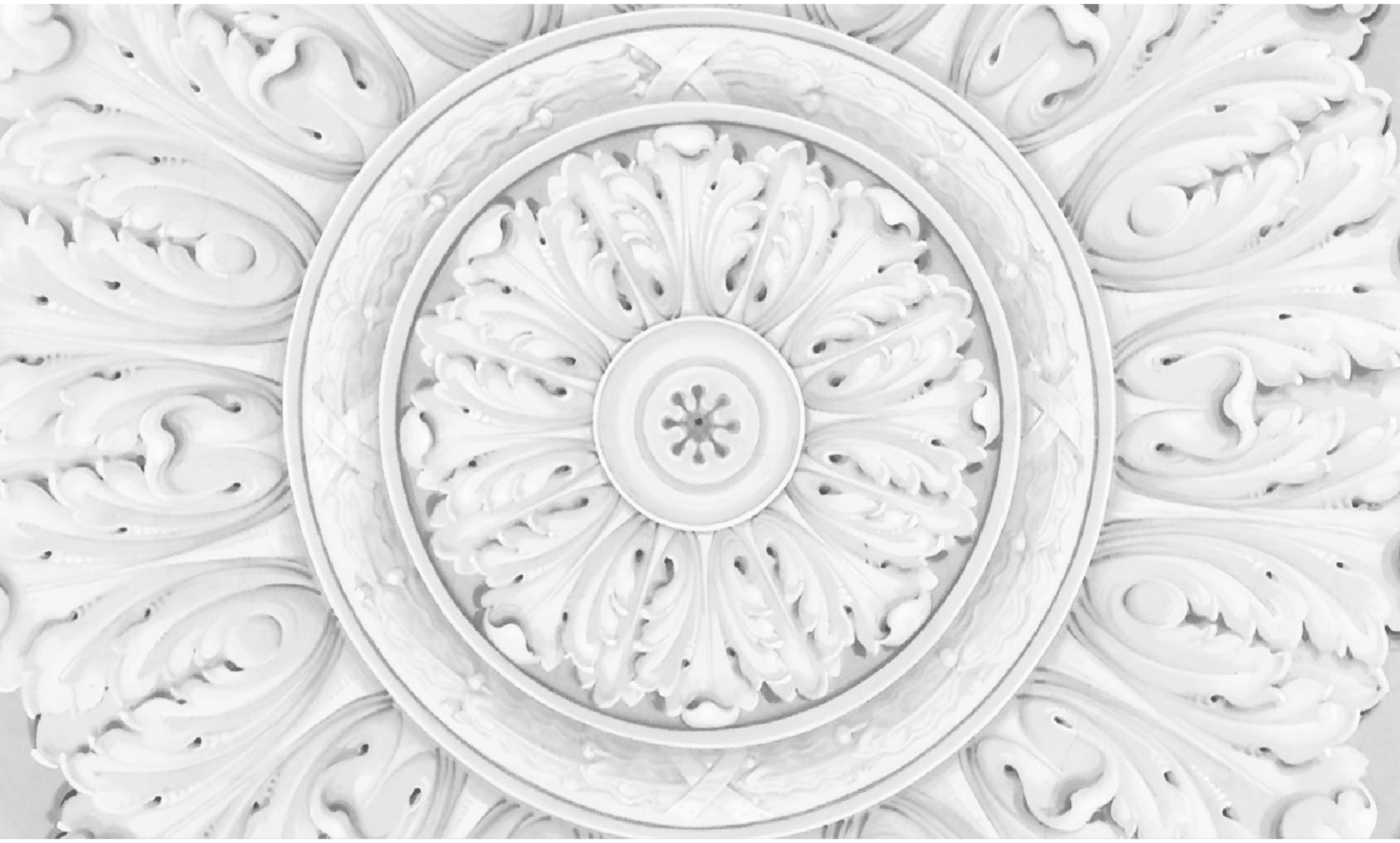
Bank of Russia
The Central Bank of the Russian Federation



WORKING PAPER SERIES

Ksenia Yakovleva

Text Mining-based Economic Activity Estimates

№25 / October 2017

Ksenia Yakovleva
Bank of Russia, Research and Forecasting Department
E-mail: YakovlevaKV@cbr.ru

The author would like to express her gratitude to A. Kruglova and to all participants of seminars held at the Bank of Russia for their valuable suggestions and comments.

