Zuzana JANKOVÁ, PhD
E-mail: zuzana.jankova@vutbr.cz
Faculty of Business and Management
Brno University of Technology

LATENT DIRICHLET ALLOCATION (LDA) APPROXIMATION ANALYSIS OF FINANCIAL-RELATED TEXT MESSAGES

Abstract. *Topic modeling is one of the most widely used NLP techniques for discovering hidden patterns and latent relationships in text documents. Latent Dirichlet Allocation (LDA) is one of the most popular methods in this field. There are different approaches to tuning the parameters of LDA models such as Gibbs sampling, variational method, or expected propagation. This paper aims to compare individual LDA parameter tuning approaches with respect to their performance and accuracy on textual data from the financial domain. From our results, it can be concluded that for text datasets obtained from financial social platforms, stochastic solvers are the most suitable and especially less time consuming than approximation methods.*

Keywords*: Financial Messages, LDA, Latent Dirichlet Allocation, NLP, Text analysis, Topic Modeling.*

JEL Classification: C39, C55, G17

1. Introduction

During the last decade, a huge amount of economic or financial textual data has been published, which is available online through e-mails, discussion forums, social networks, or news. Text documents are considered a pivotal source of information delivery through multimedia tools and applications (Shao et al., 2022). The appropriate use of this textual data can provide valuable information and important insight into the research issue (Chen et al., 2017). Finding meaningful insights in unstructured text data is difficult if the researcher cannot search, process, and classify it correctly and efficiently. Properly processed unstructured data in the form of text can be combined and supplemented with quantitative data to provide more accurate insights. In the financial literature, the vast majority of research in the field of text analysis is devoted to the prediction of market movements, most often to the prediction of stock prices.

One of the common techniques for analysing text documents is to find related topics. However, the modeling of topics in the financial field remains a relatively unexplored area of textual analysis. Topic modeling ranks among unsupervised text document classification approaches. It is the process of identifying the themes present and inferring hidden patterns in large volumes of text documents. Latent Dirichlet Allocation (LDA for short) is one of the most widespread and widely used topic modeling approaches (Aziz et al., 2022). The

DOI: 10.24818/18423264/57.1.23.17

key task of topic modeling is for the researcher to use feedback from observed words to uncover hidden topics. According to Jelodar et al. (2018) is a highly prominent technique for demonstrating discrete data in particular. There are many areas of research research in this part of natural language processing, including literature (Chen et al., 2020; Chen et al., 2021), software engineering (Agrawal et al., 2018), and political science (Jelodar et al., 2018).

The LDA algorithm iteratively partitions the topics for each document, as well as the words of each topic, until it generates the best set of topic and word distributions. LDA is capable of processing a large number of documents that would have caused unnecessarily high costs in manual coding. The algorithm thereby provides a reliable classification of topics, while not requiring prespecification of rules or keywords (Huang et al., 2017; Goloshchapova et al., 2019).

This paper focuses on modeling topics in the financial domain. The challenge of research in the field of finance still remains efficient, fast and accurate work in the processing of natural language, especially in a period characterised by a dynamic increase in text documents. Therefore, the motivation of the paper is to investigate parameter estimation, inference, and training approaches for modeling financial topics through LDA. The main objective of the presented paper is to provide a comparative analysis of latent Dirichlet allocation in order to compare multiple approaches to the estimation of LDA model parameters and evaluate their effectiveness for the analysis of a large set of financial reports. This goal is approached because of previous research (Sun et al., 2015; Agrawal et al., 2018), who commented that topic modeling results can be significantly affected by tuning LDA control parameters. Even a small deviation can play a major role in financial decision-making. For that reason, the correct tuning of the LDA parameters is considered crucial for the correct and especially timely identification of core topics discussed on online social networks. Correctly identified topics can fundamentally influence subsequent sentiment analysis and determine the influence and direction of financial market development. The main contributions of the presented research can be summarised as follows: (i) a review of scholarly articles on topic modeling in the field of finance to discover development trends in the field and provide an intellectual basis for LDA-based topic modeling; (ii) a summary of the current challenges of thematic modeling of textual documents in the research area; (iii) application of different LDA model parameter estimation approaches to a financial data set; (iv) comparison of these selected LDA approaches in terms of performance and effectiveness.

The paper is structured as follows. Section 2 provides a theoretical background synthesising the outputs of modeling topics in finance. Section 3 describes LDA methods including model parameter tuning approaches. Section 4 provides the outputs of the practical application of the above methods and their comparison in terms of performance. Section 5 then discusses the outputs and describes the conclusions drawn.

