Bank of Russia
The Central Bank of the Russian Federation

# Texts of Economic News: a Useful Addition to Official Statistics?

Analytical note of the
Research and Forecasting
Department
Bank of Russia

November 2018

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

All rights reserved. This note expresses the authors' personal opinions and does not necessarily reflect the Bank of Russia's official position. The Bank of Russia shall not be responsible for the contents of this note. No reproduction of information provided is permitted without the authors' consent.

## *ABSTRACT*

*The dynamically developing contemporary world and continuous rapid expansion of incoming information predetermines the steadily growing interest in economic studies, which are based on a massive set of unstructured information. It is obvious that the amount of this information is many times that of structured data that economists are more at home with and presents more difficulties from the perspective of technical processing. This, however, opens a radically new range of opportunities to researchers.*

*With economists paying greater attention to the analysis and processing of unstructured information, scientific interest is rising in the analysis of all kinds of documents using computer text processing.*

*Building on the earlier published methodology of constructing Russia's index of economic activity based on the texts of economic news, we attempt to briefly outline the key features of text mining analysis , as well as to provide an answer to the question regarding the scope of potential practical application of this index.*

*The analysis that we have conducted has shown that news stories from internet sources are capable of predicting key short-term business activity trends quite accurately. Moreover, the news index can be regarded as a meaningful and self-sufficient indicator of the economy's condition, which can promptly capture useful information that official statistics and survey data fail to provide. The proposed methodology of constructing the news index can also be used to develop other analytical indicators instrumental in analyzing the current economic situation and, among other things, be useful for a central bank to conduct monetary policy.*

Text analysis is gaining increasing popularity in scientific and professional communities. The rationale for this is that all kinds of information on facts, observations, and developments is mainly provided in an unstructured form.[1] Based on some studies, companies keep about 80%–90% of information they have in an unstructured form, and the body of this information is rapidly expanding.[2]

Amid these developments, scientific interest is growing with respect to analyzing a great variety of textual documents using computer text processing, which has lately produced a large number of studies in this area. For example, in *marketing*, customers' textual comments are studied. In the *financial area*, texts from financial news and social media are used to forecast asset price movements. *In macroeconomics*, texts are analyzed to forecast the fluctuations of inflation, economic growth, and unemployment. That said, text mining in the economic area is a field which leaves ample room for further research.

Text mining allows handling a large variety of tasks, including classification, clusterization, extraction of opinions and facts, construction of expert and question-answer systems, keyword searching, etc. Economists may find additional information from unstructured bodies of data useful for studying the properties of various economic indicators and

[1] Unstructured information is information which does not have a predetermined data structure or is not organized in line with established procedure.

[2] Kambies T., Roma P. Dark analytics: illuminating opportunities hidden within unstructured data. Tech Trends 2017. URL: https://www2.deloitte.com/insights/us/en/focus/tech-trends/2017/dark-data-analyzing-unstructured-data.html (дата обращения – 02.10.2018).

phenomena in more depth, as well as forecasting these. Detailed analysis is also important to central banks and their monetary policy (Bholat D., Hansen S., 2015).

Many countries' central banks have already introduced text mining methods in their research. The main sources of this research are news stories, social media, various reports, statements by officials, and other information.

Specific recent examples include the Bank of England, which has used text mining to reveal the labor market's heterogeneity and its impact on output and productivity. In particular, in one of recent studies, Turrell A. and Speigner B., (2018) used online vacancies taken from a job search engine and estimated to what extent output and labor productivity would rise if mismatches between supply and demand across occupations and regions were to be eliminated. The Bank of Canada conducted a study to see how the regulator's official statements affected changes in yields and volatility of short- and long-term interest rates (Hendry S., Madeley A., 2010). Some economic studies dealing with text mining offer models for nowcasting various economic indicators. These studies primarily contributed to an improvement in the quality of short-term estimates and forecasts of observable macroeconomic indicators using news sources (Ardia D., Bluteau K., 2017; Doms M., Morin N., 2004; Nyman R., Ormerod P., 2014; Shapiro A., Sudhoh M., 2017; Bloom N., Baker S., Davis S., 2016). For example, using texts from daily business newspapers, the Bank of Norway constructed a business cycle index which on a daily basis assesses GDP performance for a given quarter (Thorsrud A., 2016).

As a matter of fact, one almost every week or so may come across research publications or at least references to the results of text mining in analytical reports of both research institutions and central banks.

The text mining model proposed in our study[3] allows obtaining on a daily basis an estimate of the monthly composite business activity index (PMI[4]) using economic news stories. Moreover, the construction of high-frequency (daily) index can, based on short-term data, timely monitor important trends calling for a prompt policy response.