Address 12 Neglinnaya Street, Moscow, 107016
Telephone +7 495 771-91-00, +7 495 621-64-65 (fax)
Website www.cbr.ru

The views expressed in the paper(s) are solely those of the author(s) and do not necessarily represent the official position of the Bank of Russia. The Bank of Russia is not responsible for the contents of the paper(s). Any reproduction of this paper shall be subject to the consent of the author(s).

Abstract

This paper outlines the methodology for calculating a high-frequency indicator of economic activity in Russia. News articles taken from Internet resources are used as data sources. The news articles are analysed using text mining and machine learning methods, which, although developed relatively recently, have quickly found wide application in scientific research, including economic studies. This is because news is not only a key source of information but a way to gauge the sentiment of journalists and survey respondents about the current situation and convert it into quantitative data.

Key words: economic activity estimates, text mining, machine learning.

JEL classification: C51, C81, E37.

INTRODUCTION

Big Data has gained particular popularity recently due to the huge rise in the volume of digital information available thanks to increasingly widespread access to technology and the Internet, which allows users to continually create new information. This includes practically all freely-available online data, in particular: online retailers, job search engines, news websites, social media, blogs, etc. However, a great deal of the information available online is unstructured, that is a text. This makes it impossible to independently process a large volume of information. Therefore, researchers in various fields, including economics, are developing new statistical approaches to unstructured online data mining and analysis.

The main advantage of online data compared to conventional statistics is their great variety and the possibility it offers of calculating an indicator that is not factored into official statistics. It is also important to note that online data has a much higher frequency and immediacy compared to official statistics.

The aim of this paper is to construct a high-frequency indicator calculated using daily news in order to assess the dynamics of economic activity in Russia. There is a need for this indicator due to the lack of similar indicators and the delayed release of official data on economic dynamics. For example, the main indicator of economic growth, GDP, is published on a quarterly basis 1-1.5 months after the quarter-end. This makes it impossible to monitor economic dynamics in real time and make appropriate decisions.

All this testifies to the relevance of using big data in economic research. Moreover, it is set to become increasingly relevant due to the great demand for this type of research.

Big Data in Economics is a new area especially in Russia: there is practically no research on this subject currently conducted in Russia, which demonstrates its novelty. This paper is based on the slightly modified approach outlined in an article by A. Thorsrud (2016).

Section 1 of this paper introduces the methodology of the model and the benchmark data; Section 2 presents the outcomes of the model.

1. BENCHMARK DATA AND MODEL SPECIFICATIONS

1.1. Data

This paper uses two types of data: unstructured and structured. Daily news articles taken from an online resource are an example of unstructured data i.e. data that are not organised in a set data structure. The second type of data is a monthly statistical indicator - the composite PMI (Purchasing Managers Index) index of business activity. The PMI index of business activity is used as a proxy for GDP (due to the insufficient time series of news articles[1]).

News articles were selected from an economic website chosen due to the wide coverage of economic news, lack of irrelevant topics, and the simplicity of web scraping[2]. Approximately 50,000 articles were selected with a total volume of 20-25 million words: an acceptable volume for analysis.

Composite PMI index of business activity data were taken from the Bloomberg agency website.

The time sample from January 2014 through August 2017 was used to construct the indicator.

1.2. Model

There are three main stages in the process of constructing a news index. The first stage involves extracting a list of topics contained in the news texts. The second stage involves determining the tone of the news texts, categorising topics as positive or negative, and tracking their trends. The third stage involves constructing a linear regression, where the dependent variable is PMI, and regressors are news topics converted using principal component analysis.

