Text Mining-based Economic Activity Estimation

Ksenia Yakovleva, *Bank of Russia*
yakovlevakv@cbr.ru

This paper outlines a methodology for constructing a high-frequency indicator of economic activity in Russia. News stories from internet resources are used as data sources. News data is analysed using text mining and machine learning methods, which, although developed only 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.

Keywords: *economic activity estimates, nowcasting, text mining, machine learning, Big Data, data mining, topic modelling, sentiment analysis*

JEL Codes: *C51, C81, E37*

Citation: Yakovleva K. (2018). Text Mining-based Economic Activity Estimation. *Russian Journal of Money and Finance*, 77(4), pp. 26–41.

doi: 10.31477/rjmf.201804.26

1. Introduction

Understanding the current economic situation and its future dynamics is extremely important for implementing timely and effective economic policy. Of special value to politicians, businesses, and other economic agents is the ability to track real-time economic and financial indicators to enable prompt decision-making.

Scientific literature has recently devoted increasing amounts of attention to the development of high-frequency internet-based indicators. This data includes virtually all publicly available information: online-shopping websites, job search engines, news sources, social media, etc. That said, the internet presents practically all of this information in an unstructured form, i.e., as text. This makes it impossible for the individuals to promptly process large information volumes on their own, so researchers in a variety of areas, economics among them, are developing new statistical methods to extract and analyse unstructured internet data.

Interest in this kind of information stems from its advantages over conventional statistical data. These advantages are provided by the sheer volume of information and the speed of its collection, as well as data variety and reliability. On top of that, it can be used to construct new, unique indicators that are not available in official statistical reports.

In this study, we constructed a high-frequency indicator (computed using a daily news flow), which enabled a country's economic activity trends to be assessed. The PMI index[1] was chosen as a target indicator reflecting economic activity. The results obtained suggest that text analysis allows for the monitoring of the current economic situation on a daily basis with fairly high accuracy.

Section 2 of the paper reviews the key literature dealing with the use of text analysis and Google Trends in the financial and economic areas. Section 3 provides a methodology for constructing the news index. The Conclusion sums up the results of the study and considers the ways of refining the news index.

2. Literature review

The foreign literature provides a fairly large number of papers studying text analysis in economics. These papers mainly deal with short-term forecasting – nowcasting – using text information. Although forecasting stock market movements is a challenging task because of the market's high volatility and heavy data noise, quite a few papers explore this subject. Alsing and Bahceci (2015), for instance, attempt to forecast movements in share prices of three major companies (Walmart, Netflix, and Microsoft) using relevant publications in the Twitter social network. This social network was chosen because it provided the opportunity to extract user opinions and conduct their sentiment analysis[2] in order to predict future share price movements. The study found an artificial neural network that predicted Walmart stock price movements with 80-percent accuracy to be the best model.

Economic research papers on text analysis usually deal with short-term forecasting (nowcasting) of GDP and industrial output growth, inflation, the unemployment rate, sales, and other indicators.

Ardia and Bluteau (2017) attempted to predict U.S. industrial output growth based on newspaper stories covering the period from 2001 to 2016. The model thus constructed provides evidence that text analysis dramatically improves the accuracy of industrial output growth forecasting over horizons of nine months and one year.

Another study (Nyman et al., 2014) suggests that text analysis substantially improves the accuracy of forecasting the performance of the Michigan Consumer Index[3] performance. The study only used texts containing the words 'anxiety' and 'excitement'. Two time series were constructed using the average number of occurrences of each word per story in each month, with the difference between them established. The indicator thus constructed was included in the regression, enhancing the model's predictive power.

[1] Purchasing Managers Index is a macroeconomic indicator of business activity in the manufacturing and services sectors calculated based on surveys of managers.

[2] Sentiment analysis identifies and categorises the author's emotional attitude to a particular object.

[3] MCI is a monthly index of American consumers' confidence constructed by Michigan University.

Entirely new, high-frequency indexes can be constructed based on text analysis. One example is the Uncertainty Index (Bloom and Baker, 2016). The idea behind constructing this index was to aggregate newspaper stories containing the three keywords related to uncertainty, the economy, and policy, so as to monitor the instability of economic policy. This index was constructed for various countries (including Russia) and various activity areas, such as health service and national security.

