			month	surp	29.21	29.21	0.00
	X	EPS			day	flag if surp $\neq 0$	29.21	29.21	0.00
	X	TR news	low	earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	29.21	29.21	0.00
X		CPI			month	flag for announcement	28.09	28.09	0.00
X		FOMC			month	flag if surp $\neq 0$	26.97	26.97	0.00
	X	SA news		all	flow: day-to-day	flag if Δ n. words $\neq 0$	26.97	26.97	0.00
	X	SA news		all	day	flag if n. news stories ≥ 2	26.97	26.97	0.00
	X	SA news		earnings	flow: day-to-day	flag if Δ n. words $\neq 0$	25.84	25.84	0.00
	X	SA news		earnings	flow: day-to-day	flag if Δ n. news stories < 0	25.84	25.84	0.00
	X	TR news	low	earnings	flow: day-to-day	flag if Δ n. words $\neq 0$	24.72	24.72	0.00
X		PPI			week	flag if surp < 0	23.60	23.60	0.00
X		NFP			month	surp	23.60	0.00	23.60
X		FOMC			month	flag if surp > 0	23.60	23.60	0.00
	X	SA news		earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	22.47	22.47	0.00
	X	SA news		earnings	day	flag if n. news stories ≥ 2	21.35	21.35	0.00
X		JOBLESS			overnight	flag if surp < 0	20.22	20.22	0.00
X		JOBLESS			overnight	flag for announcement	20.22	20.22	0.00
	X	SA news		all	day	n. news stories	20.22	20.22	0.00
		Google Trends			week	log GSI	19.10	17.98	1.12
X		JOBLESS			day	flag if surp > 0	19.10	19.10	0.00
	X	SA news		up/downgrades	flow: day-to-day	flag if Δ n. words $\neq 0$	19.10	19.10	0.00
	X	SA news		all	flow: day-to-day	flag if Δ n. news stories < 0	19.10	19.10	0.00
	X	SA news		all	flow: day-to-day	persistence/interaction	17.98	17.98	0.00
		Google Trends			week	GSI	16.85	12.36	4.49
	X	TR news	high	earnings	flow: day-to-day	flag if Δ n. news stories $\neq 0$	16.85	16.85	0.00

Notes: Ranking of regressors (past volatility components plus the thirty most frequently selected news measures) by percentage of stocks for which they are selected by LASSO in the log HAR-TCJN model, percentage of positive and percentage of negative coefficients. Sample: Jun 2009 – Feb 2015 (expansion).

A.4 Outliers adjustment for the HAR-TCJN model

In a very few cases, the HAR-TCJN model yields RV forecasts that are extremely close to zero or higher than some thousands, which are in both cases unreliable values. In order to overcome these degeneracies, we adopt a smoothing adjustment based on the comparison of the RV forecasts from the HAR-TCJ model ($\widehat{RV}_{HAR-TCJ,t}$) and from the HAR-TCJN model ($\widehat{RV}_{HAR-TCJN,t}$), and obtain an adjusted forecast that we call $\widehat{RV}_{HAR-TCJNadj,t}$. The adjustment process is illustrated in Table A.7.

Table A.7

Condition	$\widehat{RV}_{HAR-TCJNadj,t}$
$\widehat{RV}_{HAR-TCJN,t} \leq t_{LL}$	$(t_{LL} + t_L)/2$
$t_{LL} < \widehat{RV}_{HAR-TCJN,t} < t_L$	$t_L + (\widehat{RV}_{HAR-TCJN,t} - t_L)/2$
$t_L \leq \widehat{RV}_{HAR-TCJN,t} \leq t_H$	$\widehat{RV}_{HAR-TCJN,t}$
$t_H < \widehat{RV}_{HAR-TCJN,t} < t_{HH}$	$t_H + (\widehat{RV}_{HAR-TCJN,t} - t_H)/2$
$t_{HH} \leq \widehat{RV}_{HAR-TCJN,t}$	$(t_H + t_{HH})/2$

where:
$t_{LL} = \widehat{RV}_{HAR-TCJ,t}/4$
$t_L = \widehat{RV}_{HAR-TCJ,t}/2$
$t_H = \widehat{RV}_{HAR-TCJ,t} \cdot 2$
$t_{HH} = \widehat{RV}_{HAR-TCJ,t} \cdot 4$