Textual information we used was daily economic news. News stories were chosen because they reflect all developments within the country and abroad, affecting the sentiment and behavior of economic agents taking economic decisions.

The key criteria for choosing a news source were:

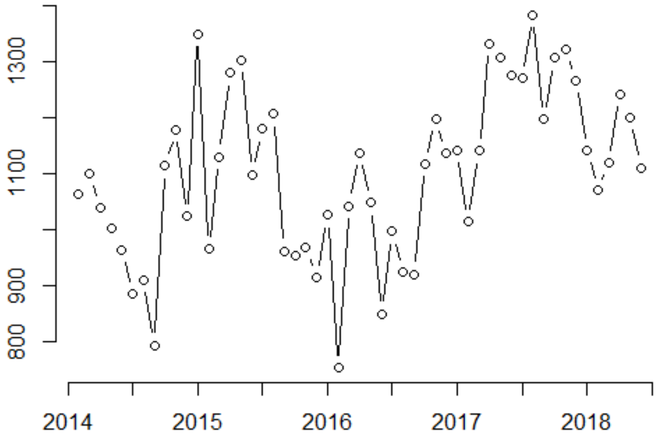
1. news is supposed to be concerned with economic issues;
2. a large body of news data should be available on the internet covering a sufficiently long period (at least 3–4 years);
3. web scraping[5] should be simple, i.e. information should be easy and fast to extract from the website.

[3] For more details see Yakovleva K. (2017) Text mining-based estimation of economic activity // Bank of Russia Working Paper Series, No 25.
[4] Purchasing Managers Index is a macroeconomic indicator of business activity in the manufacturing and services sectors calculated based on managers' surveys.
[5] Web scraping is a technology of extracting data from websites.

Given the above criteria, we chose news resource *vestifinance.ru* solely devoted to economic developments in Russia and abroad. We should make a point that this study could be further developed by, among other things, increasing the number of new sources used in text mining.

In the process constructing the Bank of Russia news index, over 60 thousand stories from the economic news resource for the 2014–2018 period were used. The number of stories varies from month to month, averaging roughly 1,100 (Figure 1 below). We collected news stories spanning the period from 2014 to present.

**Figure 1. The number of news stories per month**



*Source: internet media news, Research & Forecasting Department estimates.*

News stories are broken down into topics for further analysis. Each of these news stories reflects one or several topics to a varying degree. To reveal this, topic modeling was used, enabling all the news stories to be sorted out by topic using computer processing.

The baseline topic models are Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). Each has its advantages and disadvantages, which are, above all, related to the features of training on historical data and those of linguistic processing (Vorontsov, K., 2013). We will not dwell on these methodological aspects in detail but will only note that at the current stage of the study aiming to construct and analyze the news index we used LDA as a probabilistic model the most widely used in similar studies for other countries.

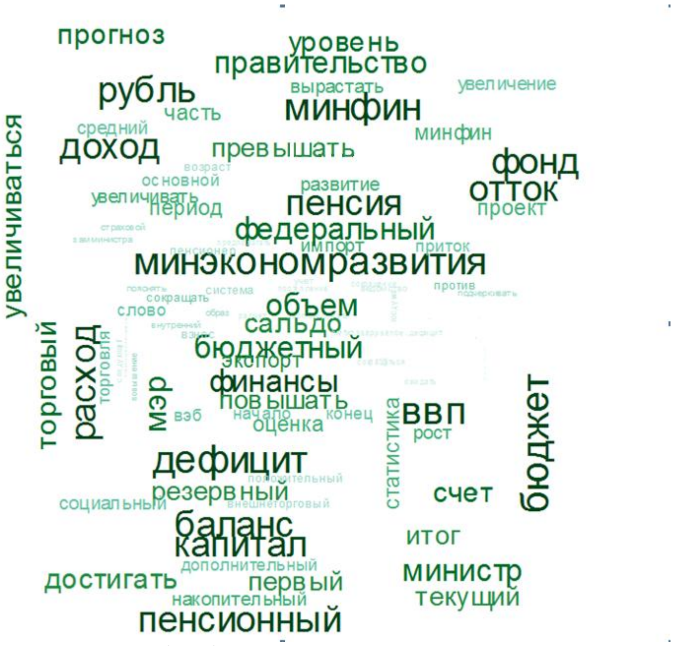
The LDA model produced a list of topics identified in news stories, each represented by the most characteristic individual words (*unigrams*) (Blei D., 2003).

The analysis that we conducted showed that the optimal number of topics for all news data collected was 50. Where there are more topics, they tend to overlap and to be duplicated, where there are fewer topics, several tend to be merged into one. We applied the LDA model to the preprocessed *corpus*.[6]

The LDA model produced a list of words (unigrams), relevant to all of the 50 topics.