Also worth noting among the tools for nowcasting economic indicators is Google Trends[4], which analysts have started using extensively for short-term forecasting of various indicators. Information obtained from Google Trends is used primarily to get an insight into the current economic activity and sentiment.

Most of the papers claim that using Google Trends as part of forecasting models improves the accuracy of forecasting results. Choi and Varian (2012) used Google search query statistics for nowcasting car sales. The study finds that the first-order autoregression model with the Google Trends time series has a better predictive power than that of the conventional first-order autoregression model.

Aside from car sales, Google Trends improves the accuracy of forecasting jobless claims, consumer confidence, and tourist statistics. The regression results have shown that autoregression models (AR-models) that include the Google Trends variables yield a 5 – 20% higher prediction accuracy than conventional AR-models (Choi and Varian, 2012).

Studies by Russian researchers have also started to use text analysis and the Google Trends service. Goloshchapova and Andreev (2017), for example, suggested a new approach to constructing inflation expectations indicators based on the algorithms of internet user opinions (comments) regarding inflation. It was found that the index measuring inflation expectations based on the internet may become a viable analogue of official survey data.

3. Main results

The key macroeconomic indicator reflecting a country's economic activity is GDP growth, which all analysts, the business community, and various government agencies keep track of. However, GDP data is published on a quarterly basis, and with a significant lag, which prevents essential information from being obtained in a timely manner. A large number of papers concerned with GDP nowcasting have sought to address this drawback. With digital information and Big Data technology development, GDP nowcasting has also started to be gradually modified. (Kapetanios and Papailias, 2018). GDP nowcasting can rely on a great variety of data, ranging from textual information

[4] Google website analysis of users' searches of a particular term relative to the total number of search queries.

to electronic payment system data on credit and debit card transactions (Galbraith and Tkacz, 2015).

Thorsrud (2016), for instance, only used textual information from daily business newspapers to forecast GDP growth. The author deconstructs news topics into time series using latent Dirichlet allocation and employs the dynamic factor model for forecasting GDP.

Thorsrud (2016) was taken as a basis for this study but with a number of differences. One of them is that it is not the quarterly GDP number that is forecast but the diffusion index of business activity[5] (Purchasing Managers' Index, PMI)[6] calculated on a monthly basis. GDP was replaced because of its limited news database, as quarterly GDP numbers are insufficient for forecasting.

Diffusion indices are superior in that they have a strong predictive power and are leading indicators closely correlated with business cycles.

PMI, released on the first business day of the month[7], is a key diffusion index, which fairly accurately tracks business cycle dynamics, as evidenced by its close correlation with GDP (Figure 1).

3.1. News data collection and processing

News data is the basis of this study, as the business activity index will be forecast based on them. An important problem to be addressed in dealing with text analysis is the choice of a news source, as this should reflect the current economic situation with maximum accuracy and cover a significant proportion of the population.

The following key criteria are to be met in choosing a news source. First, the news should be concerned with economic issues; second, a sufficiently long time series should be available on the internet (at least 3 – 4 years); and third, web scraping[8] should be simple, i.e., information should be easy and fast to extract from the website.

A news resource solely devoted to economic developments both in Russia and abroad was found to meet these criteria. It should be noted from the start that using just one data source is most probably insufficient for such analysis to be carried out, as it may be affected by a number of factors, such as less than comprehensive coverage of economic topics, a specific target audience, etc. On the other hand, Thorsrud (2016) also uses just one business newspaper, the fourth

[5] A diffusion index constructed based on survey data. Each respondent answers questions related to business conditions: new orders, the labour market, contract performance time, and so on.