Before embarking on the three stages of index construction, it is first necessary to prepare the data i.e. convert an unstructured text into a structured format. Data preparation is an important step in text analysis as, firstly, it reduces data dimensionality,

[1] A longer time-series is required to construct the regression analysis due to the fact that GDP is calculated on a quarterly basis.

[2] Software used to automatically extract and save data from websites.

which accelerates information processing; secondly, comprehensive text preparation at the outset results in higher-quality and more thoroughly-interpreted topics in the end.

The preparation process involves several steps. The first step is stemming - reducing inflected or derived words to their root form using the MyStem program, free software that conducts morphological analysis of the Russian language created by Yandex in 1997. An article written by one of Yandex's founders, I. Segalovich, outlines how it works. The second step involves text processing: removing punctuation, numbers, unnecessary spaces, and stop words[3]. Filtering news texts in this way significantly reduces the benchmark data without losing the semantic component.

The processed words in the filtered texts are called terms and are used as the basis for a document-term matrix (dtm). Each row of the matrix denotes an individual term, and each column is a separate document.

The dtm is used in the first stage of the analysis - identifying topics in the corpus of texts. The topics in this paper were identified using probabilistic topic modelling, a statistical method which defines the themes that arise from a set of words and order the words according to its specificity to the theme.

One of the most popular probabilistic topic modelling methods is the latent Dirichlet allocation (LDA) method. It was proposed for use in text analysis by D. Blei, A. Ng, and M. Jordan in their 2003 article that demonstrated that each document has several topics and that the LDA method can be used to identify the probabilistic distribution of each one using the model employed in the data retrieval system.

Latent Dirichlet allocation is a three-level hierarchical Bayesian model, in which each document in the corpus is modelled as a collection of underlying, latent topics. According to LDA, each word in a text document belongs to an unknown topic, and each topic is modelled from the initially specified probabilities of the topics:

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta)\, p(w_n \mid z_n, \beta) \quad ,$$

where α и β are specified vectors – Dirichlet distribution parameters;

$\theta \sim \mathrm{Dir}(\alpha)$ is topic proportion in each document;

$z_n \sim \mathrm{Dir}(\beta)$ is word proportion in each topic;

w_n is the correspondence between words within a document and the topics.

[3] Stop words are link words bearing no sense load. Stop words include conjunctions and connective words, pronouns, prepositions, particles, interjections, demonstrative and parenthetical words, as well as some nouns, verbs and adverbs.

The end process of the LDA model produces vectors showing how the topics are distributed in each document, and distributions showing which words are more likely to appear in each topic. The result of the LDA model is a list of words most characteristic for each topic. The list of words is ranked on a decreasing scale and the first five words are selected to form the basis used to calculate the frequency of topics that occur.

In order to obtain data on a daily basis, all of the day's articles are collected and analysed. Then, the first five words of each topic are used to calculate their frequency per day. The data obtained shows the frequency at that each topic is mentioned per day but they do not reflect the tone of the topic (positive or negative). Therefore, the second step involves constructing an index to determine the tone of the text.

In literature, there are four basic approaches to defining text tonality:

1) Rule-based approaches;
2) Dictionary-based approaches;
3) Supervised learning;
4) Unsupervised learning.

The work by A. Thorsrud (2016) that was used as the basis for the construction of this index determines text tonality using a dictionary-based approach. The Harvard IV-4 Psychological Dictionary, which has already defined positive and negative words, was used to identify a positive or negative tone. However, this dictionary does not reflect the specifics of economic terminology and the Russian language. Therefore, the approach employed in this paper to define the text tonality was based on supervised learning as it generally demonstrates a high quality of classification. The chosen method was a support vector machine (SVM), which marks samples as belonging to two categories using a separating hyperplane, so that the distance from it to the nearest data points of the set is maximised.

However, this method requires a training collection, i.e. a set of similar texts that have already been classified as positive or negative. Therefore, the first step involved compiling a training collection. Each news article in the training collection was manually attributed to either a positive or a negative class. In the second step, this training model was applied to the complete set of texts to be evaluated.

