The Use of News Reports to Predict the Values of Macroeconomic Indicators and Indices Represented by Time Series

Artur Mikhailov and Natalia Gergel

[1]*PermNational Research Polytechnic University,Komsomolsky ave. 29, Perm, Russia*
mihailovarthur@rambler.ru, natalia_gergel@mail.ru

Keywords: Forecast, Prediction, Model, Text-mining, Machine Learning, Classification, Time Series.

Abstract: The use of forecasts and predictive models highly affects the process of making decisions. The use of given forecasts allows to increase economic effectiveness of individual entities as well as the corporations. The aim of the article is the investigation of the influence of the weakly formalized factors on the forecasts' accuracy. The study is based on the problem of classification for determining the trends of changing the indicators and the levels of external factors' influences on a change of the referencing parameter. The dataset which contains 25 daily news headings gathered during 8 years was used to make the calculations. The chosen news headlines are related to the stock market and were published by the most authoritative sources such as: Russia Today, Reuters, Scientific American, The Guardian. It was demonstrated that the record of the influence of the information in the news reports on the change of the referencing parameter (using the example of the NASDAQ index) allows clarifying the forecasts taken with the use of functional methods. Therefore, it leads to minimizing mistakes and maximizing the forecasts' reliability.

1 INTRODUCTION

Increasing of forecasting accuracy due to the use of earlier unrecorded factors and the work with large amounts of information are becoming more valuable, because the decisions that are being made directly depend on its quality.

As for now, when working with the macroeconomic indicators, the fact of the influence of scientific discoveries, political changes and the public figures' opinions have to be considered. This type of information mainly comes from the mass media. That is why the use of sources of information where the data is presented as a text deserves a special attention.

The method of machine learning is used in order to work with weakly structured information. When working with the given methods the quality of solutions depends on the quality of the data, amount of information and algorithms that are being used.

The processing of textual information is divided into two stages: preparation and processing of the data.

When preparing the data, some certain questions should be considered: 1) selection of data [1], 2) clearing of data (minimization of noises), 3) choosing the type of textual information for its use in machine algorithms (for example, LDA [2] – Latent Dirichlet Allocation, an approach based on using n-grams [3]), 4) reduce the number of attributes used (frequency algorithms that work with the concepts [4] and terms [5] are used for it), 5) setting correspondence between the dataset and numerical measures (for example, binary feature representation [6], Inverse Document Frequency Method [7]).

In order to process the textual information, the authors are trying to use methods such as: Support Vector Machine (SVM) [8], Regression Algorithms[9], Naïve Bayes [10], Decision Rules or Trees [11], k-NN [12] and [13]. The common factor of these methods is the detection of the relationships between features (which are usually words or phrases), such as input data and target.

2 DATA PREPARATION

For the purposes of this paper the opportunity of using these news reports in order to solve the task of forecasting macroeconomic index using NASDAQ

index as an example has to be considered (the data can be taken from the website https://www.finam.ru).

2.1 Description of a data feature format

For the experiment, a dataset (taken from https://www.kaggle.com/aaron7sun/stocknews), which contains 25 news annotations for each day of the previous 8 years from 08.08.2008 to 01.07.2016 was used (the example of data is in the table 1). The data was selected from the authoritative mass media sources on the topic related to economics. Their news headlines became the data, since in comparison to the full news reports, the information used in these headings is straight to the point.

Table 1: The stricter of the news headings data set.

Date	Heading 1	Heading 2	…	Heading 25
08.08.2008	b"Georgia 'downs two Russian warplanes' as countries move to brink of war"	b'BREAKING: Musharraf to be impeached.'	…	b"No Help for Mexico's Kidnapping Surge"
11.08.2008	b'Why wont America and Nato help us? If they wont help us now, why did we help them in Iraq?'	b'Bush puts foot down on Georgian conflict'	…	b'All signs point to the US encouraging Georgia to invade South Ossetia. Goddamnit Bush.'
12.08.2008	b'Remember that adorable 9-year-old who sang at the opening ceremonies? That was fake, too.'	b"Russia 'ends Georgia operation'"	…	b"BBC NEWS | Asia-Pacific | Extinction 'by man not climate'"
⋮	⋮	⋮	⋱	⋮
01.07.2016	A 117-year-old woman in Mexico City finally received her birth certificate, and died a few hours later. Trinidad Alvarez Lira had waited years for proof that she had been born in 1898.	IMF chief backs Athens as permanent Olympic host	…	Ozone layer hole seems to be healing - US & UK team shows it's shrunk & may slowly recover.