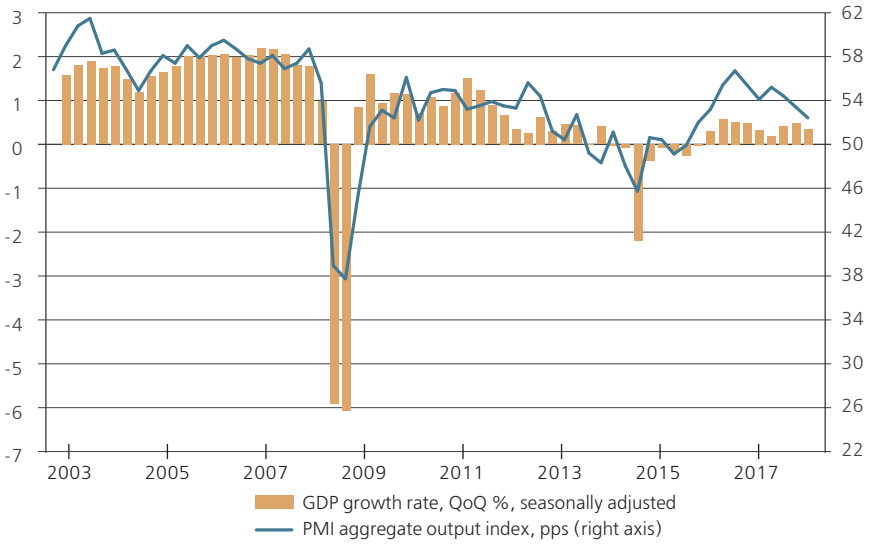
[6] The Composite PMI Index is a weighted average of the Manufacturing Output Index and the Services Business Activity Index. The Manufacturing PMI Index is based on five key indicators with the following weights: new orders – 0.3, output – 0.25, employment – 0.2, timing of raw materials and supplies deliveries – 0.15, raw materials and supplies inventories – 0.1. The Services PMI Index is calculated by weighing percentages of respondents' answers with the following weights: improvement/growth – 1.0, unchanged conditions – 0.5, worsening/decline – 0.0.

[7] https://www.markiteconomics.com

[8] Web scraping – a technology for extracting data from web sites.

largest in terms of readership. Still, reliance on just one source may become a problem, so further research is needed to increase the number of sources used.

Figure 1. GDP growth rate and PMI



Source: Rosstat, IHS Markit

The resulting sample had just over 59,000 news items (Table 1). News items were collected over a time span of four and a half years – from 2014 to 2018 – and contained 54 month-long periods[9]. The sample was from the outset split into training and testing parts, with the former used to construct the model and optimise its parameters, and the latter serving as a tool to assess the quality of the model constructed. The learning sample, as a rule, accounts for 75 – 80% of the raw data, although there are no strict rules in this regard.

Table 1. Descriptive statistics of news data

Period	January 2014 – June 2018
Number of news stories mined, total	59 172
Number of news stories per month (on average)	1 096
Learning sample	January 2014 – January 2017 (37 months) – 68.5%
Test sample	February 2017 – June 2018 (17 months) – 31.5%

As the text is unstructured data, it needs to be converted into a structured form, allowing for the identification of unknown patterns and construction of topics.

[9] https://www.cbr.ru/Content/Document/File/33591/wp25_e.pdf

Before textual data is converted into a quantitative structured vector, it needs to be processed. The processing of news data is an important stage of text analysis: first, it allows for the reduction of data dimensionality, thereby drastically accelerating information processing; and second, a better text preparation from the outset helps obtain final results of higher quality that can be more easily interpreted.

Processing involves several steps. The first step is stemming: reducing words to their stems using the free MyStem[10] software developed by Yandex in 1997 to conduct morphological analysis of the Russian language. It's operating principles are outlined in an article by one of Yandex's founders, Ilya Segalovich (Segalovich, 2003). The second step involves text processing: removing punctuation, numbers, unnecessary spaces, and 'stop words'[11]. Filtering news texts in this way significantly reduces the raw data without losing its semantic component. Table 2 provides examples of raw and processed texts. The processed words in the filtered texts are called terms and 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.

Table 2. Input and processed news texts

Raw news text	Processed news text
U.S. initial jobless claims rose last week (July 21–28) to 365 thousand, according to U.S. Labor Department report, up 8 thousand from last week's revised number of 357 thousand (versus an initial estimate of 353 thousand). Experts surveyed by Bloomberg expected the jobless claims to increase by 17 thousand over the week of July 21-28 from the earlier announced level to reach 370 thousand. Thus preliminary data suggest that jobless claims rise was twice as low as forecast.	*U.S. initial jobless claim last report US labor department revise last initial expert survey bloomberg expect rise claim announce level preliminary low forecast*

3.2. Construction of a quantitative indicator

Two indicators were constructed as part of this study. The first indicator reflects the quantitative component of the topics, i.e., how frequently the topic is mentioned in news data. The second represents the emotional component, i.e., the tone of the news.

Since we are interested in all news items rather than a specific topic, this study analyses all news data collected by web-scraping.