Yet, not all texts can be unequivocally classified as positive or negative, some of them are written in a neutral tone. SVM modelling assigns a tonality classification “+1” or “-1” and the probability of the text belonging to a particular tonality classification. Consequently, if the model defined the text tonality with a probability of less than 60%, its tone was determined to be neutral. Texts classified as neutral were excluded from the analysis as they do not contain information relevant to this paper.

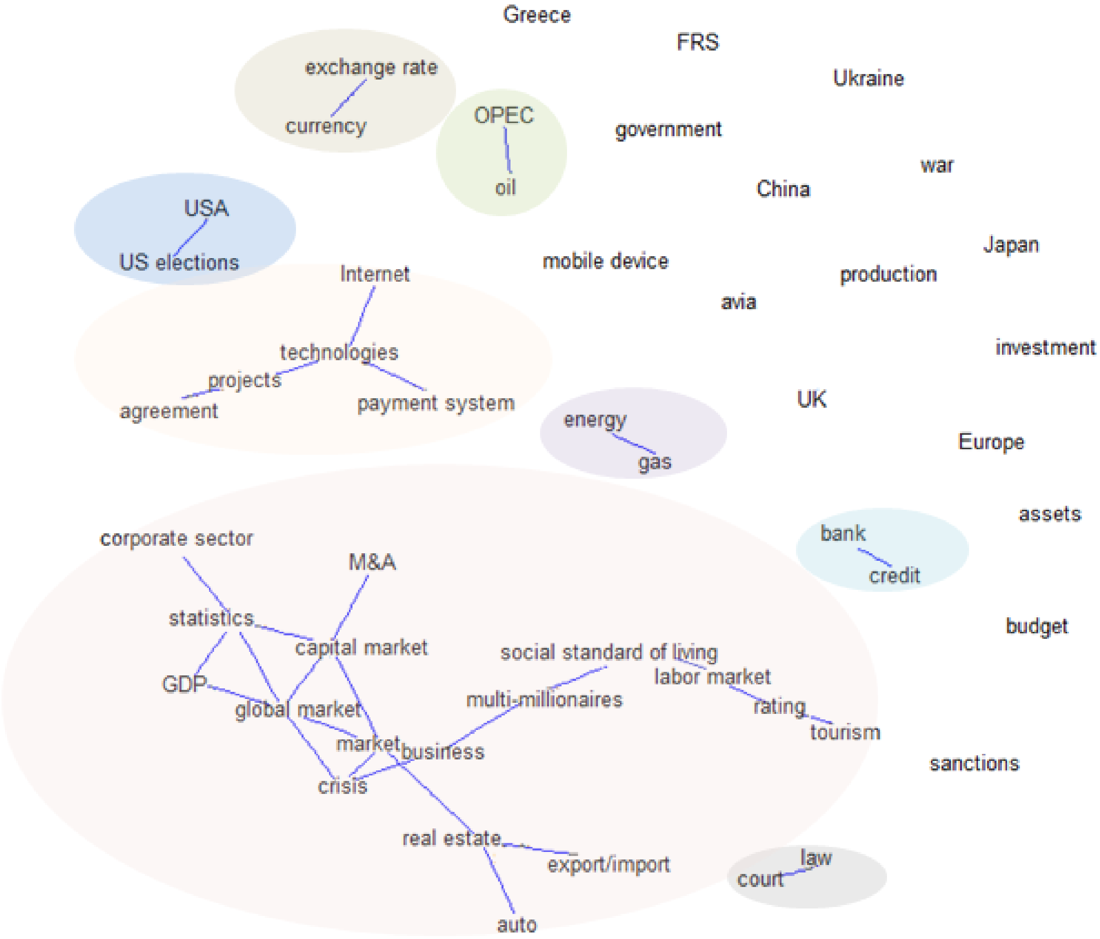
As a result, all the topics obtained in the first stage were adjusted to account for tonality and were regressors in the econometric model constructed in the third stage. Multiple linear regression was used as the econometric model.