2. Literature review

Most of the research is mainly focused on the role of sentiment analysis in predicting stock returns and stock market volatility. However, topics discussed in financial reports may contain no less important information (Hájek and Baruska, 2018). Topic modeling makes it possible to find vocabulary associated with the topic of a text corpus and to capture the semantic relationship across documents. The paper is focused on the LDA algorithm, which according to Blei et al. (2003) is one of the most popular topic models. Recently, several researches and studies dealing with topic modeling in text processing have been proposed in many economically oriented fields, such as customer reviews (Korfiatis et al., 2019), banking sector (Chen et al., 2017), consumer complaints of financial services (Bastani et al., 2019). Research focused on the financial market is particularly relevant for this paper, as the subject of interest is financial reports. Recent studies related to the subject matter are discussed below.

Ploessel et al. (2021) analysed real estate topics and then used a dictionary-based approach to determine the tone of coverage of a specific topic and analysed the impact on the real estate market. Feuerriegel and Pröllochs (2021) examined the influence of different themes generated through LDA from regulatory 8-K filings from the US companies on share prices, or stock market reaction. Edison and Carcel (2021) applied LDA to reveal the evolution of various topics discussed by members of the U.S. Federal Open Market Committee (FOMC). Goloshchapova et al. (2019) modeled themes from Corporate Social Responsibility reports for MSCI Europe stock index companies. On the basis of topic modeling, they identified the sectoral bias of companies. The most recent study by Balaneji and Maringer (2022) looked at the combination of themes and sentiment scores extracted from financial news and its effect on predicting the evolution of the implied volatility index. Their results show that adding a topic model has a positive effect on model accuracy. Larsen and Thorsrud (2022) investigated the causal and predictive role of news topics and asset prices. Their research findings show that news published online has a significant and potentially profitable predictive power of asset returns. According to the authors, there is a significant drop in the prices of the assets of companies that are very often mentioned and discussed in the news. According to the results of Chen et al. (2017) thematic modeling is an effective tool for financial decision-making. Morimoto and Kawasaki (2017) add that incorporating topics using a dynamic model can contribute to improving the prediction of volatility in financial markets. Similar results can be found in the work of Kanungsukkasem and Leelanupab (2019).

It follows from the conducted literature research that the area of this interest is not sufficiently covered. Moreover, the number of publications dealing with modeling topics in financial markets is insufficient. To the best knowledge of the authors of this paper, methods for estimating topic modeling parameters via

LDA have not been explored in this field. For that reason, it is necessary to pay more attention to the modeling of topics in the financial markets, because according to Chen et al. (2018) one of the most challenging problems in topic-based sentiment analysis is how to get relevant topics, especially in the financial market. Identified topics can be used for semantic computing and behavioural financial research.

3. Topic Modelling via Latent Dirichlet Allocation

To model topics, probabilistic topic models such as Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) were developed by Blei et al. (2003). The principle of these models is based on the assumption that they model a document as a mixture of hidden (latent) topics. Individual topics are defined by the probabilistic distribution of words in the vocabulary.

3.1. Latent Dirichlet Allocation

LDA is a generative statistical model to explain why some data are similar. First, the number of searched topics is defined in advance, and then the Feuerriegel and Pröllochs (2021) topic is selected based on the highest probability with respect to a specific set of words. Two probability distributions are examined to find LDA model topics:

- $\alpha = P(t|d)$, indicating the probability of topic t in the document d;
- $\beta = P(w|t)$, indicating the probability of the word w in the topic t.

Blei et al. (2003) and Blei (2012) describe this two-phase process in more detail as follows:

1. For each text document d in corpus C, topics θ_d are randomly distributed, where $\theta_{d,k}$indicates the proportion of the number of topics k in document d. The random variable θ_d follows the Dirichlet probability distribution with α given by the relation:

$$\theta_d \sim Dir(\alpha), \qquad \alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)$$

2. Words are assigned to topics in such a way that word n is selected from document d for the selected topic $t_{d,n}$ from θ_d. Subsequently, words are $w_{d,n}$ selected across a fixed dictionary, which is conditional on the selected topic $t_{d,n}$. The distribution is given by the relation $\beta_k \sim Dir(\eta)$ with prior η.

