Latent Dirichlet allocation assisted time series of financial news sentiments

Oleg Y. Rogov[1], Elena A. Fedorova[2], and Igor S. Demin[2]

[1] Moscow Institute of Physics and Technology, Dolgoprudniy, Russia fintech@gmx.ch
[2] Financial University under the Government of the Russian Federation, Moscow ecolena@mail.ru

Abstract. Quantifying the financial stability of a country's financial system plays one of the dominant roles in the market risks evaluation. Ordinary methods that help measure financial instability may have instinctive shortcomings due to their unavailability or one-sided nature. Thus, finding new paths to quantifying and analyzing financial instability is of a high priority today. Financial texts data feature specific sentiments and informative patterns that can be extracted with modern natural language processing techniques and machine learning algorithms. We report a LDA-assisted sentiment analysis of the 3 million Thomson Reuters news texts collection. Decomposing the textual data into relevant topics we calculate the sentiment-based indicator of a critical response in the selected news articles.

Keywords: Text analysis · LDA models · natural language processing.

1 Introduction

Dealing with the raw texts sources available via social networks or news agencies has become gradually crucial for a variety of business intelligence tasks [1], as well as challenges of evaluating the financial stability of a country in general or a set of commercial entities [2]. With this trend, a problem arises to find appropriate tools [3, 4] to processing the extensively unstructured textual data that holds a significant share of the information data stream today.

We employ LDA [6] approach to perform a topics modeling and analysis for the news dataset formed with the Thomson Reuters Agency. The result could give readers an overall but specific analytics report. Furthermore, the approach features clustering of the news texts into different groups, which could present a comprehensible stem of news texts and be used to define unknown documents. Given that, the relevant sentiment analysis could be applied.

2 Methodology

Latent Dirichlet Allocation is the one of the basic yet most used probabilistic topic modeling algorithms. The primary assumption behind the algorithm is that the given news text is a mixture of multiple themes (topics). Given a dataset comprised by the news texts, one can employ the LDA statistical framework to explore beyond the thematic distribution representing each news text. The probability that tokenized word z_i is assigned to topic j, assuming that C^{WT} is the term-topic matrix, α is the parameter that sets the theme distribution over the news texts. The higher α the more widespread the texts shall be over the specific number of topics K. Likewise, η is the parameter describing the distribution of the words throughout the theme:

$$P(t_i = j|t_{-i}, w_i, d_i) = \frac{C^{WT}_{w_i j} + \eta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\eta} \times \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \tag{1}$$

3 Empirical results

3.1 Data

The data is collected via the Thomson Reuters agency website [7]. The dataset contains articles in the date range January 2012 - June 2018, 2.9 mln news texts in total. Each news text is marked with the corresponding publication date-time stamp. This fact allows one to construct a time-dependent function of textual sentiments. Prior to sentiment calculation, all the news texts were subject to pre-processing scripted routines such as removing the stop-words, and text normalization.

Applying the LDA algorithm yields 37 dominant topics with the relevant most frequent terms. The latter are calculated based on TF-IDF frequencies.

3.2 Sentiment-based indicator

The sentiments of the Russian news subset are obtained with the nltk-based author-modified library for every news text, and included the following polarity variations: positive, neutral, negative, compound. The sentiments of the financial texts are calculated according to the Loughran-McDonald dictionary.

Given the polarities of the texts, we construct a sentiment-based indicator that describes the coverage of the events related to the Russian companies and banks (see Fig. 1).

The Critical Response Indicator (CRI) is primarily based on two major text polarities: the text negativeness and neutrality. The indicator's dynamics over the past decade shows major events that affect the sentiment polarity of the news. For example, the dip of the curve in April, 2016, corresponds to the publication date of so-called leaked 'Panama papers'.

4 Conclusion

Although unstructured textual data is usually difficult to analyze, we employ the pre-processed unstructured news data to extract sentiment polarities. We

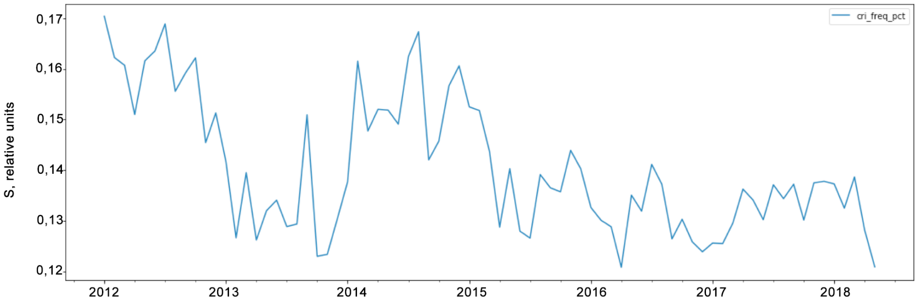

Fig. 1. The CRI indicator performance over time (re-sampled quarterly, mean), highlighting the most significant events covered in news media.

use the LDA to determine the most plausible number of the news topics in order to create a relevant subset of texts. Based on the subset, a sentiment time series indicator is then constructed thus describing the news media coverage of the specific topic over time. In this regard, the proposed approach can be efficiently used for decomposing and classifying news texts to the relevant topics with a subsequent sentiment series construction which are commonly used as a robust and effective analytic tool.

While the overall volume of financial media reports and relevant market data continues to grow at a rapid pace, a rational and well-informed decision making process becomes a cornerstone with the practitioners of financial analysis and risk managers.

References

1. Larsen, V. H.; Thorsrud, L. A.: The value of news for economic developments. Journal of Econometrics, 2018, 207(1): pp. 140-154.
2. Li, G.; Shi, F.; Tu., J.: Textual analysis and machine leaning: Crack unstructured data in finance and accounting. The Journal of Finance and Data Science, 2016, 2(3): pp. 153-170.
3. Chan, S.W.K.; Chong, M.W.C.: Sentiment analysis in financial texts. Decision Support Systems, 2017, 94(1): pp. 53-64.
4. Moreno-Ortiza, A.; Cruz, J.F.: Identifying polarity in financial texts for sentiment analysis: a corpus-based approach. Procedia - Social and Behavioral Sciences, 2015, 198(1): pp. 330-338.
5. Loughran, T.; McDonald, B.:Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research, 2016, 54: pp. 1187-1230.
6. Blei D. M. et al.: Textual Analysis in Accounting and Finance: A Survey. Annals of Applied Statistics, 2007, 1(2): pp. 634-634.
7. URL: Thomson Reuters, https://www.reuters.com/.