Figures 2 below presents 100 unigrams for one of the topics. It is not necessary to scan all the words to understand what topic they refer to. According to Kholodilin K. and Thomas T. (2017), the first 10 words contain 30% of information about the topic, which is adequate for fully understanding what about the text is all about. Hence the most conspicuous words immediately suggest that the *unigrams* refer to the fiscal system.

**Figure 2. Unigrams in one of the topics identified through LDA topic modeling**



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

*Source: Internet media news, Research & Forecasting Department estimates*

The list of the main topics which we identified using the LDA model is presented in Table 1 below. *Topic* 1 includes words such as “bond”, “loan”, “yield”, suggesting the debt securities market. Words occurring in *Topic 2* the most frequently are “dollar”, “ruble”, “the exchange rate”, indicating a topic dealing with the exchange rate. The model therefore only indicates keywords (*unigrams*) based on which we should “name” each of the listed topics on our own.

[6] A corpus is a set of textual documents. A preprocessed 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 noise in the documents.

**Table 1**

**Results of topic modeling for news data**

| Topic | Keywords (unigrams) |
|---|---|
| Topic 1 | Bond, loan, yield, security, amount |
| 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, production, 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 | Deposit, oil field, project, Rosneft, production |
| Тема 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 |
| Topic 16 | Interest rate, Federal Reserve, policy, bank |

*Source: Internet media news, Research & Forecasting Department estimates*

Among the words referring to the above topics are those whose connotations are so broad that their specific meaning only emerges in a certain context. This can be seen in *Topic* 10, which contains *unigram* “establishment”. Its connotation is fairly broad and one can come across it in several topics. But the likelihood of this word appearing in other topics is much lower, hence the word “establishment” is characteristic of Topic 10 rather than any other of them.

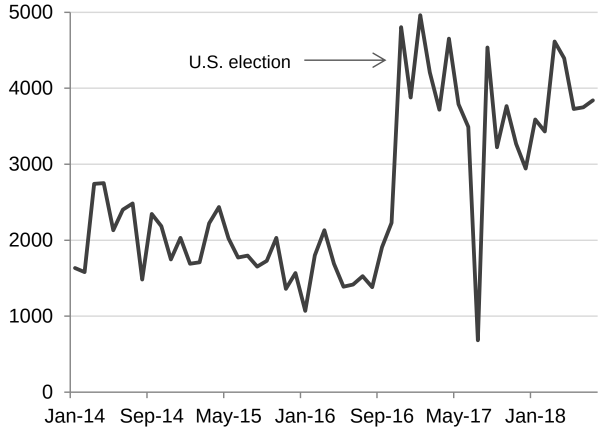
It is important to note that topics represented in news stories change over time. Their changes can be interpreted from two perspectives. *First*, some topics are replaced by other ones. *Second*, their intensity changes, reflecting media’s degree of interest in them.

As for the replacement of some topics by others, economic topics are fairly general in nature (for example, exchange rate, labor market, inflation, and others) and they therefore do not change fast. We will briefly outline this aspect in the final part of the note, which discusses the results of forecasting business activity in the Russian economy using the news index. Only intensity, i.e. media’s degree of interest, will change in the topics under discussion.

To illustrate how intensity can change, we will look at topics *The UK, The USA, and The Exchange Rate* relying on a simple calculation of word occurrence frequency in news stories.

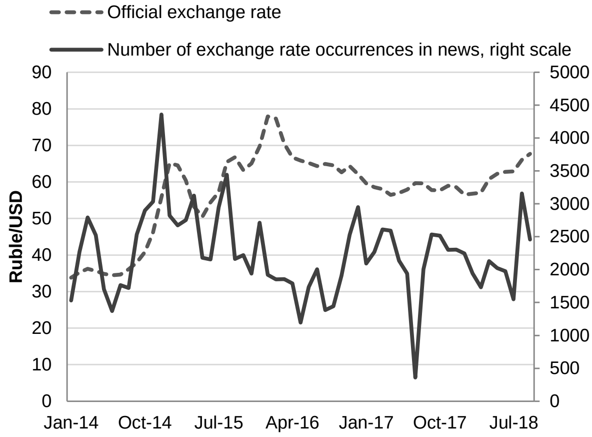
The intensity of the topics drops or rises appreciably over the period from January 2014 to July 2018, depending on a particular time stretch involved. One can see a surge in the number of news stories for topic *The USA* (Figure 3), more than doubling because of the November 2016 presidential election in this country. As regards topic *The UK* (Figure 5), it is very well seen in what periods the UK was in the media’s spotlight. June 2016 accounts for the largest number of news stories due to the UK European Union membership referendum held on June 23, 2016.