To identify which topic a particular news story represents, we use topic modeling enabling data to be automatically sorted out by topic.

[10] https://tech.yandex.ru/mystem/ [in Russian]

[11] 'Stop words' are connective words of minor semantic importance that connect and make logical transitions between sentences, paragraphs, etc. They include conjunctions, prepositions, interjections, particles, parenthetic words, demonstrative pronouns, as well as some nouns, verbs, and adverbs.

The most popular topic models can be divided into two groups: algebraic and probabilistic. Among algebraic models are the Vector Space Model (VSM) and Latent Semantic Analysis (LSA). Algebraic models allow for the identification of the weights of words and the assessment of similarities between the texts under consideration.

The base probabilistic models are Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). The PLSA develops the LSA model and allows a better identification of possible topics of the document. But the number of parameters increases with an increase in the number of documents, necessitating the overfitting of the model (Vorontsov and Potapenko, 2013). The LDA model was designed to eliminate this overfitting problem.

In exploring the key topic model, algebraic models were found to be unsuitable for handling the data mining task, as they largely aim to extract the key words from the text and compare them. The best among the probabilistic models is LDA, because it eliminates the main PSLA drawbacks, and it is the most widespread probabilistic model – used, among other places, in Thorsrud (2016).

The LDA model, however, also has some drawbacks, one of them being the absence of linguistic notions: it only takes into account the frequency of word occurrence, rather than the order and meaning. Another drawback is that it assumes that words and documents follow a normal distribution, whereas the Poisson distribution is closer to reality.

LDA is a three-layer hierarchical Bayesian model in which each document in the corpus is modelled as a collection of nonobservable, latent topics. According to LDA, each word in a textual document belongs to an unknown topic, and each topic is modelled from the initially specified probabilities of the topics:

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

where α and β are given vectors – the Dirichlet distribution parameters;
$\theta \sim Dir(\alpha)$ is the distribution of topics in each document;
$z_n \sim Dir(\beta)$ is the distribution of words in each document;
w_n is the conformity of the words in the document to the topics.

The LDA model requires the initial parameters, or the number of topics here, to be set. The rest of the parameters were used by default. The analysis that we conducted showed that the optimal number of topics for all news data collected was 50. When there are more topics, they tend to overlap and to be duplicated, and when there are fewer topics, several tend to be merged into one. We therefore apply the LDA model to the pre-processed corpus[12], taking the number of topics as 50.

[12] A corpus is a set of textual documents. A pre-processed corpus means that all words that do not make sense from the perspective of machine learning, such as numbers, punctuation marks and other characters are eliminated to filter out noise in the documents.

The LDA model produced a list of words (unigrams) relevant to all of the 50 topics. Figures 2 and 3 present 100 unigrams for two topics in the form of a word cloud. To identify the topics of the first and second word clouds, one needn't scan all the words. According to Lau and Newman (2010), the first 10 words contain 30% of information about the topic, which is adequate for full understanding. The most conspicuous words suggest that documents related to the first word cloud are concerned with oil, while in the second example they deal with fiscal policy.

Figure 2. Unigrams for oil

The most conspicuous words: oil, price, oil product, petrol, demand, oil inventories, volume, grow up, growth, extraction, OPEC, U.S., Saudi, Brent, market

Source: author's estimates

Figure 3. Unigrams for fiscal policy

The most conspicuous words: deficit, finance ministry, pension, revenue, reserve, current, outflows, fund, ruble, first, level, GDP, balance, budget, export

Source: author's estimates

The examples of 15 major topics constructed using LDA are given in Table 3, which provides five words for each topic. The words are then used to sort out the topics. It was decided to select five words rather than ten, in order not to overload the model, and to accelerate information processing; and also because words relevant to economic issues are informative enough, so their automatic classification should not present any difficulties.