The uniform probability is given by:

$$P(\theta, \beta, w, t) = \prod_{d=1}^{D} P(\theta_d|\alpha) \prod_{k=1}^{K} P(\beta_k|\eta) \prod_{n=1}^{N} P\left(t_{d,n}|\theta_d\right) P\left(w_{d,n}|t_{d,n}\right)$$

However, it is not possible to directly maximise the likelihood, since topics cannot be observed, only documents.

$$P(\theta, \beta, t|w, \alpha, \eta) = \frac{P(\theta, \beta, w, t|w, \alpha, \eta)}{P(w|\alpha, \eta)}$$

The problem is with the denominator of the fraction, because there is a dependency between β and θ. Thus, LDA uses approximate inference techniques. It is necessary to choose the Dirichlet priors α and η, as they determine the distribution between the document and the topic; and topic and word.

In a subsequent step, LDA assigns a topic name identifier to each extracted topic. According to Feuerriegel and Pröllochs (2021), a list of the three to thirty most likely words in the given topic is examined for correct identification.

3.2. LDA parameter estimation, inference and training

There are multiple methods for estimating topic modeling parameters via LDA. Among the most widespread are Gibbs sampling (Griffiths and Steyvers, 2004), variational method (Blei and Jordan, 2003), or expected propagation (Minka and Lafferty, 2002).

- The Gibbs sampling is based on the Monte Carlo Markov-chain algorithm. It is a powerful technique based on statistical inference in which the conditional distribution of each variable can be efficiently calculated.
- The expected propagation is used to estimate the parameters of graphical models. It is based on finding the maximum likelihood of estimating model parameters that are dependent on some variables.
- The variational method is a extension of the aforementioned expected propagation based on a parametric approximation to the posterior distribution and optimises the fit (Jelodar et al., 2018).

4. Data and Methodology

In the following section, the text data source as well as the preprocessing method and processing process of topic modeling through LDA are introduced. The software MATLAB 2021a, Text Analytics Toolbox, and Parallel Computing Toolbox are used for the calculation.

4.1. Corpus extraction

For the purpose of this paper, the social platform StockTwits was chosen. Stocktwits is the largest social network for investors and traders, with over five million community members and millions of monthly visitors. As the definitive voice of "social finance", Stocktwits is the best way to find out what is happening in the markets right now. Due to the real-time nature of StockTwits, it offers an ideal source of public data. An extensive dataset is freely available from Jaggi et al. (2021). The StockTwits dataset has nearly 6.4 million records spread over ten years.

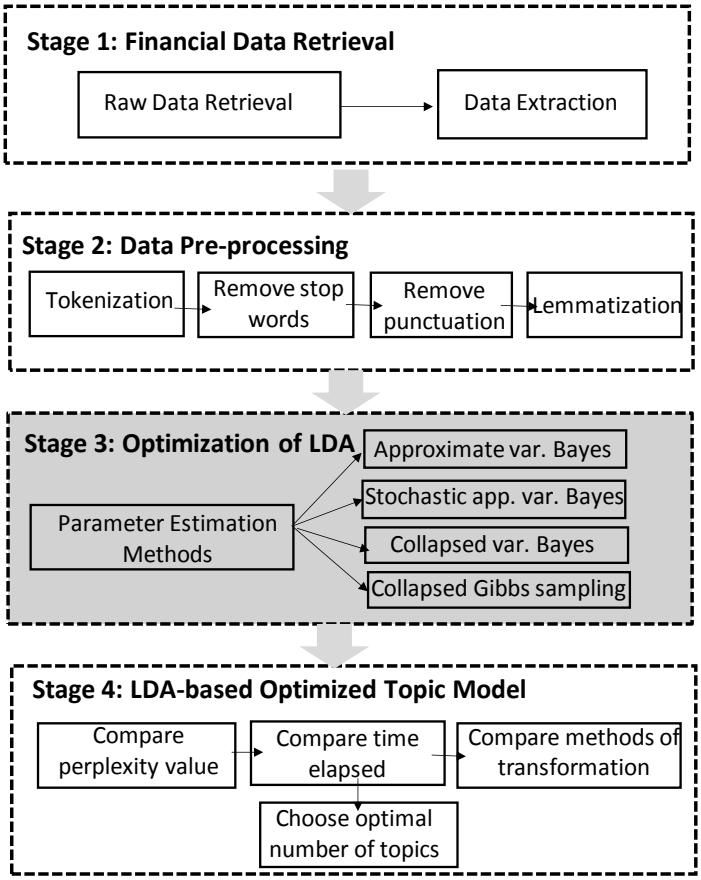
4.2. Corpus pre-processing

The first step in topic modeling is the pre-processing of the text corpus. The data preprocessing stage is an important stage in the text data mining and processing process. This step aims to remove irrelevant terms that occur in the source documents. Text documents are first (1) tokenised into a collection of individual tokens. Subsequently, (2) stop words that provide no significant context are removed. To reduce the dimensionality of the documents, words that are too long with 15 or more characters or words that are too short with 2 characters or less are also (3) removed. Furthermore, (4) normalisation, more precisely, lemmatisation, is used. This means that the suffixes and prefixes of the words are removed for the obtained root words. A summary of these preprocessing tasks is shown in Fig.1.