2.2 Preparation of a training sample

In order to solve the task of classification the following classes of interconnection with the values of NASDAQ index need to be emphasized: strong growth, average growth, weak growth, weak decline, average decline, and strong decline. In order to train the model, it is necessary to link every day of the data set to the corresponding classes. In order to do so, the following steps were taken: 1) the difference of NASDAQ values taken from the current and the previous day was calculated, (delta value), 2) a step of growth and a step of decline were calculated in order to determine conditions for each of the classes in a training sample. Step of growth: $S_g = 2/3 \cdot \bar{G}$ (decline $S_d = 2/3 \cdot \bar{D}$), where $\bar{G}$ - is an average value of growth, $\bar{D}$ - is an average value of decline.

The classes were identified based on the following conditions:

- strong growth $\Delta \geq 2 \cdot S_g$ (value +3);
- average growth $2 \cdot S_g > \Delta \geq S_g$ (value +2);
- weak growth $S_g > \Delta \geq 0$ (value +1);
- weak decline $0 > \Delta \geq S_d$ (value -1);
- average decline $S_d > \Delta \geq 2 \cdot S_d$ (value -2);
- strong decline $2 \cdot S_d > \Delta$ (value -3).

As a result, there is a table 2 in which each of the classes corresponds to the following values: strong decline – 241, average decline – 200, weak decline – 431, weak growth – 479, average growth – 316, strong growth – 280.

Table 2: The example of the NASDAQ index classes' table.

Date	Nasdaq (close)	Δ	Class
08.08.2008	2414,1	-	-
11.08.2008	2439,95	25,85	1
12.08.2008	2430,61	-9,34	-1
⋮	⋮	⋮	⋮
01.07.2016	4862,693	19,953	1

2.3 Preparation of a training sample

In order to use the headings of the news reports (that were presented in English) aspects such as: articles,

punctuation marks, numbers and other meaningless words, were excluded. The whole data set was written in a lowercase. These changes were needed in order to conduct private analysis of the whole textual data and the following private analysis of the headlines of each day.

3 DEVELOPMENT OF A MODEL FOR SOLVING THE TASKS OF CLASSIFICATION

3.1 Choosing a machine learning model

Using the inductive approach to the analysis of the results, it can be stated that the models in most of the cases were able to solve the task of classification. The best results were made by the Naïve Bayes model. One of the particular qualities of this method is not being able to work with new features that were not a part of the testing data, which was used to train a model. Based on this, the model that was used in order to solve the task of classification was the Random Forest Method. Its results were better than all of the other methods, except Naïve Bayes.

The models such as Logistic Regression, Naïve Bayes, Random Forest, k-NN were compared in order to choose the method of solving the task of classification. The work of these models was evaluated using different amounts of the data set: the whole data set and partial sample (25% of the whole data set). This investigation showed the behaviour of each of the models of machine learning when using different amounts of data set. It allowed to choose a model of machine learning in order to solve the task of classification. (Table 3)

Table 3: The results of learning and checking the work of methods for the classifying weak growth and average decline of the retrospective data.

Model	Partial sample (25% of the whole dataset)				Full sample (100% of the whole dataset)			
	Probability of the prediction (%)		numerical value (result out of the whole value)		Probability of the prediction (%)		numerical value (result out of the whole value)	
	Chosen classifier	Other classifiers	Chosen classifier	Other classifiers	Chosen classifier	Other classifiers	Chosen classifier	Other classifiers
Weak growth (+1)								
Naïve Bayes	100	92	121/121	337/365	100	97	478/478	1426/1470
Random Forest	44	99	53/121	361/365	69	99	328/478	1452/1470
Logistic Regression	95	99	115/121	364/365	97,5	99	466/478	1469/1470
k-NN	39,5	83,5	48/121	305/365	42,5	84,5	204/478	1244/1470
Average decline (-2)								
Naïve Bayes	100	76	47/47	335/439	100	95	202/202	1655/1746
Random Forest	93,5	100	44/47	439/439	100	100	202/202	1746/1746
Logistic Regression	13	99	6/47	435/439	18	99	37/202	1745/1746
k-NN	15	98	7/47	430/439	22	97	44/202	1692/1746