Table 3. Key topics

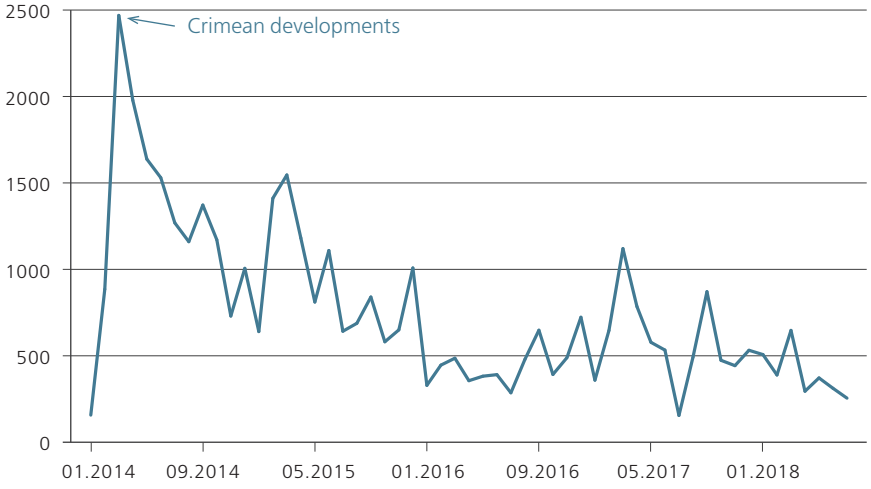
Topic	Key words (unigrams)
Topic 1	Bond, credit, yield, securities, volume
Topic 2	Dollar, ruble, exchange rate, euro, currency
Topic 3	Ukraine, Ukrainian, Crimea, Kiev, hryvna
Topic 4	Fund, investor, asset, financial, investment
Topic 5	Exports, commodity item, products, imports, ton
Topic 6	Bank, banking, Sberbank, capital, VTB
Topic 7	Debt, IMF, creditor, support, default
Topic 8	Natural gas, Gazprom, delivery, cubic meter, stream.
Topic 9	Oil field, oil, project, Rosneft, oil production
Topic 10	Development, project, investment, business, establishment
Topic 11	China, Chinese, People's Republic of China, yuan, Asia
Topic 12	Source, energy, coal, electricity
Topic 13	USA, Trump, American, Obama, state
Topic 14	Oil, price, barrel, oil production, OPEC
Topic 15	Finance Ministry, budget, revenue, expenditure, deficit

The quality of topics thus produced was assessed by comparing them with actual events. Figures 4 and 5 present fluctuations in the intensity of topics from January 2014 to July 2018, declining in some cases and rising in others. Figure 4 shows the Ukrainian topic, the intensity of which surged in March 2014, when the Crimean referendum was held. In subsequent periods, the focus on the Ukrainian topic started tapering off, and its intensity in economic media is now at its lowest since early 2014, based on our estimates.

A surge in the number of news stories is also evident in the American topic (Figure 5). Their number more than doubled because of the November 2016 presidential election in the U.S.

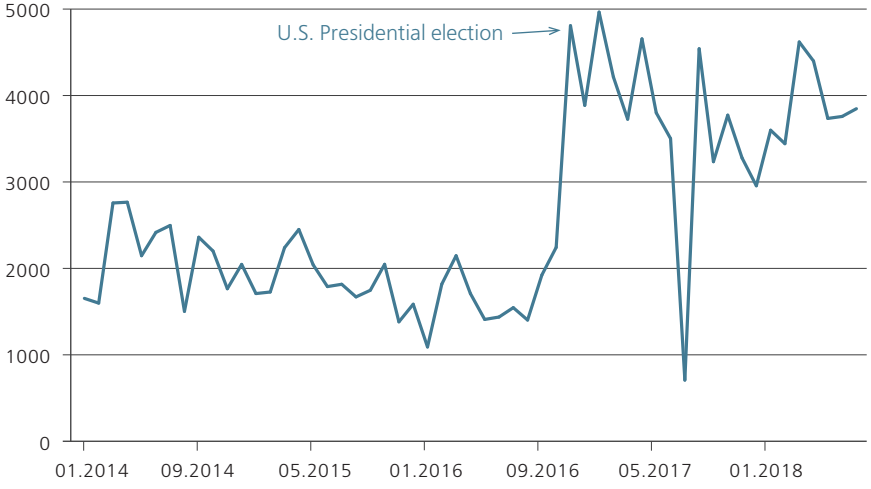
Quality assessment shows that the first five words identified using the LDA model help capture key peaks of developments, forming quantitative indicators.

Figure 4. Ukrainian topic



Source: author's estimates

Figure 5. American topic



Source: author's estimates

3.3. Construction of the emotional indicator

To recognise the emotional colouring of the text, the problem of its classification needs to be addressed. In our case, classification should seek to determine the tone of the text as 'positive' or 'negative'.