Preprocessing plays a vital role in the natural language processing process, as it reduces noise in the data, which can significantly affect the performance of LD. In addition to noise, there are also several redundant and meaningless words in text documents that unnecessarily increase the dimensionality of the text and are removed. Thus, the classification accuracy can be improved by the appropriate selection of preprocessing techniques.

The pre-processed textual content is then transformed into numerical feature vectors using the TF-IDF approach, which gives more weight to feature expressions. For a more fine-grained analysis, we extracted BoW and N-gram features from the corpus. Specifically, each occurring sequence of words of length "1", "2", and "3" is extracted from the tweets to create a dictionary of words and phrases. Some simple statistical analysis shows that the corpus shows a number of terms that are too frequent and indicate certain aspects related to market activities such as "buy", or the amplification of the call to a buying investment strategy with the trigrams "buy, buy, buy". While BoW indicates individual activities on the market without an obvious and unambiguous connection, N-grams give context to, for example, a specific company "tesla, short", "aapl, new, article" or an important personality "elon musk". The same analysis can be performed for other companies, i.e., the entire stock market.

Figure 1. The process of setting up the LDA model



5. Empirical Findings

The Latent Dirichlet Allocation (LDA) semantic model is used in the classification of text data. The latter falls under the category of machine learning without a teacher. This means that it only needs the data, the number of topics parameter, and the model identifies the appropriate number of topics in the data by itself. It is a straightforward, albeit relatively computationally demanding task. To test the function of the LDA algorithm, documents divided into two corpora intended for training the model and for comparing the trained model are used. Each corpus consists of already pre-processed words.

5.1. Selection of the number of topics

The following part of the paper shows how to decide on an appropriate number of topics for the LDA model. To choose an appropriate number of topics, it

is convenient to compare different LDA models with different numbers of topics. The suitability of the LDA model can be evaluated using the perplexity of text documents. Perplexity expresses the suitability of the model with respect to the selected text document. The lower the value, the better the model fit. The goal is to select several topics that minimise the perplexity of the model. To evaluate the trade-off between these two effects, both accuracy and computation time are calculated. Approximation methods are used in LDA to select the correct number of topics. A comparison of the performance of these methods for textual data using the BoW textual representation is shown in Figure 2 indicating the degree of perplexity, resp. 3 showing the elapsed time.

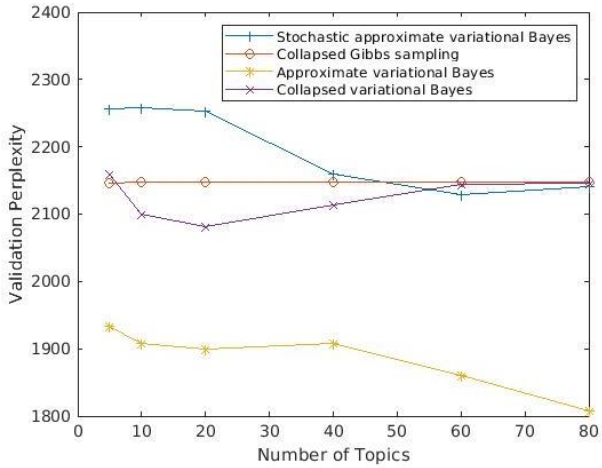


Figure 2. Comparison of approximation methods according to perplexity

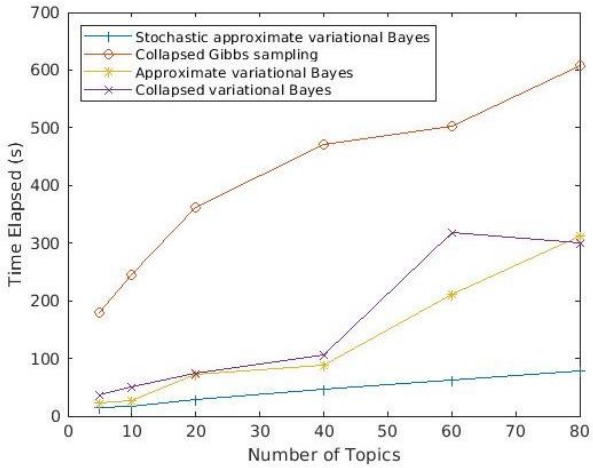


Figure 3. Comparison of approximation methods by time

The most commonly used methods that are compared include variational methods and Gibbs sampling. For large data sets, the approximation via stochastic variational Bayes is most appropriate. As can be seen from the pictures, it is less accurate in terms of the perplexity indicator, but the most time efficient. Stacked Gibbs sampling shows a slight improvement in accuracy, however, the highest time consumption. The approximation variational Bayes achieved the most accurate results, but with the increasing number of topics, the calculation time increases sharply. Packed variational Bayes is more accurate than approximate Bayes, but at the cost of taking longer to run.