Bibliography

Andersen, T.G.; Bollerslev, T.; Diebold, F.X.; Labys, P. Modeling and Forecasting Realized Volatility. *Econometrica* **2003**, *71*, 579–625.

Andersen, T.G.; Bollerslev, T.; Diebold, F.X. Roughing It Up: Including Jump Components in the Measurement, Modeling, and Forecasting of Return Volatility. *Rev Econ Stat* **2007**, *89*, 701–720.

Andersen, T.G.; Bollerslev, T.; Diebold, F.X.; Vega, C. Real-time price discovery in global stock, bond and foreign exchange markets. *J Int Econ* **2007**, *73*, 251–277.

Andersen, T.G.; Bollerslev, T.; Dobrev, D. No-arbitrage semi-martingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: Theory and testable distributional implications. *J Econometrics* **2007**, *138*, 125–180.

Andrei, D.; Hasler, M. Investor Attention and Stock Market Volatility. *Rev Financ Stud* **2015**, *28*, 33–72.

Audrino, F.; Knaus, S.D. Lassoing the HAR Model: A Model Selection Perspective on Realized Volatility Dynamics. *J Economet Rev* **2016**, *35*, 1485–1521.

Antweiler, W.; Frank, M.Z. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *J Financ* **2004**, *59*, 1259–1294.

Bajgrowicz, P.; Scaillet, O.; Treccani, A. Jumps in High-Frequency Data: Spurious Detections, Dynamics, and News. *Management Sci*, **2016**, *62*, 2198–2217.

Baker, S.; Fradkin, A. What Drives Job Search? Evidence from Google Search Data. *Discussion Papers, Stanford Instititute for Economic Policy Research* **2011**.

Baklaci, H.F.; Tunc, G.; Aydogan, B.; Vardar, G. The Impact of Firm-Specific Public News on Intraday Market Dynamics: Evidence from the Turkish Stock Market. *Emerg Mark Financ Tr* **2011**, *47*, 99–119.

Berry, T.D.; Howe, K.M. Public Information Arrival. *J Financ* **1994**, *49*, 1331–1346.

Barndorff-Nielsen, O.E.; Shephard, N. Power and Bipower Variation with Stochastic Volatility and Jumps. *J Financ Economet* **2004**, *2*, 1–37.

Barndorff-Nielsen, O.E.; Shephard, N. Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation. *J Financ Economet* **2006**, *4*, 1–30.

Beine, M.; Lahaye, J.; Laurent, S.; Neely, C.J.; Palm, F.C. Central Bank Intervention and Exchange Rate Volatility, its Continuous and Jump Components. *Int J Financ Econ* **2007**, *12*, 201–223.

Bjursell, J.; Wang, G.H.K.; Webb, R.I. Jumps and Trading Activity in Interest Rate Futures Markets: The Response to Macroeconomic Announcements. *Asia-Pac J Financ St*, **2013**, *42*, 689–723.

Birz, G.; Lott, J.R., Jr. The Effect of Macroeconomic News on Stock Returns: New Evidence from Newspaper Coverage. *J Bank Financ* **2011**, *35*, 2791–2800.

Bollerslev, T. Generalized Autoregressive Conditional Heteroskedasticity. *J Econometrics* **1986**, *31*, 307–327.

Bollerslev, T.; Ghysels, E. Periodic Autoregressive Conditional Heteroscedasticity. *J Bus Econ Stat* **1996**, *14*, 139–151.