The automated analysis of the text tone often uses the following approaches:

1. rule-based approaches;
2. dictionary-based approaches;
3. supervised machine learning;
4. unsupervised machine learning.

An analysis of each approach provided in Voronina and Goncharov (2015) suggests that the rule-based approach is the most accurate but also the most difficult to implement, as well as the most labor- and time-intensive. The dictionary-based approach has substantive constraints, and currently there is no Russian-language dictionary of economic terminology. The unsupervised approach offers the lowest prediction accuracy. This study therefore used the supervised learning approach to determine the tone of the text. This approach, as a rule, provides for a relatively high quality of classification. The chosen method was the 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.

A number of empirical studies suggest that the SVM method is well suited to text classification, offering advantages over other methods (see, for example, Joachims, 1998). Basu and Walters (2003), exploring automated classification of news texts, found the SVM method to be superior to neural networks as regards classification quality. Sassano (2003) used the Reuters database to show how the SVM method could improve text classification accuracy.

To train the classifier using the SVM method, relevant tones were identified in the sample. The symbols '1' and '-1', corresponding to positive and negative tones, were assigned manually to individual news stories. A total of 3,438 news stories were used in constructing the model, of which 2,600 (76%) were chosen as 'training' and 838 – as 'test' ones (24%).

Table 4. Tones in the sample

		Model	
		Positive	Negative
Actual value	Positive	350 (TP)	96 (FN)
	Negative	206 (FP)	186 (TN)

Using values from Table 4, a metric can be computed showing the quality of the classifier that we have constructed.

$$\text{Accuracy} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} = 64\%$$

The accuracy of the classifier is computed as the share of accurately predicted values in the total number of values in the test sample. The tone was correctly predicted for 536 out of 838 news stories (64%).

The result of the SVM method is the distribution of tone estimates ('1', '-1') and their probability. If the probability is less than 60%, these texts are excluded from the sample, because they identify the tone with low accuracy and may produce biased results. For the remaining texts, the tone was multiplied by probability in

order to assign greater weights to news stories which are classified with the highest accuracy.

3.4. Construction of the forecast model

The quantitative series with topics are multiplied by the relevant tone, forming new time series which will be subsequently used to construct the forecast model for the PMI index.

To eliminate the daily noise in the time series of each topic, all the data is smoothed using the three-month (87 days) moving average. The three-month moving average was chosen because the PMI index itself has a fairly strong predictive power, since the managers surveyed take into account, among other things, the economic situation anticipated in the future. So it is not only the current month's values that need to be taken into account (conducting data smoothing for 30 days) but also data for previous months. The results suggest that the model using the three-month moving average has the best predictive power.

To compare daily topics with the monthly PMI index, the topics are converted into monthly series by identifying the average monthly value. As a result, 50 monthly time series regressors were obtained, each characterising a particular topic.

When the number of regressors (50 regressors) is larger than that of observations (34 observations), the usual linear regression cannot be used. One way of addressing this problem is to employ machine learning methods enabling a larger number of regressors to be used. They help reduce data dimensionality, minimising information losses. Among them are models with regularisation and factor models.

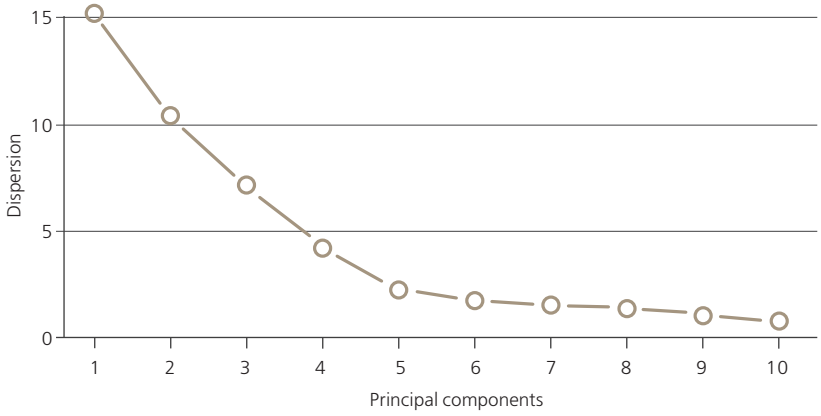
We tested LASSO and Ridge regressions as regularised models. The models' key parameters, lambda[13] and alpha[14], were chosen in such a way as to minimise the mean square errors of the equation. Lambda equalled 0.23, alpha – 0.1. As a result, out of the 50 topics initially given, the regularised model left 24, which is also a fairly large number. A large number of regressors may result in the problem of model overfitting, so the regularised models had to be rejected.