As mentioned above, the stochastic solver is most suitable for large data sets and especially the least time-consuming than approximation methods. For that reason, the evaluation of different textual representations is carried out by means of stochastic variational Bayes. Perplexity normalisation was performed for visualisation. According to Figures 4 and 5, Trigram shows the lowest error rate for 0-20 topics, but the calculation time increases fundamentally with the increasing number of topics. Bigram, which is less accurate and more time-consuming than Trigram, also shows similar results. In the case where more than 20 topics are determined in the corpus, overfitting occurs, and the model begins to deteriorate. As for BoW, from 40 or more topics, it shows the lowest perplexity and, moreover, a fundamental change in computational complexity is noticeable. It can be concluded that setting the model with 20 topics can be a good choice for N-grams, while for the BoW text representation, setting the number of topics in a document to 40-60 seems appropriate. However, human judgment is still required as to exactly how many topics to use.

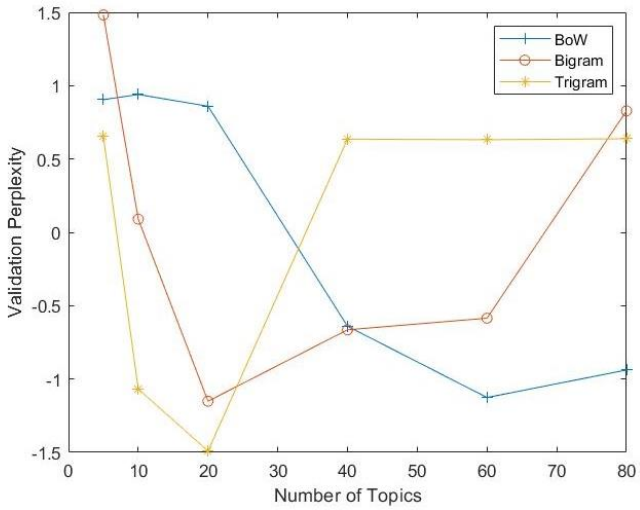


Figure 4. Comparison of BoW, N-gram by perplexity

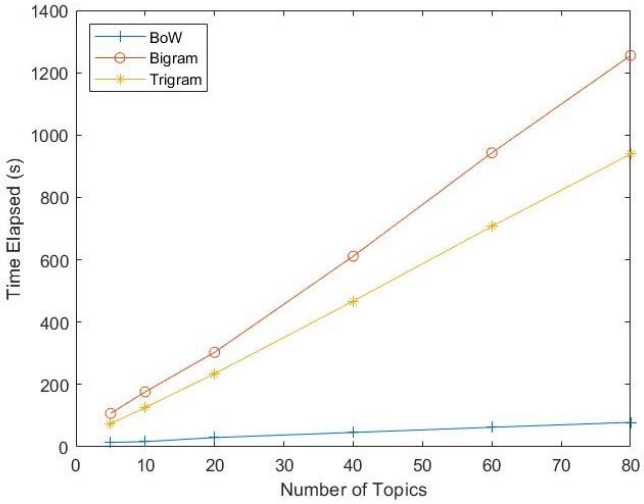


Figure 5. Comparison of BoW and N-gram by time

However, it should be noted that the entire calculation process took several hours. Especially when calculating the other methods, with the exception of stochastic approximation variational Bayes, there was a noticeable increase in the time required for the calculations. The computational complexity was further increased with the use of Bigrams and Trigrams. In addition, the Parallel Computing Toolbox and MATLAB Parallel Server were also used for the calculation, allowing to solve demanding mathematical tasks in the environment of multiprocessor stations, GPUs, and computer clusters, but without significantly speeding up the calculation. Exactly how the number of documents, topics, and words contributes to the time cost and working memory requirements of fitting a topic model should be the subject of future research.