### Figure 3. Topic *USA*



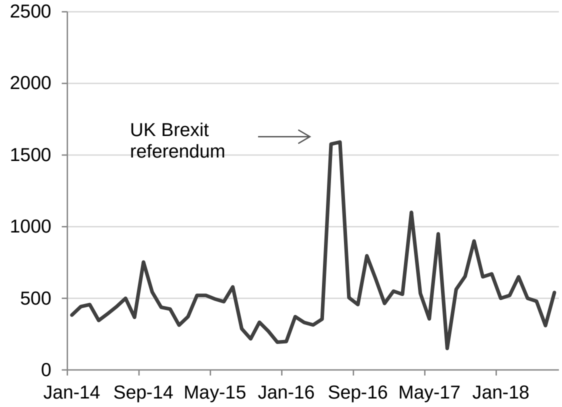
*Source: internet media news, Research & Forecasting Department estimates*

### Figure 4. Topic *The Exchange Rate*



*Source: internet media news, Research & Forecasting Department estimates*

### Figure 5. Topic *The UK*



*Source: internet media news, Research & Forecasting Department estimates*

Meanwhile the effect of news stories on Russia's economy varies. The UK referendum and rising geopolitical risks are example of news stories within the same topic which can differ in the strength of their effect on Russia. To take account of this aspect in text mining, news stories are adjusted for their tone. Specifically, in calculating the news index, we assigned a value of 1 to a news item if it positively affected Russia and -1 if the effect was negative.

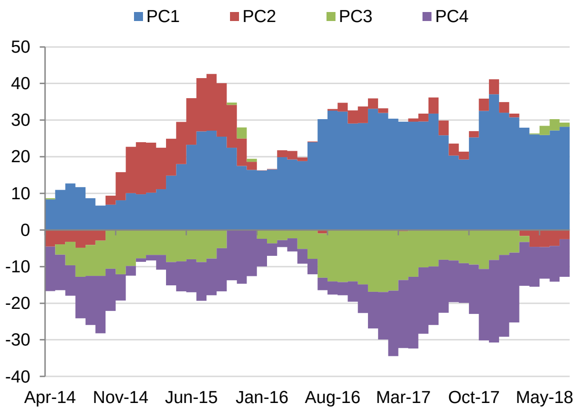
Topic *The Exchange Rate* (Figure 4) features *unigrams* such as "dollar", "ruble", "exchange rate", "euro", "currency", so this topic will not adequately reflect the movements of a particular currency's official exchange rate but will provide a general picture of the situation on the forex market. One can see, for example, that in late 2014 changes in the number of news stories correlated with the movements in the Bank of Russia's official exchange rate of the ruble. The explanation is that ruble depreciation amid the high geopolitical uncertainty and the oil price drop gave rise to a large number of relevant media publications. The subsequent bouts of ruble weakening as the economy gradually adapted to the worsening economic conditions, with economic agents responding to exchange rate movements more moderately, received less extensive media coverage

It is important to note however that text classification through unigrams can also be inaccurate because the same words can have different connotations. We, therefore, emphasize the importance of research aiming to use other methods for constructing a topic model. One example is the use of *bigrams*, i.e., word combinations comprised of two words, such as “exchange rate”, “debt security”. But the use of *bigrams* may in some situations make this work more difficult AND/OR produce a worse result instead. An attempt would then be worthwhile to find a compromise model which would allow an accurate interpretation of the text’s meaning without making the model more complicated.

Under the procedure used to construct our news index, topics obtained through topic modeling are transformed to time series by means of the method of Principal Component Analysis (PCA[7]). The data conversion mainly seeks to compress it by eliminating its redundancy while minimizing information losses.

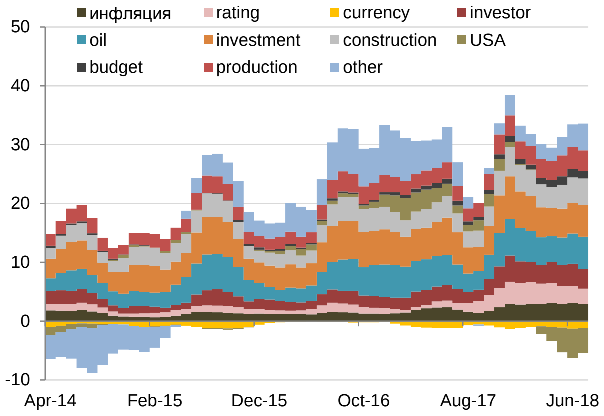
A regression analysis suggests that statistically significant are four principal components, whose contribution to the news index is presented in Figure 6 below. The four principal components explain 83% of the overall variance.