Another way of reducing dimensionality is factor analysis. Principal Component Analysis is one of the most frequently used factor analysis methods. It uses an algorithm to move from the initial data to new groups where data have similar relationships (Orlov and Lutsenko, 2016).

[13] Lambda is a regularisation parameter introducing a complexity penalty: at $\lambda = 0$, LASSO regression reduces to the conventional method of least squares; and as λ increases, the number of variables declines until it becomes zero.

[14] Alpha determines which model type is the most suitable: at $\alpha = 0$, Ridge regression is suitable, at $\alpha = 1$ – LASSO regression.

Figure 6. Share of principal component dispersion



Source: author's estimates

The first five principal components usually account for the main share of dispersion (Figure 6). The regression equation of the PMI index on the first five principal components shows that the first principal component is nonsignificant at the 10% significance level. The following four principal components are significant and explain 85% of the regression (Table 5).

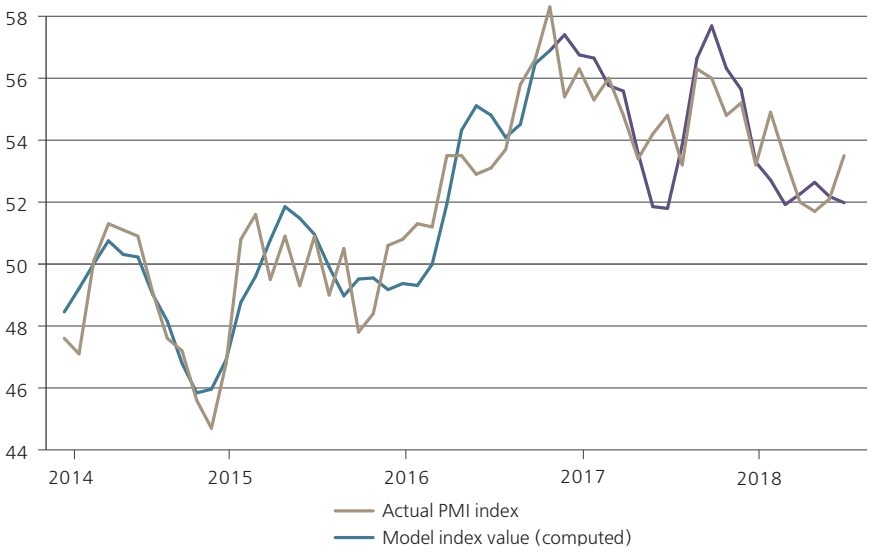
Table 5. Regression components

Regression result		
Dependent variable – PMI index		
Variable	B-coefficient	t-statistic
Principal component (2)	0.23	10.28 ***
Principal component (3)	-0.11	-3.87 ***
Principal component (4)	-0.13	-3.60 **
Principal component (5)	-0.24	-5.00 ***
F-statistic	39.67 ***	
Adjusted R^2	0.85	

Note: ***, **, and * denote the significance of coefficient estimates at 0.1, 1, and 5% respectively.

The model constructed was tested on a test sample covering the period from February 2017 to August 2018. The comparison of predicted results with the actual PMI values demonstrates the fairly strong predictive power of the model, and its adequately chosen specification (Figure 7). To assess the quality of the model constructed, mean absolute error (MAE) was used. The MAE value for this forecast equaled 1.0 percentage point, while in using the first-order autoregression model AR(1), it was 2.7 percentage points.

Figure 7. Actual PMI and computed news index



Source: HIS Markit, author's estimates

4. Conclusion

This paper presents a model estimating economic performance based on news data. The calculations presented in the paper show that the use of unstructured information such as news is as important a component of economic activity forecasting as the use of conventional statistical indicators.

The methodology developed addresses the task of forecasting economic performance fairly accurately, as evinced by the model quality metrics obtained. This suggests that news data has a fairly strong predictive power.