5.2. Fitting the LDA model

The LDA model is a topic model that discovers underlying topics in a collection of documents and infers the probabilities of words in the topics. According to the previous section, a stochastic approximation variational Bayesian solver is chosen for the extension of the LDA model, because, as was illustrated, it is usually more suitable especially in terms of computational complexity. For example, considering Figure 4, a break in the curve is considered a signal for an appropriate number of topics. Here, the break corresponds optimally to the choice of 20 topics. Due to semantic association, the most likely words for each topic are derived. Such a semantic relationship, which is defined as a rough association of words, does not necessarily correspond to linguistic theories, but must also be comprehensible to human judgment. However, it is difficult for humans to

distinguish between optimal and nonoptimal topics. A sample of topics according to the textual representation of BoW is given in Table 1, while the number of words generated for each topic is set to 5. The distribution of topics according to the textual representation of bigrams and trigrams is given in the Appendix.

Table 1. Word distribution for each topic via Bow

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
against	new	source	china	coronavirus	billion	share	job	back	finance
0.025	0.039	0.053	0.119	0.026	0.023	0.035	0.038	0.016	0.034
court	amazon	exclusive	trade	job	apple	sale	plan	johnson	ceo
0.025	0.018	0.042	0.118	0.019	0.020	0.026	0.036	0.015	0.033
fund	uber	billion	trump	british	sell	tesla	cut	claim	chief
0.024	0.017	0.033	0.082	0.017	0.019	0.025	0.031	0.013	0.024
billion	tax	million	tariff	firm	firm	growth	bank	call	nil
0.023	0.015	0.031	0.054	0.015	0.016	0.024	0.021	0.011	0.021
saudi	launch	pay	deal	industry	well	see	store	bill	pratley
0.022	0.014	0.021	0.050	0.014	0.016	0.024	0.017	0.010	0.021
Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
facebook	rate	business	oil	car	profit	cramers	street	boeing	cramer
0.026	0.040	0.087	0.082	0.033	0.050	0.037	0.061	0.053	0.087
datum	brexit	brexit	price	nissan	share	round	wall	airline	stock
0.021	0.024	0.047	0.049	0.019	0.038	0.030	0.053	0.044	0.073
project	economy	live	rise	new	sale	bank	high	max	market
0.019	0.023	0.042	0.034	0.019	0.036	0.028	0.044	0.037	0.056
business	fed	coronavirus	cut	electric	hit	lightning	record	737	jim
0.019	0.022	0.034	0.030	0.018	0.018	0.027	0.021	0.036	0.030
regulator	bank	happen	fall	ceo	forecast	buy	stock	airbus	investor
0.014	0.021	0.032	0.021	0.017	0.018	0.021	0.021	0.026	0.026

From a human judgment point of view, it can be argued that there is a certain redundancy of words between topics and, as a result, a lack of homogeneity in the content of topics. Intuitively, selecting a combination with too many topics could generate topics that are too specific and redundant, while selecting fewer topics based on larger corpus blocks will make the topics too broad and diverse.

4. Conclusions

The paper dealt with the application and comparison of different approaches to tuning the parameters of Latent Dirichlet Allocation (LDA) models on a financial text set. Several LDA models applying Gibbs sampling, variational method, or expected propagation were created, and their performance was monitored. The default LDA shows significant instability in the number of generated topics, so it is necessary to focus on the correct tuning of the LDA

model. From our experimental results, it follows that for large data sets, in terms of accuracy and time-consuming, the most appropriate approximation is by means of stochastic variational Bayes. The packed Gibbs sampling shows the highest time consumption. The packed variational Bayes is indeed more accurate than the approximation variant of Bayes, but at the price that its computational complexity is higher. From the above, it can be concluded that for text datasets obtained from financial social platforms, stochastic solvers are the most suitable and especially less time-consuming than approximation methods.

Bag-of-Words, Bigram and Trigram were investigated in terms of methods of transforming text into data structures. The trigram shows the lowest error rate, but the computation time increases substantially as the number of topics increases. Similar results are shown by Bigram, which is less accurate and even more time-consuming than Trigram. BoW shows the lowest perplexity, and, moreover, a fundamental change in computational complexity is evident. Thus, from the point of view of transformation methods, taking performance into account appears to be a suitable BoW method for the financial domain.

Based on the themes generated, it can be argued that there is some redundancy between themes due to a lack of homogeneity in their content. Combinations with a large number of topics generate topics that are too specific and redundant. Selecting a lower number of topics will cause too broad and diverse topics to be generated.