Principal Component Analysis (PCA) was used to reduce data dimensionality. PCA was chosen because it reduces the number of regressors while maximally avoiding the loss of information.

2. ANALYSIS OUTCOMES

According to the LDA model, 50 topics were identified that provide the best statistical decomposition of the corpus. The LDA model does not name the topics but it is possible to identify and subsequently name each topic by reviewing the most frequent words in each group. For example, from January 2014 to January 2017, the main topics covered in news articles were related to the exchange rate, oil, the banking sector, events in the USA, etc. (Figure 1).

Figure 1. Topics revealed using the LDA method and the connections between them



Source: authors' calculations.

The figure also shows a network visualisation of the topics where all connections that have a correlation of less than 20% have been removed. Correlation was calculated using a matrix where the rows were made up of 50 topics, and the columns were all the terms used by a news site during the analysis period resulting in a 50x243577 matrix, where the probability of each word (term) appearing in each topic was calculated at the intersection of the rows and columns. Then, a 50x50 correlation matrix was calculated between the themes based on these probabilities.

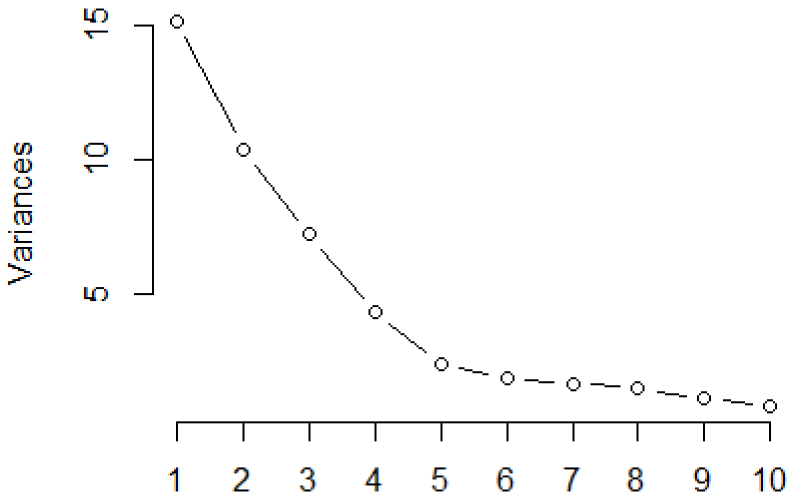
The figure shows that half of the topics have been grouped into several large and small clusters. These clusters will not be used in further analysis but it helps visualise the topics being analysed.

As a result, 50 daily time series were calculated based on the 50 identified topics, which were later adjusted for tonality as categorised by the SVM method. In order to assess the quality of the SVM method, the training sample was divided into two parts. A model was built using the first part (9/10 of the training data), and the accuracy of the algorithm was verified using the second part (1/10 of the training data). Accuracy was calculated as the proportion of correctly predicted values out of the total values and came to 68%, which is a fairly good result.

In order to eliminate the daily statistical noise in the time series of each topic, the data was smoothed using a moving average over 80 days. Daily topics were compared with the monthly PMI by converting daily topics into monthly ones by finding the monthly average. As a result, 50 monthly time series regressors were obtained, each one characterising a particular topic.

The PCA method was used to construct the linear regression, which reduced the dimensionality of the regressors (Figure 2).

Figure 2. Topics transformed using PCA



Source: authors' calculations.

Out of 50 initial topics, there are four regressors that clearly represent the dependent variable, the PMI economic activity indicator. Regression analysis showed that the first component does not have significance in the equation constructed, while the second to fifth components have significance and account for 85% of the regression (Table 1).