**Figure 6. Decomposition of the news index into principal components**



*Source: media news stories, Research & Forecasting Department estimates*

**Figure 7. Decomposition of the first principal component into topics**



*Source: media news stories, Research & Forecasting Department estimates*

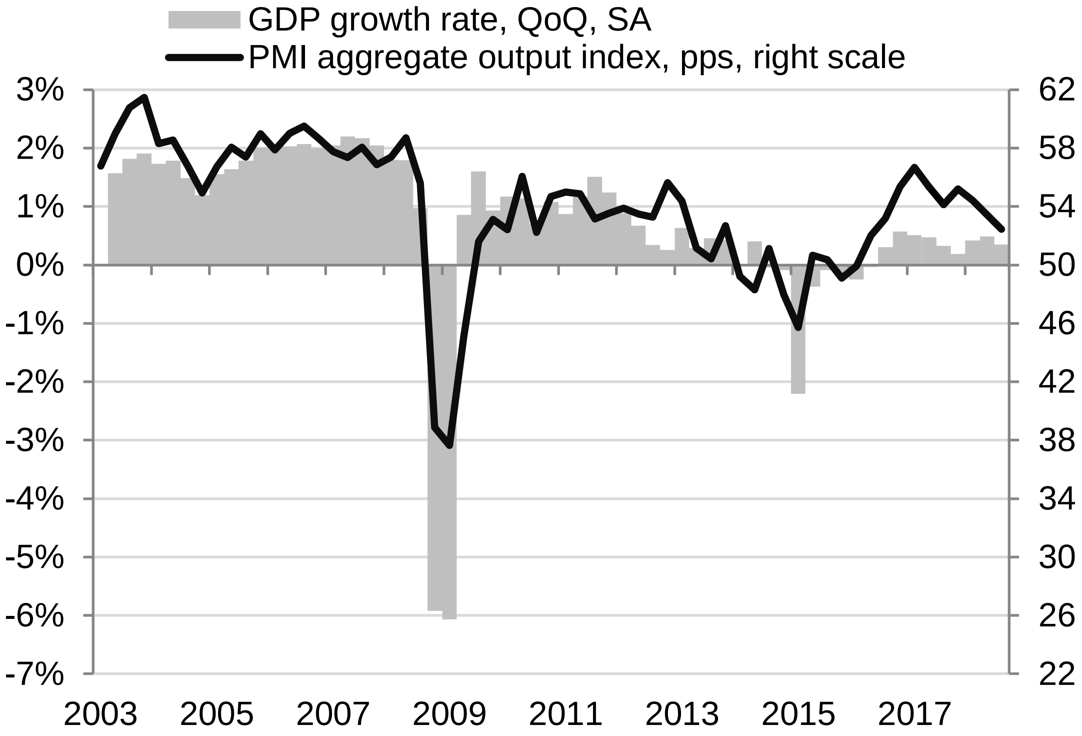
In decomposing the first principal component into topics, which, by definition, explains the largest share of the variance (up to 40%–50%), one can see that its key drivers are the words “oil”, “production”, “investment”, and “ exchange rate” (Figure 7). Overall, the first component’s performance remains positive, which is good from the perspective of economic activity. In other words, other things equal, the more frequently the above words, which are the drivers of almost half of the news index, occur in news stories, the higher our assessment of business activity based on economic news.

[7] Principal Component Analysis is one of factor analysis methods aiming to present data in reduced dimensionality, thereby minimizing a loss of information. Its algorithm provides for going from initial data to several new groups in which data has similar relationships.

*So how useful is the news index constructed in analyzing and forecasting short-term business activity changes in Russia?*

To answer this question, we used all of the 50 topics which the LDA model identified and attempted to match this database with the observable business activity benchmark – that of the composite PMI business activity index.[8] The PMI index was chosen as a criterion of news index quality because of its strong correlation with the key economic indicator, GDP (Figure 8). GDP was not used as a forecasting indicator because GDP data is published on a quarterly basis, i.e. not often and with a considerable lag. We chose the period from February 2017 to September 2018 as the reference point, with the rest of the months (From April 2014 to January 2018) used as the period of training observations.

**Figure 8. GDP and composite PMI index in Russia**



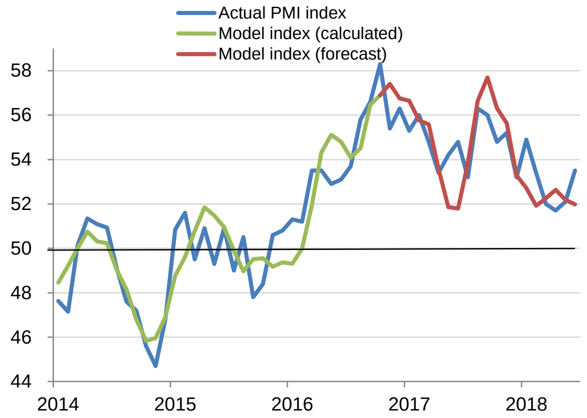
*Source: Rosstat, HIS Markit.*

The results illustrating the accuracy of including data in the composite PMI index based on the news index constructed are presented graphically in Figure 9 and Figure 10 below.