REFERENCES

[1] **Agrawal, A., Fu, W., Menzies, T. (2018),** ***What is wrong with topic modeling? And how to fix it using search-based software engineering.*** *Information and software technology*, 98, 74-88, https://doi.org/10.1016/j.infsof.2018.02.005;

[2] **Aziz, S., Dowling, M., Hammami, H., Piepenbrink, A. (2022),** ***Machine learning in finance: A topic modeling approach.*** *European financial management: the journal of the European Financial Management Association*, 28(3), 744-770, https://doi.org/10.1111/eufm.12326;

[3] **Balaneji, F., Maringer, D. (2022),** ***Applying Sentiment Analysis, Topic Modeling, and XGBoost to Classify Implied Volatility.*** In: *2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr), IEEE, 2022*, 1-8, https://doi.org/10.1109/CIFEr52523.2022.9776196;

[4] **Blei, D.M. (2012),** ***Probabilistic topic models*****.** *Communications of the ACM*, 55(4), 77–84;

[5] **Blei, D.M., Ng, A.Y., Jordan, M.I. (2003),** ***Latent Dirichlet allocation.*** *Journal of Machine Learning Research*, 3, 993–1022;

[6] **Blei, D.M., Jordan, M.I. (2003),** ***Modeling annotated data.*** In: *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval*. ACM;

[7] **Chen, X., Xie, H., Cheng, G., Li, Z. (2021),** ***A Decade of Sentic Computing: Topic Modeling and Bibliometric Analysis.*** *Cognitive computation*, 14(1), 24-47. https://doi.org/10.1007/s12559-021-09861-6;

[8] **Chen, X., Xie, H., Cheng, G., Poon, L.K.M., Leng, M., Wang, F.L. (2020),** ***Trends and deatures of the applications of natural language processing techniques for clinical trials text analysis***. *Appl Sci.* 10(6), 2157–93;

[9] **Chen, W., Lai, K., Cai, Y. (2018),** ***Topic generation for Chinese stocks: a cognitively motivated topic modelling method using social media data.*** *Quantitative Finance and Economics*, 2(2), 279-293, https://doi.org/10.3934/QFE.2018.2.279;

[10] **Chen, Y., Rabbani, R.M., Gupta, A., Zaki, M.J. (2017),** ***Comparative text analytics via topic modeling in banking.*** In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1-8, https://doi.org/10.1109/SSCI.2017.8280945;

[11] **Edison, H., Carcel, H. (2021),** ***Text data analysis using Latent Dirichlet Allocation: an application to FOMC transcripts.*** *Applied Economics Letters*, 28(1), 38-42, https://doi.org/10.1080/13504851.2020.1730748;

[12] **Feuerriegel, S., Pröllochs, N. (2021),** ***Investor Reaction to Financial Disclosures across Topics: An Application of Latent Dirichlet Allocation***. *Decision sciences*, 52(3), 608-628, https://doi.org/10.1111/deci.12346;

[13] **Griffiths, T.L, Steyvers, M. (2004),** ***Finding scientific topics.*** *Proc Natl Acad Sci*, 101(1), 5228–5235;

[14] **Goloshchapova, I., Poon, S.H., Pritchard, M., Reed, P. (2019),** ***Corporate social responsibility reports: topic analysis and big data approach.*** *The European journal of finance*, 25(17), 1637-1654, https://doi.org/10.1080/1351847X.2019.1572637;

[15] **Huang, A., Lehavy, R., Zang, A., Zheng, R. (2017),** ***Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach.*** *Management Science*;

[16] **Hajek, P., Barushka, A. (2018),** ***Integrating Sentiment Analysis and Topic Detection in Financial News for Stock Movement Prediction.*** In: *Proceedings of the 2nd International Conference on Business and Information Management* [online]. New York, NY, USA: ACM, 2018-09-20, 158-162, https://doi.org/10.1145/3278252.3278267;

[17] **Jaggi, M., Mandal, P., Narang, S., Naseem, U., Khushi, M. (2021),** ***Text mining of stocktwits data for predicting stock prices***. *Applied System Innovation*, Basel: MDPI, 4(1);

[18] **Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L. (2018),** ***Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey.*** *Multimedia tools and applications*, 78(11), 15169-15211, https://doi.org/10.1007/s11042-018-6894-4;

[19] **Kanungsukkasem, N., Leelanupab, T. (2019),** ***Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction.*** *IEEE Access*, 7, 71645-71664, https://doi.org/10.1109/ACCESS.2019.2919993;

[20] **Korfiatis, N., Stamolampros, P., Kourouthanassis, P., Sagiadinos, V. (2019),** ***Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews.*** *Expert systems with applications*, 116, 472-486, https://doi.org/10.1016/j.eswa.2018.09.037;

[21] **Larsen, V.H., Thorsrud, L.A. (2022),** ***Asset returns, news topics, and media effects.*** *The Scandinavian Journal of Economics*, 124(3), 838-868, https://doi.org/10.1111/sjoe.12469;