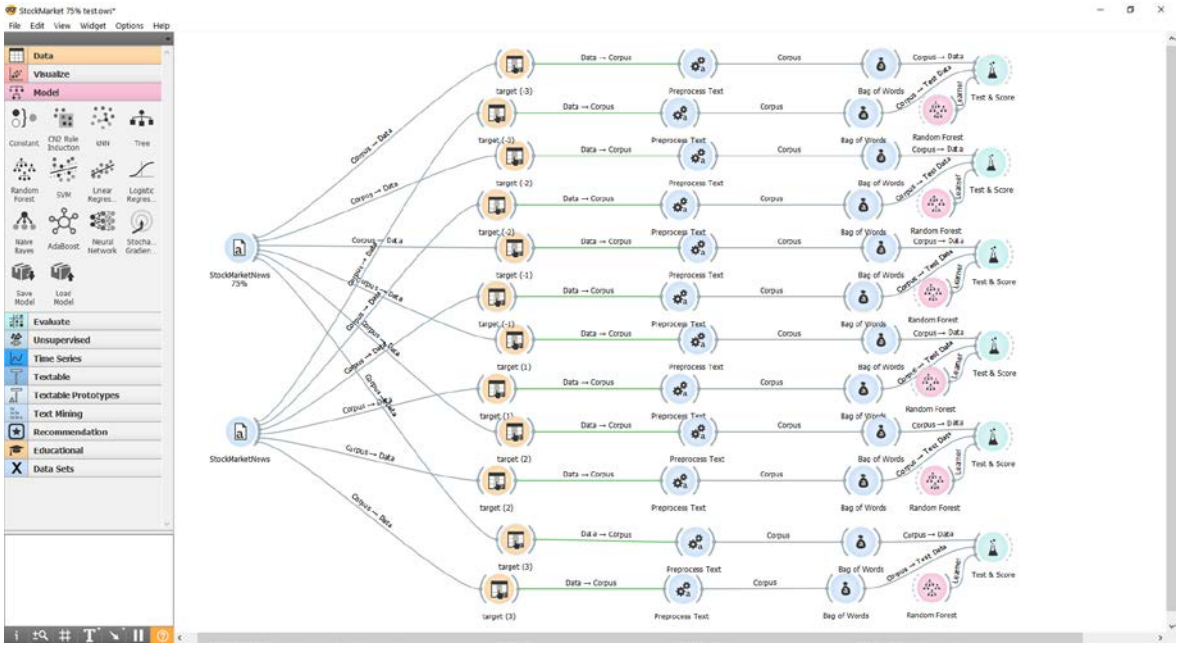

Figure 1: The model of solving the tasks of classification, which was developed in Orange.

3.2 Solving the tasks of classification

As a result, the model (figure 1) that was trained on 75% of data was developed to solve the task of classification. When checking on a retrospective data (for the 6 classes stated above) the method solved the task of classification correctly in 1826 out of 1949 cases (it corresponds to the probability of the correct classification which is more than 90%).

4 USING THE RESULTS OF THE SOLVED TASKS OF CLASSIFICATION TO FORECAST THE TIME SERIES VALUES

The main factor of the time series that describes the stock indices (such as NASDAQ) is the lack of seasonality, periodicity and known sequences. In this case the functional methods of forecasting [14] of the reference parameter do not work (the methods give bad results and do not pass the test of checking the adequacy of significant amount of steps – figure 2).

Our algorithm has to be built based on the following recurrent formula:

NASDAQ (t+1) = NASDAQ (t) + C, where C is the corrective coefficient that depends of the values of the expressions $3 \cdot S_g$, $2 \cdot S_g$, S_g, S_d, $2 \cdot S_d$, $3 \cdot S_d$, S_d and S_g (which are described in the 2nd part) and can be picked up based on the first values of the testing data set (or their parts).

The results were obtained using the recurrent formula (figure 3).

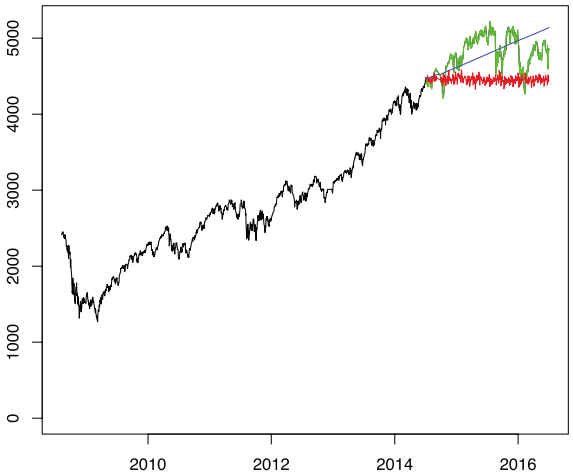

Figure 2: These changes of the NASDAQ index (black line – training sample, green line - testing sample) and the forecasts values that were obtained using the autoregressive method (blue line), fractal method (red line).