[8] 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; time 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.

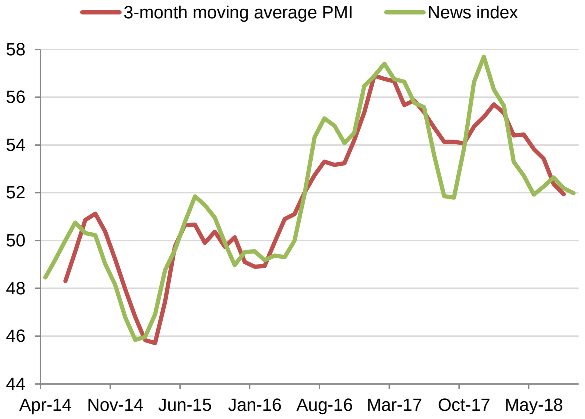
.

**Figure 9. PMI index and News index**



*Source: IHS Markit, internet media news, Research & Forecasting Department estimates.*

**Figure 10. News index and PMI business activity index (three-month moving average)**



*Source: IHS Markit, internet media news, Research & Forecasting Department estimates.*

The PMI index forecast shows that the text mining-based model has a satisfactory predictive power. It can be seen that topics which we have identified often capture general trends but in some periods may fail to reflect some turning points in economic performance or reflect them not accurately enough.

It should be noted that the news index which we have constructed shows a more stable performance versus actual PMI index changes: estimates calculated from business surveys may change from month to month more substantially than the rhetoric of economic news, which is not as volatile. As a consequence, we are prepared to see significant temporary discrepancies between the movements of the news index and the composite PMI index over a horizon of one to two months. But over slightly longer horizons, the news index continues to consistently capture trends in changes of overall assessment of business conditions suggested by the above surveys.

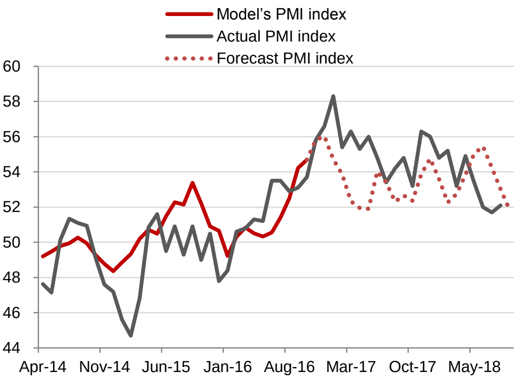
As regards specific assessment of the current business activity level based on economic news data, we note that the value of the news index which we have constructed has stood at around 52 since May 2018. This suggests the maintenance of steady, albeit moderate, economic growth but indicates the risks of its marginal deceleration in the second half of 2018. That said, most of the news topics showed a decline in positive dynamics. Our estimates suggest that the number of positive news stories decreased, above all, in topics related to inflation, the debt market, and the banking sector. This decline was, however, largely compensated by a volatility decline in financial markets. This observation emphasizes the fact that situation normalization in financial markets is an important stabilizing factor indicative of an improvement in the economic environment. The Bank of Russia's Research and Forecasting Department arrived at a similar conclusion in

May 2016 based on surveys of industrial and agricultural businesses conducted jointly with the Institute for Economic Policy.[9]

It should also be noted that the index is at times subject to downward pressure from the topic related to the USA and geopolitics. We, however, emphasize that, based on our estimates, the impact of this topic on the news index is temporary and does not have a sustainable negative effect on our real-time assessment of business activity in the Russian economy.

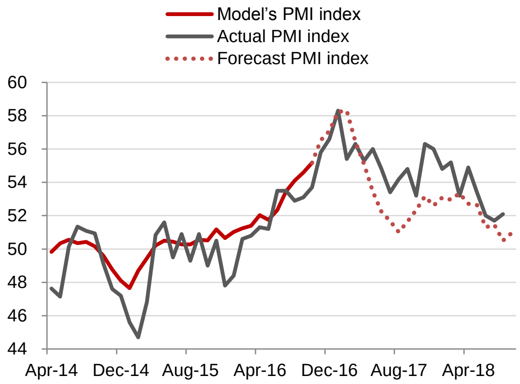
It is also noteworthy that we have found several other predictive properties of the news index. Thus, the index on average shows the highest predictive power in the second week of a month (Figure 12). One explanation is that managers (based on whose answers the PMI business activity index is constructed) usually start at the end of the second – the beginning of the third week of a calendar month. Therefore, the opinions of the surveyed managers are to a greater extent affected by developments occurring in the middle of the month.