Bollerslev, T.; Law, T.H.; Tauchen, G. Risk, jumps, and diversification. *J Econometrics* **2008**, *144*, 234–256.

Bomfim, A.N. Pre-announcement Effects, News Effects, and Volatility: Monetary Policy and the Stock Market. *J Bank Financ* **2003**, *27*, 133–151.

Borovkova, S.; Mahakena, D. News, volatility and jumps: the case of natural gas futures. *Quant Financ*, **2015**, *15*, 1217–1242.

Bouchaud, J.P.; Kockelkoren, J.; Potters, M. Random walks, liquidity molasses and critical response in financial markets. *Quant Financ* **2006**, *6*, 115–123.

Boudt, K.; Croux, C.; Laurent, S. Robust estimation of intraweek periodicity in volatility and jump detection. *J Empir Financ*, **2011**, *18*, 353–367.

Boudt, K.; Petitjean, M. Intraday Liquidity Dynamics and News Releases Around Price Jumps: Evidence from the DJIA Stocks. *J Financ Mark*, **2014**, *17*, 121–149.

Brailsford, T.J. The Empirical Relationship between Trading Volume, Returns and Volatility. *Account Financ* **1996**, *36*, 89–111.

Brenner, M.; Pasquariello, P.; Subrahmanyam, M. On the Volatility and Comovement of U.S. Financial Markets around Macroeconomic News Announcements. *J Financ Quant Anal* **2009**, *44*, 1265–1289.

Busse, J.A.; Green, T.C. Market Efficiency in Real Time. *J Financ Econ* **2002**, *65*, 415–437.

Caporin, M.; Kolokolov, A.; Renò, R. Systemic Co-jumps. *J Financ Econ*, forthcoming.

Caporin, M.; Poli, F. Building News Measures from Textual Data and an Application to Volatility Forecasting. *Working Paper*

Chatrah, A.; Miao, H.; Ramchander, S.; Villupuram, S. Currency jumps, cojumps and the role of macro news. *J Int Money Financ*, **2014**, *40*, 42–62.

Clark, P. A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. *Econometrica* **1973**, *41*, 135–155.

Corsi, F. A Simple Approximate Long-Memory Model of Realized Volatility. *J Financ Economet* **2009**, *7*, 174–196.

Corsi, F.; Pirino, D.; Renò, R. Threshold Bipower Variation and the Impact of Jumps on Volatility Forecasting. *J Econometrics* **2010**, *159*, 276–288.

Cui, J.; Zhao, H. Intraday jumps in China's Treasury bond market and macro news announcements. *Int Rev Econ Financ*, **2015**, *39*, 211–223.

Cutler, D.M.; Poterba, J.M.; Summers, L.H. What Moves Stock Prices? *J Portfolio Manage* **1989**, *15*, 4–12.

Da, Z.; Engelberg, J.; Gao, P. In Search of Attention. *J Financ* **2011**, *66*, 1461–1499.

Da, Z.; Engelberg, J.; Gao, P. The Sum of All FEARS Investor Sentiment and Asset Prices. *Rev Financ Stud* **2015**, *28*, 1–32.

D'Amuri, F.; Marcucci, J. The Predictive Power of Google Searches in Forecasting Unemployment. *Banca D'Italia Working Papers, n. 891* **2012**.

Dewatcher, H.; Erdemlioglu, D.; Gnabo, J-Y. The intra-day impact of communication on euro-dollar volatility and jumps. *J Int Money Financ*, **2014**, *43*, 131–154.

Diebold, F.X.; Mariano, R.S. Comparing Predictive Accuracy. *J Bus Econ Stat* **1995**, *13*, 253–263.

Dimpfl, T.; Jank, S. Can Internet Search Queries Help to Predict Stock Market Volatility? *Eur Financ Manag* **2016**, *22*, 171–192.

Dougal, C.; Engelberg, J.; García, D.; Parsons, C.A. Journalists and the Stock Market. *Rev Financ Stud* **2012**, *25*, 639–679.