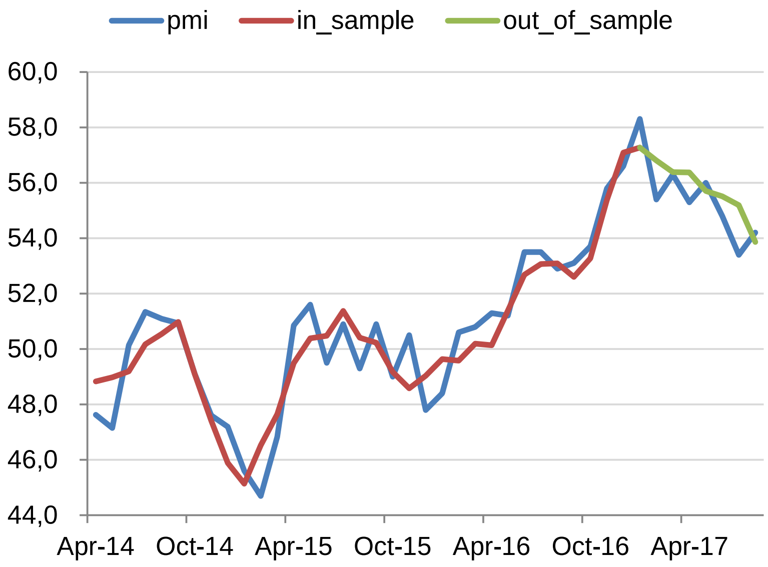
Table 1. Regression analysis on PMI economic activity indicator

Dependent variable = PMI index		
Variable	β	t-statistic
PCA(2)	0.2329	10.28 ***
PCA(3)	-0.1054	-3.87 ***
PCA(4)	-0.1268	-3.60 **
PCA(5)	-0.2368	-5.00 ***
F-statistic	39.67 ***	
R 2 – adjusted	0.8455	

*Note: ***; ** and * are the significance of the coefficient estimate at 0.1; 1 and 5% respectively.*

In order to validate the linear regression, the sample was divided into training (in sample) and control (out of sample) sets. The training sample covers the period from January 2014 to January 2017, the control sample covers February 2017 to August 2017. Comparing the control sample with actual PMI data indicates that the model has a reasonably good predictive strength (Figure 3).

Figure 3. PMI and the calculated news index



Sources: IHS Markit, authors' calculations.

The mean absolute error (MAE) was also used to estimate the quality of the model. The MAE for this forecast was 0.81%. Moreover, the mean absolute error of forecasting using a different model, the first-order autocorrelation model AR(1), was 2.7% (Table 2).

Table 2. Mean absolute error of forecasting models

	Mean absolute error (MAE)
News-based model	0.81
AR(1) Model	2.7

CONCLUSION

This paper presents a model that estimates economic dynamics based on news articles. The calculations given above show that using unstructured data such as news is just as important as conventional statistical indicators when forecasting economic activity.

The methodology developed successfully offers a solution to the problem of forecasting economic dynamics, which is evidenced by estimates of the model's quality. Therefore, we can conclude that news-based data have a fairly good predictive power. The news index presented in this paper can be used to monitor the dynamics of economic activity on a daily basis as well as develop other indicators that will make it possible to react more promptly to the current economic situation and make appropriate decisions.

REFERENCE

1. Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. January, 2003.
2. Bloom N., Baker S., Davis S. Measuring Economic Policy Uncertainty // THE QUARTERLY JOURNAL OF ECONOMICS. November, 2016.
3. Doms M., Morin N. Consumer Sentiment, the Economy, and the News Media // Finance and Economics Discussion Series (FEDS). September, 2004.
4. Kholodilin K., Thomas T., Ulbricht D. Do media data help to predict German industrial production? // Journal of Forecasting. 2017.
5. Shapiro A., Sudhoh M., Wilson D. Measuring News Sentiment // FEDERAL RESERVE BANK OF SAN FRANCISCO WORKING PAPER SERIES. January, 2017.
6. Thorsrud A. Words are the new numbers: A newsy coincident index of business cycles // Norges Bank Research. Working Paper. February, 2016.
7. Голощапова И., Андреев М. Оценка инфляционных ожиданий российского населения методами машинного обучения // Вопросы экономики, июнь 2017 г., № 6.