**Figure 11. Index constructed on data related to the *first* week of each month**



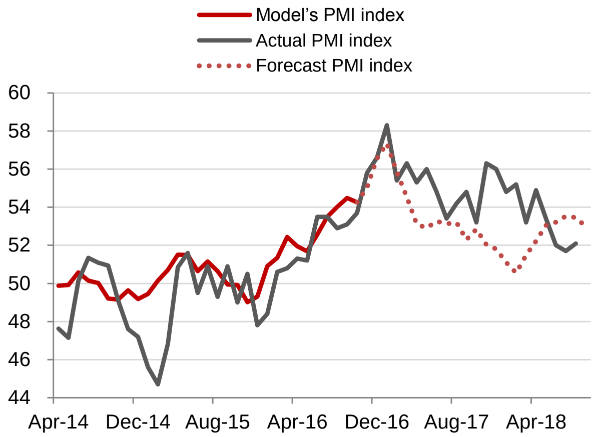
*Source: Internet media news, Research & Forecasting Department estimates.*

**Figure 10. Index constructed on data related to the *second* week of each month**



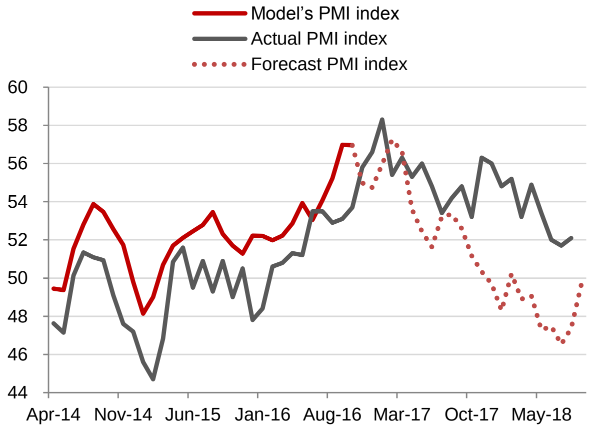
*Source: Internet media news, Research & Forecasting Department estimates.*

**Figure 11. Index constructed on data related to the *third* week of each month**



*Source: internet media news, Research & Forecasting Research and Forecasting Department estimates.*

**Index constructed on data related to the *fourth* week of each month**

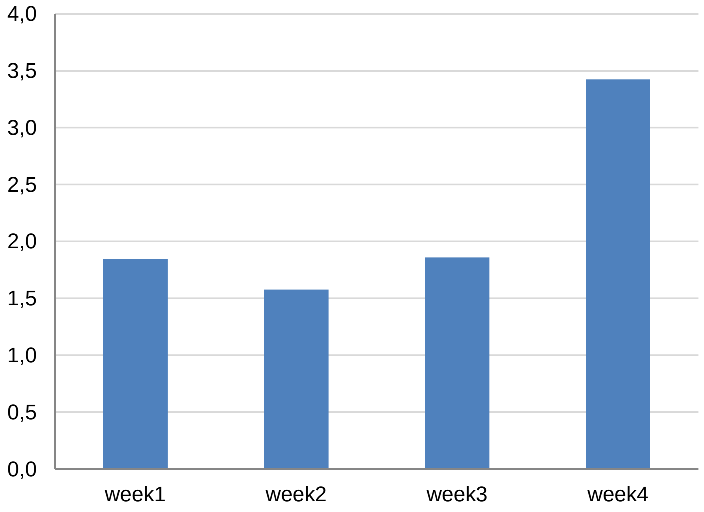


*Source: internet media news, Research & Forecasting Research and Forecasting Department estimates.*

[9] See Talking Trends (No 6, May 2016). In focus. Companies need various ruble exchange rates: survey results.

On average, though, the accuracy of forecasting business activity estimates in a given month does not show an improvement trend as news is accumulated during each following week of a month (Figure 15). The body of economic news for a full month covering a period of one–two weeks after the surveys shows a noticeably worse PMI index forecasting quality.

**Figure 12. Weekly mean absolute errors of the index**



*Source: internet medial news, Research & Forecasting Department estimates*

There may be a number of explanations for that. The key ones are in our view as follows:

1. *Managers' surveys periodicity.* It is after the surveys that developments in the fourth week of a calendar month take place, and, from the perspective of the PMI index data, they will affect respondents' assessment of business activity no earlier than the following calendar month.
2. *Rosstat's short-term statistics are released in the second half of the month.* Comments in economic news stories are published several days after Rosstat's monthly economic activity data is released. Under Rosstat's official schedule, this data is always released in the second half or even towards the end of the month. Moreover, for a number of reasons (including methodological ones), short-term economic statistics traditionally show high volatility: it is often problematic to identify important trends AND/OR turning points in economic activity changes on high-frequency data. In economic media, one can come across unjustifiably negative comments regarding short-term data. For example, a negative tone can be assigned to a comment on a decline in a given month of a short-term indicator *in year-on-year terms* (whereas this may be due to last year's high base) or in *month-on-month terms* without seasonal adjustment (while this decline is in fact part of the usual seasonality of the short-term economic activity indicator). More generally, though, it is important that Rosstat's

short-term data is released already after the managers' surveys and is bound to affect their decisions no earlier than the following survey's month if this development proved to be very important.