Dungey, M.; McKenzie, M.; Smith, L.V. Empirical Evidence on Jumps in the Term Structure of the US Treasury Market. *J Empir Financ*, **2009**, *16*, 430–445.

Ederington, L.H.; Lee, J.H. How Markets Process Information: News Releases and Volatility. *J Financ* **1993**, *48*, 1161–1191.

Elder, J.; Miao, H.; Ramchander, S. Jumps in Oil Prices: The Role of Economic News. *Energ J*, **2013**, *34*, 217–237.

Engle, R.F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **1982**, *50*, 987–1007.

Engle R.F.; Rangel J.G. The Spline-GARCH Model for Low-Frequency Volatility and Its Global Macroeconomic Causes. *Rev Financ Stud* **2008**, *21*, 1187–1222.

Epps, T.W.; Epps, M.L. The Stochastic Dependence of Security Price Changes and Transaction Volumes: Implications for the Mixture-of-Distributions Hypothesis. *Econometrica* **1976**, *44*, 305–321.

Evans, K.P. Intraday jumps and US macroeconomic news announcements. *J Bank Financ* **2011**, *35*, 2511–2527.

Fang, L.; Peress, J. Media Coverage and the Cross-section of Stock Returns. *J Financ* **2009**, *64*, 2023–2052.

Flannery, M.J.; Protopapadakis, A.A. Macroeconomic Factors do Influence Aggregate Stock Returns. *Rev Financ Stud* **2002**, *15*, 751–782.

Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **2010**, *33*, 1–22.

Frömmel, M.; Han, X.; van Gysegem, F. Further Evidence on Foreign Exchange Jumps and News Announcements. *Emerg Mark Financ Tr*, **2015**, *51*, 774–787.

Gallo, G.; Pacini, B. The Effects of Trading Activity on Market Volatility. *Eur J Financ* **2000**, *6*, 163–175.

García, D. Sentiment during Recessions. *J Financ* **2013**, *68*, 1267–1300.

Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S.; Brilliant, L. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature* **2009**, *457*, 1012–1014.

Gloß-Klußmann, A.; Hautsch, N. When Machines Read the News: Using Automated Text Analytics to Quantify High Frequency News-Implied Market Reactions. *J Empir Financ* **2011**, *18*, 321–340.

Goddard, J.; Kita, A.; Wang, Q. Investor Attention and FX Market Volatility. *J Int Finan Markets Inst Money* **2015**, *38*, 79–96.

Guo, J.F.; Ji, Q. How does Market Concern Derived from the Internet Affect Oil Prices? *Appl Energ* **2013**, *112*, 1536–1543.

Hamid, A.; Heiden, M. Forecasting Volatility with Empirical Similarity and Google Trends. *J Econ Behav Organ* **2015**, *117*, 62–81.

Hautsch, N.; Hess, D.; Veredas, D. The Impact of Macroeconomic News on Quote Adjustments, Noise, and Informational Volatility. *J Bank Financ* **2011**, *35*, 2733–2746.

Ho, K.Y.; Shi, Y.; Zhang, Z. How Does News Sentiment Impact Asset Volatility? Evidence from Long Memory and Regime-Switching Approaches. *N Am J Econ Financ* **2013**, *26*, 436–456.

Huang, X. Macroeconomic News Announcements, Systemic Risk, Financial Market Volatility and Jumps. *Economics Discussion Series 2015-097. Washington: Board of Governors of the Federal Reserve System*, **2015**.

James, G.; Hastie, T.; Witten, D.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer-Verlag: New York, NY, USA, 2013.

Janssen, G. Public Information Arrival and Volatility Persistence in Financial Markets. *Eur J Financ* **2004**, *10*, 177–197.

Jiang, G.J.; Lo, I.; Verdelhan, A. Information Shocks, Liquidity Shocks, Jumps, and Price Discovery: Evidence from the U.S. Treasury Market. *J Financ Quant Anal*, **2011**, *46*, 527–551.

