Diagnostics for topic modelling. The dubious joys of making quantitative decisions in a qualitative environment

Andrea Sciandra, Matilde Trevisani and Arjuna Tuzzi

Abstract Diagnostics is a crucial component of any topic modelling application. However, available measures seldom offer indisputable and consistent solutions. We analyse the score distribution of a large set of intrinsic measures by varying two model inputs: text length and topic number. The first aim is to identify an ideal text length (or range of) by exploring per-length diagnostic distributions over the topic number. The second aim, once the optimal text length has been set, is to select the best model (or candidates) by comparing different specifications that include document metadata. We will also detect any conflict or ambivalence in the solutions produced by the different diagnostics.

Key words: diagnostic measures, topic modelling, structural topic modelling, model selection

1 Introduction

Diagnostic measures are a crucial component of any topic modelling application. This is because several decisions need to be made before estimating the final model for meaningful results to be obtained. However, identifying and justifying these choices is a challenging journey, and available measures seldom offer indisputable and consistent solutions. In general, selecting an appropriate topic model (TM) involves a variety of trade-offs and judgments by the human researcher. In this study,

Andrea Sciandra
Dipartimento FISPPA, Università di Padova, e-mail: andrea.sciandra@unipd.it

Matilde Trevisani
DEAMS "Bruno de Finetti", Università di Trieste, e-mail: matilde.trevisani@deams.units.it

Arjuna Tuzzi
Dipartimento FISPPA, Università di Padova, e-mail: arjuna.tuzzi@unipd.it

we discuss the role of diagnostic measures towards selecting the most appropriate topic structure of a diachronic corpus.

Using a TM as an unsupervised tool involves focusing on how the learned topics align with human evaluations and help differentiate between aspects of a discourse. Until recently, the evaluation of such models has been ad hoc and application-specific, ranging from a fully automated intrinsic approach to a manually crafted extrinsic approach. Intrinsic evaluation, based on statistical measures, can be problematic because the measures do not account for domain relevance. Meanwhile, extrinsic evaluations are hand-constructed and often costly to perform for domain-specific topics. In any case, the real-world deployment of topic models requires time-consuming expert verification and model refinement to gain semantically meaningful topics within the domain of analysis.

Because of the ability of intrinsic measures to standardise, automate and scale the evaluation of TMs, the analyst generally picks one or more diagnostics of this type to be guided in the landscape of possibilities from which to choose the best model. Two broad classes can be envisaged: diagnostics that measure the predictive accuracy of the model (of which perplexity and marginal probability are the most well-known and widely applied) and diagnostics that assess the quality of topics (of which semantic coherence and Kullbach-Leibler (KL) topic divergence are among the most frequent instances, although the diversity of metrics is greater in this class). Moreover, any diagnostic can be variously implemented. Importantly, assessing a TM based on its predictive ability generally involves choices that are misleading, if not conflicting, in their judgements based on the quality of topics.

In this paper, we analyse the score distribution of a large set of intrinsic measures by varying two model inputs: text (or text partition) length and topic number. While topic number selection has been extensively studied, the impact of text length on a model's performance has been rarely addressed. In particular, some studies focus on the relationship between text length and topic number [8] or propose a document partition to improve model estimation [5]. In this study, we widen the research scope by first seeking to identify an ideal text length (or range) that optimises the selected diagnostics by exploring their distribution over the topic number. Once we have determined the optimal document chunking, we will analyse what the same diagnostics suggest for model selection by comparing alternative specifications that include document metadata. We will also detect any conflict or ambivalence in the solutions produced by the different diagnostics.

Within the panorama of topic modelling, we chose for our simulation study the structural topic model (STM, [7]) because it is a natural extension of the most famous and widely used TM, i.e. the Latent Dirichlet Allocation (LDA): it allows for correlation both between topics and between the topics and document-level covariates. At present, STM is very popular among topic modelling practitioners compared to other LDA generalisations. Moreover, it suits our application in which the effect of corpus metadata on topic determination is of interest.