3. *Specifics of PMI index construction.* At a month's end, economic media often focus on a limited group of business activity indicators. For example, Rosstat's short-term output statistics are often commented on in economic media with a focus on manufacturing sector output, whereas manufacturing is given a lower weight than the services sector in surveys on which the PMI index is based.

As a result, the behavior of the news index constructed on a database including the last week of a month may run counter to the performance of the PMI index. Nevertheless, the news index properties described above, in our view, mean that the quality criterion of the analytical indicator which we have constructed should under no circumstances be confined to the error of forecasting official statistics' and survey-based indicators, especially over short horizons. What is important is that the news index, as we have already pointed out, captures overall trends in economic activity performance but can also be regarded as a meaningful independent instrument of business activity assessment that has potential to serve as an alternative to existing statistical and survey-based indicators or complement them. A more detailed study of this issue should in our view provide the basis for further research aiming to develop text mining-based indicators of economic conditions.

***

Our analysis suggests that news is in itself capable of sufficiently high-quality description of developments occurring in an economy. There are undoubtedly news topics in the index that we will not always be able to accurately interpret. This is due to both the human factor and the specifics of machine learning methods. Despite the apparent advantages of text mining techniques, it also has its weak points, which are primarily associated with the problems of identifying topics and tones of a textual document and those of text preprocessing texts, especially with respect to Russian words, whose spelling is governed by a lot of rules subject to many exceptions.

Given objective methodological difficulties in dealing with unstructured bodies of data and the ongoing extensive development of text mining techniques, there is certainly ample room for refining and improving approaches to processing texts as part of economic analysis and forecasting. As regards practical conclusions for economists, they so far suggest the need for a balanced and cautious interpretation of text mining results.

The proposed news index produces, already at a first approximation, satisfactory real-time estimates of business activity in the Russian economy . These estimatescan be extensively used in central bank practices. But are text mining methods actually capable of adding a useful component to information about economic conditions contained in official statistics? It is difficult at this point to give a conclusive answer to this question, as it is only further refinement of methods for processing unstructured bodies of data that could bring us closer to it. Hence it would be more appropriate at the current stage to regard the news index which we have constructed as a food for thought, a starting point in a broad complex of economic research dealing with an analysis of large volumes of unstructured information that the Bank of Russia'splans to carry out in the future..

## REFERENCES

1. Vorontsov K. (2013), Probabilistic topic modeling. URL: http://pratsi.opu.ua/app/webroot/articles/1414145257.pdf (as of 01.10.2018).
2. Yakovleva K. (2017). Оценка экономической активности на основе текстового анализа // Серия докладов об экономических исследованиях в Банке России, № 25
3. Ardia D., Bluteau K. (2017). Questioning the News About Economic Growth: Sparse Forecasting Using Thousands of News-Based Sentiment Values. Preprint submittes to SSRN, July 21
4. Bholat D., Hansen S. (2015). Text mining for central banks. Centre for Central Banking Studies
5. Blei D., Ng A., Jordan M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research
6. Bloom N., Baker S., Davis S. (2016). Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics
7. Doms M., Morin N. (2004). Consumer Sentiment, the Economy, and the News Media. Finance and Economics Discussion Series (FEDS)
8. Hendry S, Madeley A. (2010). Text Mining and the Information Content of Bank of Canada Communications. Staff Working Paper 2010–31
9. Kholodilin K., Thomas T., Ulbricht D. (2017). Do media data help to predict German industrial production? Journal of Forecasting
10. Nyman R, Ormerod P. (2014). Big Data and Economic Forecasting: A Top-Down Approach Using Directed Algorithmic Text Analysis
11. Shapiro A., Sudhoh M., Wilson D. (2017). Measuring News Sentiment. Federal Reserve Bank of San Francisco Working Paper Series
12. Thorsrud A. (2016). Words are the new numbers: A newsy coincident index of business cycles. Norges Bank Research. Working Paper
13. Turrell A., Speigner B. (2018). Using job vacancies to understand the effects of labour market mismatch on UK output and productivity. Staff Working Paper. № 737

**Research and Forecasting Department**

Ksenia Yakovleva<br>
Alexey Porshakov