Jones, C.M.; Lamont, O.; Lumsdaine, R.L. Macroeconomic News and Bond Market Volatility. *J Financ Econ* **1998**, *47*, 315–337.

Joulin, A.; Lefevre, A.; Grunberg, D.; Bouchaud, J.P. Stock price jumps: News and volume play a minor role. *Wilmott Magazine*, **2008**, 1–7.

Kalev, P.S.; Liu, W.M.; Pham, P.K.; Jarnecic, E. Public Information Arrival and Volatility of Intraday Stock Returns. *J Bank Financ* **2004**, *28*, 1441–1467.

Kim, D.; Kon, S.J. Alternative Models for the Conditional Heteroscedasticity of Stock Returns. *J Bus* **1994**, *67*, 563–598.

Kraussl, R.; Mirgorodskaya, E. Media, Sentiment and Market Performance in the Long Run. *Eur J Financ* **2016**, *22*, 1–24.

Lahaye, J.; Laurent, S.; Neely, C.J. Jumps, cojumps and macro announcements. *J Appl Econom* **2011**, *26*, 893–921.

Lamoureux, C.G.; Lastrapes, W.D. Heteroskedasticity in Stock Return Data: Volume versus GARCH Effects. *J Financ* **1990**, *45*, 221–229.

Larkin, F.; Ryan, C. Good News: Using News Feeds with Genetic Programming to Predict Stock Prices. In *Genetic Programming. EuroGP 2008. Lecture Notes in Computer Science, vol 4971*; ONeill M. et al. (eds); Springer, Berlin, Heidelberg; 2008; pp. 49-60

Lee, S.S.; Mykland, P.A. Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics. *Rev Financ Stud* **2008**, *21*, 2535–2563.

León, Á.; Sebestyén, S. New measures of monetary policy surprises and jumps in interest rates. *J Bank Financ*, **2012**, *36*, 2323–2343.

Li, L.; Engle, R.F. Macroeconomic Announcements and Volatility of Treasury Futures. *Discussion Paper, Dept. of Economics, University of California, San Diego* **1998**.

Loughran, T.; McDonald, B. When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *J Financ* **2011**, *66*, 35–65.

Madhavan, A. Market microstructure: A survey. *J Financ Mark* **2000**, *3*, 205–258.

Maheu, J.M.; McCurdy, T.H. News Arrival, Jump Dynamics, and Volatility Components for Individual Stock Returns. *J Financ* **2004**, *59*, 755–793.

Martens, M.; van Dijk, D.; de Pooter, M. Forecasting S&P 500 Volatility: Long Memory, Level Shifts, Leverage Effects, Day-of-the-Week Seasonality, and Macroeconomic Announcements. *Int J Forecasting* **2009**, *25*, 282–303.

McMillan, D.G.; García, R.Q. Does Information Help Intra-Day Volatility Forecasts? *J Forecasting* **2013**, *32*, 1–9.

Meurer, R.; Santos, A.A.P.; Turatti, D.E. Monetary policy surprises and jumps in interest rates: evidence from Brazil. *J Econ Stud*, **2015**, *42*, 893–907.

Miao, H.; Ramchander, S.; Zumwalt, K. S$P 500 Index-Futures Price Jumps and Macroeconomic News. *J Futures Markets* **2014**, *34*, 980–1001.

Mitchell, M.L.; Mulherin, J.H. The Impact of Public Information on the Stock Market. *J Financ* **1994**, *49*, 923–950.

Mizrach, B. Jumps and Cojumps in Subprime Home Equity Derivatives. *J Portfolio Manage*, **2012**, *38*, 136–146.

Mood, A. The distribution theory of runs. *Ann Math Statist*, **1940**, *11*, 367–392.

Neely, C.J. A Survey of Announcement Effects on Foreign Exchange Volatility and Jumps. *Federal Reserve Bank of St. Louis Review*, **2011**, *93*, 361–407.