Section 2 summarises the STM and introduces the data. Section 3 presents the simulation plan and selected diagnostics. Section 4 shows the results of a preliminary pilot study carried out on reduced dataset and simulation scope.

2 Model and data

STM is an unsupervised method for identifying the topical structure of a collection of texts. The method incorporates observable metadata information (i.e. covariates at the text level) to capture their effects on topics. STM can be conceptually divided into three components. The first is a topic prevalence model, which controls how words are allocated to topics as a function of covariates:

$$\gamma_\kappa \sim Normal_p(0, \sigma_k^2 I_p), \quad for\ k = 1 \dots K-1, \tag{1}$$

$$\theta_d \sim LogisticNormal_{K-1}(\Gamma^{'} x_d^{'}, \Sigma), \tag{2}$$

where Γ is the matrix of coefficients for the topic prevalence model specified by Equations (1) and (2), d stands for document, k is the number of topics, X is the matrix for topic prevalence, and σ is a k-dimensional hyper-parameter vector; The second is a topical content model that controls the frequency of the terms in each topic as a function of covariates:

$$\beta_{d,k,v} = \frac{exp(m_v + k_{k,v}^{(t)} + k_{y_d,v}^{(c)} + k_{y_d,k,v}^{(i)})}{\sum_v exp(m_v + k_{k,v}^{(t)} + k_{y_d,v}^{(c)} + k_{y_d,k,v}^{(i)})} \quad for\ v = 1 \dots V\ and\ k = 1 \dots K, \tag{3}$$

where $k_{k,v}^{(t)} + k_{y_d,v}^{(c)} + k_{y_d,k,v}^{(i)}$ is a collection of coefficients for the topical content model, and m_v is the marginal log-transformed rate of term v; The third is a core observation model that combines models (1), (2), and (3)

$$\mathbf{z}_{d,n} \sim Multinomial_k(\theta_d), \quad for\ n = 1 \dots N_d, \tag{4}$$

$$\mathbf{w}_{d,n} \sim Multinomial_v(\mathbf{B}\ \mathbf{z}_{d,n}), \quad for\ n = 1 \dots N_d, \tag{5}$$

The core observation model allows for correlations in the topic proportions using the logistic normal distribution. The topic prevalence (which describes the association of a document with a topic) and the topical content (which describes how the words are used within a topic) components enable the expected text-topic proportion and, respectively, the topic-word probability to vary as a function of the observed text-level covariates X rather than arising from global parameters shared by all texts.

The diachronic corpus under scrutiny is a collection of all end-of-year addresses of the Italian Presidents of the Republic. Time (i.e. years of the speeches) and President (i.e. speakers) are the covariates employed to test their effect on the STM components. The corpus includes 73 addresses (1949-2021) delivered by 11 presidents. The corpus is available in both original and lemmatised versions and is continuously updated through the collation of digitalised text with audio-visual recordings. Word selection relies on part of speech (POS) information and frequency.

3 Simulation plan and selected diagnostics

We consider and discuss two main problems: 1) determining the length of the texts under scrutiny and the opportunity to work with equal-sized chunks; and 2) choosing the best model (which involves topic number selection) by comparing alternative specifications (i.e. metadata). Concerning the first, although speech length varies considerably, the presidential addresses generally represent medium-length texts (i.e. longer than a social media post, shorter than a novel). This raises the following questions: would length standardisation improve model performance? Is there a text length more appropriate to topic modelling? Given these questions, we compare per-length diagnostic score distributions over the topic number for the original setting (consisting of the whole documents) and for standardised settings (obtained by chunking the original speeches into equal text fragments). Given that we chose to first select content words (by POS tagging and frequency threshold) and then split the documents, we decided to make chunks constituted of 10 to 100 words that increment by a 10-word step. The selected content words cover roughly 50 percent of a sentence; a chunk of 10 words corresponds to a short-medium 20-word sentence, and a chunk of 100 corresponds to a 200-word text that is close to the shortest original documents. Lastly, related to the first aim, we discuss the assumption that each text is necessarily multi-topic, also in relation with the length of the text itself. As for the second objective, once we have applied the best chunking option (including no chunking), we will compare diagnostic score distributions over the topic number to select the best model from the different combinations of both covariates and model specifications (i.e. year and/or President included as prevalence and/or content model).