Moreover, this model has potential for further refinement in a variety of directions: first, expanding the available news database through the use of both other news sources and social media; and second, different topic models need to be used so as to identify the best of them. The latent Dirichlet allocation method produces fairly good results but fails to capture some relationships. Therefore, the use of bigrams (two-word combinations) instead of unigrams could be considered along with other topic methods. This also applies to the identification of the text tone, the accuracy of wich needs to be brought to 85% – 90%.

The news index obtained can be used not only to monitor economic performance on a daily basis but also to develop other indicators, allowing faster responses to the current economic situation and prompt decision-making.

5. References

Alsing, O. and Bahceci, O. (2015). *Stock Market Prediction using Social Media Analysis.* Stockholm, Sweden: KTH Royal Institute of Technology.

Ardia, D. and Bluteau, K. (2017). *Questioning the News About Economic Growth: Sparse Forecasting Using Thousands of News-Based Sentiment Values.* Preprint submitted to SSRN, July 21.

Basu, A., Watters, C. and Shepherd, M. (2003). Support Vector Machines for Text Categorization. In: *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03)*. Washington, DC: IEEE Computer Society.

Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, pp. 993–1022.

Bloom, N., Baker, S. and Davis, S. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4), pp. 1593–1636.

Choi, H. and Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88(s1), pp. 2–9.

Doms, M. and Morin, N. (2004). Consumer Sentiment, the Economy, and the News Media. *Finance and Economics Discussion Series (FEDS)*, N 51.

Galbraith J. and Tkacz, G. (2015). *Nowcasting GDP with Electronic Payments Data.* ECB Statistics Paper Series, N 10.

Goloshchapova, I. and Andreev, M. (2017). Measuring Inflation Expectations of the Russian Population with the Help of Machine Learning. *Voprosy Ekonomiki*, N 6, pp.71–93. [in Russian].

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: *Machine Learning: ECML-98*. Springer, pp. 137–142.

Kapetanios, G. and Papailias, F. (2018). *Big Data & Macroeconomic Nowcasting: Methodological Review.* ESCoE Discussion Paper, N 12. Available at: https://www.escoe.ac.uk [accessed on 07 October 2018].

Kholodilin, K., Thomas, T. and Ulbricht, D. (2017). Do Media Data Help to Predict German Industrial Production? *Journal of Forecasting*, 36(5).

Lau, J.H., Newman, D., Karimi, S. and Baldwin, T. (2010). Best Topic Word Selection for Topic Labelling. In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10).* Stroudsburg, PA: Association for Computational Linguistics, pp. 605–613.

Nyman, R., Ormerod, P., Smith, R. and Tuckett, D. (2014). Big Data and Economic Forecasting: A Top-Down Approach Using Directed Algorithmic Text Analysis. *ECB Workshop on Big Data for Forecasting and statistics*. Available at: https://www.ecb.europa.eu/events/pdf/conferences/140407/TuckettOrmerod_BigDataAndEconomicForecastingATop-DownApproachUsingDirectedAlgorithmicTextAnalysis.pdf [Accessed 07 October 2018].

Orlov, A. and Lutsenko, E. (2016). Methods of Reducing Space Dimension of Statistical Data. *Nauchnui zhurnal KubGAU*, 119(05). [in Russian].

Sassano, M. (2003). Virtual Examples for Text Classification with Support Vector Machine. In: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP '03).* Stroudsburg, PA: Association for Computational Linguistics, pp. 208–215. doi: 10.3115/1119355.1119382.

Segalovich, I. (2003). A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. In: *Proceedings of the international conference on machine learning; models, technologies and applications (MLMTA'03).* CSREA Press, pp. 273–280.

Shapiro, A., Sudhoh, M. and Wilson, D. (2017). *Measuring News Sentiment*. Federal Reserve Bank of San Francisco Working Paper Series, 1.

Thorsrud, A. (2016). *Words are the New Numbers: A Newsy Coincident Index of Business Cycles*. Norges Bank Research Working Paper, 21.

Voronina, I. and Goncharov, V. (2015). Sentiment Analysis in Social Networks. Computational linguistics and natural language processing. *Vestnik VGU, Series: System Analysis and Information Technologies, 4.* [in Russian].

Vorontsov, K. and Potapenko, A. (2013). Modifications of EM-Algorithm for Probabilistic Topic Modeling. *Machine learning and data analysis*, 1(6), pp. 657–686. [in Russian].