[22] **Minka, T., Lafferty, J. (2002),** ***Expectation-propagation for the generative aspect model.*** In: *Proceedings of the eighteenth conference on uncertainty in artificial intelligence.* Morgan Kaufmann Publishers Inc;

[23] **Morimoto, T., Kawasaki, Y. (2017),** ***Forecasting Financial Market Volatility Using a Dynamic Topic Model.*** *Asia-Pacific Financial Markets*, 24(3), 149-167, https://doi.org/10.1007/s10690-017-9228-z;

[24] **Ploessl, F., Just, T., Wehrheim, L. (2021),** ***Cyclicity of real estate-related trends: topic modelling and sentiment analysis on German real estate news***. *Journal of European Real Estate Research*, 14(3), 381-400, https://doi.org/10.1108/JERER-12-2020-00593;

[25] **Shao, D., Li, C., Huang, C., Xiang, Y., Yu, Z. (2022),** ***A news classification applied with new text representation based on the improved LDA.*** *Multimedia tools and applications*, 81(15), 21521-21545, https://doi.org/10.1007/s11042-022-12713-6;

[26] **Sun, X., Li, B., Leung, H., Li, B., Li, Y. (2015),** ***MSR4SM: Using topic models to effectively mining software repositories for software maintenance tasks.*** *Information and software technology*, 66, 1-12, https://doi.org/10.1016/j.infsof.2015.05.003

Appendix 1. Word distribution for each topic via Bigram

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
cramer remix	small business	deutsche bank	cramer remix	cramers lightning	hong kong
box office	saudi aramco	week ahead	selfdriving car	lightning round	elon musk
land rover	white house	take five	pay million	climate change	retail sale
coronavirus crisis	billion deal	trump tariff	china zte	new york	second quarter
jaguar land	well fargo	financial crisis	million euro	british airway	share plunge

Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
737 max	thomas cook	stock market	business live	due coronavirus	trade deal
boeing 737	new york	jim cramer	first time	share fall	china trade
wall street	job risk	first quarter	trade war	aston martin	trade talk
economics viewpoint	guardian view	high street	wall street	fossil fuel	trade war
white house	brief letter	five year	step down	house fraser	hong kong

Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18
fiat chrysler	business live	rate cut	project syndicate	nil pratley	amid coronavirus
next year	nodeal brexit	instant view	syndicate economist	pratley finance	trade war
business leader	house price	federal reserve	wall street	business live	fiat chrysler
central bank	oil price	share fall	donald trump	record high	electric vehicle
trade tension	guardian view	two year	carlos ghosn	wall street	stock rally

Topic 19	Topic 20
observer business	jim cramer
business agenda	business live
wall street	bank england
long read	wall street
trade tension	trade war

Appendix 2. Word distribution for each topic via Trigram

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
theme week ahead	weekly jobless claim	nil pratley finance	nil pratley finance	china trade deal
world market theme	cramers week ahead	jaguar land rover	raise interest rate	project syndicate economist
interest rate rise	global stock market	politics live andrew	london stock exchange	defense contract pentagon
hit record high	business make simple	live andrew sparrow	north american trade	hit record high
business live pound	house price rise	trade deal china	amid trade war	cramer break down
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
nil pratley finance	wall street rise	cramers lightning round	project syndicate economist	box office analysis
french finance minister	million defense contract	china trade talk	nil pratley finance	saudi crown prince
saudi energy minister	profit beat estimate	wall street week	wall street end	steel aluminum tariff
china trade talk	house price fall	street week ahead	sir philip green	china trade hope
hit new high	pay million settle	trump china trade	wall street tumble	consumer champion why
Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
boeing 737 max	china trade war	china trade deal	guardian brexit watch	boeing 737 max
737 max jet	737 max software	white house adviser	cramers game plan	737 max crash
ground boeing 737	nil pratley finance	trade war fear	brexit watch brexit	market theme week
uschina trade deal	problem anna tims	wall street rally	observer business agenda	737 max plane
amid brexit uncertainty	observer business agenda	new york city	gender pay gap	uschina trade war
Topic 16	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
nil pratley finance	everything jim cramer	observer business agenda	boeing 737 max	cramers lightning round
china trade war	jim cramer mad	business live market	william keegans view	lightning round buy
fed rate cut	cramer mad money	amid coronavirus outbreak	737 max flight	737 max grounding
business live stock	mad money include	small business loan	guardian opinion cartoon	jim cramer warn
business live bank	cramers lightning round	stock market rally	beat profit estimate	audio long read