We chose the most widely used intrinsic measures for TM evaluation to compose the set of diagnostics under investigation. Within the class of diagnostics for addressing a predictive task, we calculate the following: (p1) held-out log-likelihood, (p2) residuals, (p3) perplexity and (p4) model posterior probability. The first two are provided by the `stm` R package and correspond to the (p1) log-likelihood of the held-out set of words, according to the document completion method [9] and the (p2) multinomial dispersion of model residuals (i.e. when the model is correctly specified, the multinomial likelihood implies a residual dispersion $\sigma^2 = 1$). Perplexity (p3) is the most well-known diagnostic in topic modelling and is defined as the inverse geometric mean of the per-word probability of a held-out set of words. (p4) is the model posterior probability under a Bayesian estimation approach [4]. The class of topic quality diagnostics includes countless examples. However, a 'red thread' that allows a synthetic interpretation can be borrowed from the literature of psychology (i.e. self-definition) and organisation theory ('category' definition), which explains that a well-defined topic requires the co-presence of distinctiveness, coherence and continuity. A non-exhaustive list of quality measures contains the following: (q1) semantic coherence, (q2) exclusiveness, (q3) consistency and differentiation (CD) scores, (q4) between-topic (cosine) similarity [6], (q5) symmetric Kullbach-Leibler divergence between the singular value distribution of the topic-word matrix and the row L_1 norm of document-topic matrix [1], (q6) the Jensen-

Shannon divergence between all pairs of topics [3] and (q7) distinctiveness whence saliency [2]. (q1) and (q2) are provided by the `stm` package, though only for those specifications without content covariates (to overcome the limitations of existing packages, we developed ad hoc R procedures for each measure). We propose using (q3) to synthetise the trade-off between (q1) and (q2) by calculating the L_2 norm of the two min-max normalised measures to pinpoint the top-right region of the plane generated by the two metrics.

In this work, we first focus on the p1–p2 and q1–q3 measures. The idea is to extend the empirical part to the other diagnostics presented above. We repeat the estimation of each configuration (length, topic number, model) 30 times to control the variability associated with model initialisation (i.e. different random seed values). The topic numbers range from three to 50.

4 A pilot study

For our pilot study, we chose to work with sole nouns that occur at least 10 times (612 lemmas). Figure 1 shows the first two predictive (p1–p2; top left and middle panels) and topic quality diagnostics (q1–q2 on the same plane; synthesis q3, the CD score; in bottom left and middle panels) required for text length selection. Both the p1–p2 and q3 measures clearly favoured longer chunks, so we chose a chunk size of 50 nouns. This represented the longest chunk possible and appeared to exalt interpretability while maintaining both the highest levels of held-out log-likelihood and the lowest levels of residuals. The CD score provided two more insights: (1) equal-sized chunks ensure better results than whole documents, and (2) each CD curve reaches its optimum level around k within 10 to 15 topics. This indicates that chunking does indeed produce a form of standardisation. A topic number that is slightly higher than the number of presidents may suggest a combination of factors, such as the distinctive influence of the presidents' personal traits and the evolution of socio-political historical facts. Once the ideal chunk length was fixed, the next step was to determine the optimal model specification. We experimented with different settings to test the effect of the Year (smoothed with splines) and/or President covariates on the topic prevalence (θ) and/or the topical content (β) components. These settings were as follows: (i) no covariates for neither θ nor β; (ii) Year for θ and no covariate for β; (iii) President for θ and no covariate for β; (iv) Year for θ and President for β; (v) no covariate for θ and President for β; (vi) President on θ and Year on β; (vii) no covariate for θ and Year for β. Models that included the President factor in the topical content were found to perform far better (Figure 1, top-right). Given a topic, a different lexicon characterised the Presidents. The ideal topic number was determined at slightly above 10 or 20 across all models. By picking the best or second-best topic number, most models (except for the best couple with President in β) performed similarly on the predictive side (bottom-right).