Omran, M.F.; McKenzie, E. Heteroscedasticity in Stock Returns Data Revisited: Volume versus GARCH Effects. *Appl Finan Econ* **2000**, *10*, 553–560.

Pavlou, M.; Ambler, G.; Seaman, S.; De Iorio, M.; Omar, R.Z. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statist Med* **2016**, *35*, 1159–1177.

Preis, T.; Reith, D.; Stanley, H.E. Complex Dynamics of our Economic Life on Different Scales: Insights from Search Engine Query Data. *Philos T Roy Soc A* **2010**, *368*, 5707–5719.

Rangel, J.G. Macroeconomic News, Announcements, and Stock Market Jump Intensity Dynamics. *J Bank Financ* **2011**, *35*, 1263–1276.

Riordan, R.; Storkenmaier, A.; Wagener, M.; Zhang, S. Public Information Arrival: Price Discovery and Liquidity in Electronic Limit Order Markets. *J Bank Financ* **2013**, *37*, 1148–1159.

Roll, R. R2. *J Financ* **1988**, *43*, 541–566.

Savor, P.; Wilson, M. How Much Do Investors Care About Macroeconomic Risk? Evidence from Scheduled Economic Announcements. *J Financ Quant Anal* **2013**, *48*, 343–375.

Smales, L.A. Time-Variation in the Impact of News Sentiment. *Int Rev Finan Anal* **2015**, *37*, 40–50.

Smith, G.P. Google Internet Search Activity and Volatility Prediction in the Market for Foreign Currency. *Financ Res Lett* **2012**, *9*, 103–110.

Solomon, D.; Soltes, E.; Sosyura, D. Winners in the Spotlight: Media Coverage of Fund Holdings as a Driver of Flows. *J Financ Econ* **2014**, *113*, 53–72.

Tauchen, G.E.; Pitts, M. The Price Variability-Volume Relationship on Speculative Markets. *Econometrica* **1983**, *51*, 485–505.

Tauchen, G.; Zhou, H. Realized jumps on financial markets and predicting credit spreads. *J Econometrics* **2011**, *160*, 102–118.

Tetlock, P.C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *J Financ* **2007**, *62*, 1139–1168.

Tetlock, P.C.; Saar-Tsechansky, M.; Macskassy, S. More than Words: Quantifying Language to Measure Firms' Fundamentals. *J Financ* **2008**, *63*, 1437–1467.

Tetlock, P.C. Does Public Financial News Resolve Asymmetric Information? *Rev Financ Stud* **2010**, *23*, 3520–3557.

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J Roy Stat Soc B Met* **1996**, *58*, 267–288.

Schwert, G.W. Why Does Stock Market Volatility Change over Time? *J Financ* **1989**, *44*, 1115–1153.

Vlastakis, N.; Markellos, R.N. Information Demand and Stock Market Volatility. *J Bank Financ* **2012**, *36*, 1808–1821.

Vozlyublennaia, N. Investor Attention, Index Performance, and Return Predictability. *J Bank Financ* **2014**, *41*, 17–35.

Vrugt, E.B. U.S. and Japanese Macroeconomic News and Stock Market Volatility in Asia-Pacific. *Pac-Basin Financ J* **2009**, *17*, 611–627.

Wright, J.H.; Zhou, H. Bond risk premia and realized jump risk. *J Bank Financ* **2009**, *33*, 2333–2345.

Zhang, Y.; Swanson, P.E.; Prombutr, W. Measuring Effects on Stock Returns of Sentiment Indexes Created from Stock Message Boards. *J Financ Res* **2012**, *35*, 79–114.

Zhang, Y.; Feng, L.; Jin, X.; Shen, D.; Xiong, X.; Zhang, W. Internet Information Arrival and Volatility of SME PRICE INDEX. *Physica A* **2014**, *399*, 70–74.

Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J Roy Stat Soc B*, **2005**, *67*, 301–320.