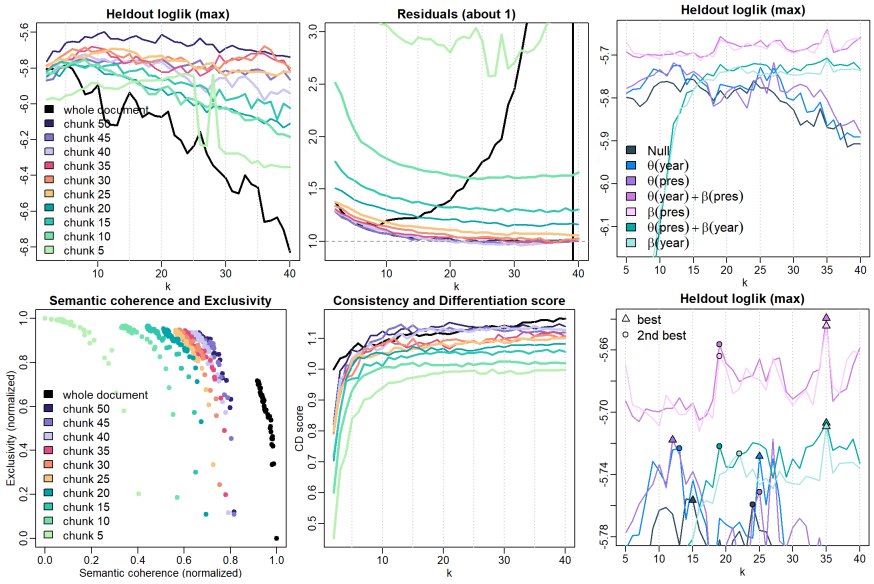

Fig. 1 (Left and middle panels) Per-length predictive (p1-p2; top) and topic quality (q1-q3; bottom) diagnostic distributions for choosing the best text length; (right panels) predictive distributions for choosing the best model structure (p1, top; p2, bottom)

References

1. Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N.: On finding the natural number of topics with latent dirichlet allocation: Some observations. In Zaki M. J., Yu J. X., Ravindran B., Pudi V. (eds) Advances in knowledge discovery and data mining, 391-402 (2010)
2. Chuang, J., Manning, C., Heer, J.: Termite: visualization techniques for assessing textual topic models. Proceedings of International Working Conference on Advanced Visual Interfaces (AVI 2012), 74-77 (2012)
3. Deveaud, R., SanJuan, E., Bellot, P.: Accurate and effective latent concept modeling for ad hoc information retrieval. Document numerique 17(1), 61-84 (2014)
4. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235 (2004)
5. Guo, C., Lu, M., Wei, W.: An improved LDA topic modeling method based on partition for medium and long texts. Annals of Data Science 8(2), 331-344 (2021)
6. Juan, C., Tian, X., Jintao, L., Yongdong, Z., Sheng, T.: A density-based method for adaptive lda model selection. Neurocomputing - 16th European Symposium on Artificial Neural Networks 2008 72(7-9), 1775-1781 (2009)
7. Roberts, M.E., Steward, B.M., Airoldi, E.M.: A Model of Text for Experimentation in the Social Sciences. J of the American Statistical Association 111(515), 988-1003 (2016)
8. Sbalchiero, S., Eder, M.: Topic modeling, long texts and the best number of topics. Some problems and solutions. Quality and Quantity 54(4), 1095-1108 (2020)
9. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimmo, D.: Evaluation methods for topic models. Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09), pp 1105-1112 (2009)