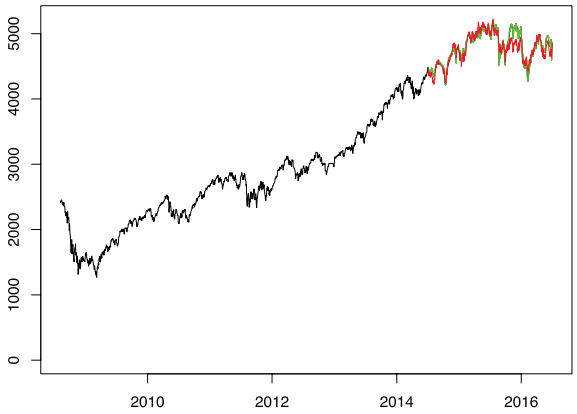

Figure 3: The results of forecasting the NASDAQ index (black line – learning sample, green line – testing sample) with the use of the offered recurrent formula (red line).

The Pearson's chi-squared test provides the best values of adequacy of the obtained result when checking the testing sample, narrowing of the confidence interval and greater forecasting horizon when using the offered recurrent dependence (Table 5).

Table 5: The results of the Pearson's chi-squared test.

Predictive model	The amount of the calculation steps in which the method stays adequate
Autoregressive method (ARIMA)	54
Fractal method	44
The offered recurrent method	338

In this case the offered method allows obtaining better results for a long period of time.

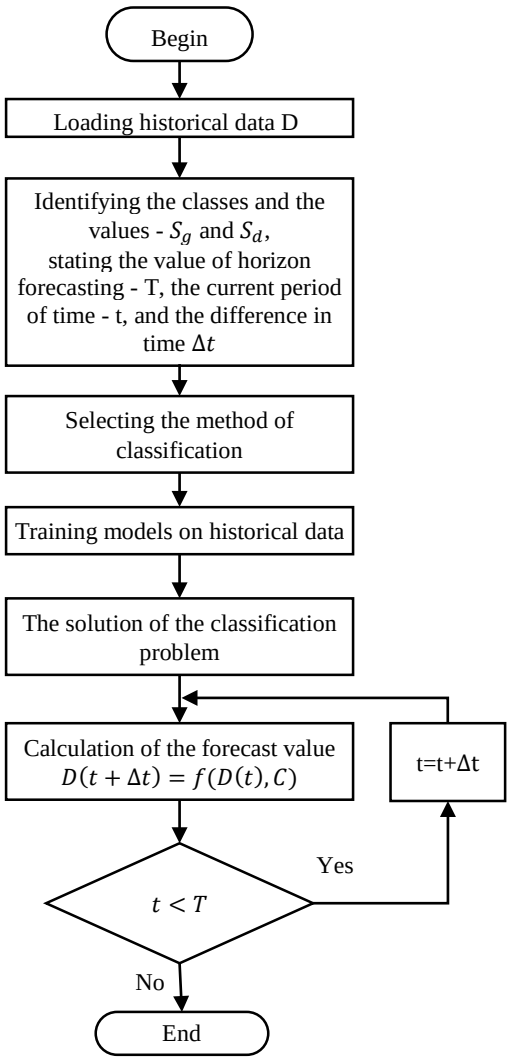

Figure 4: General scheme of using the classifiers in the time series forecasting algorithms.

5 CONCLUSIONS

The results of the investigation showed the weak fitness of the forecasting functional methods used for the data description without any expressed sequences. At the same time, the results showed that the classification data could be used to solve the tasks of forecasting for which the algorithms on the 4th figure should be used.

The perspective of the use of classification methods for solving the tasks of forecasting and the opportunity of developing the forecast based on the use of testing data can be shown in this case.

ACKNOWLEDGMENTS

The authors thank the government of Perm Krai for the support of the project for "Development of software and economic and mathematical models for supporting innovation project management processes in production systems", which is being implemented in accordance with decree №166-п of 06.04.2011.

The reported study was partially supported by the Government of Perm Krai, research project No. C-26/058.

REFERENCES

[1] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7653–7670, Nov. 2014.

[2] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, and N. Ramakrishnan, "Forex-foreteller: currency trend modeling using news articles," 2013, p. 1470.

[3] M. Butler and V. Kešelj, "Financial Forecasting Using Character N-Gram Analysis and Readability Scores of Annual Reports," in *Advances in Artificial Intelligence*, vol. 5549, Y. Gao and N. Japkowicz, Springer Berlin Heidelberg, 2009, pp. 39–51.

[4] Y. Zhai, A. Hsu, and S. K. Halgamuge, "Combining News and Technical Indicators in Daily Stock Price Trends Prediction," in *Advances in Neural Networks – ISNN 2007*, vol. 4493, D. Liu, S. Fei, Z. Hou, H. Zhang, and C. Sun, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 1087–1096.

[5] M.-A. Mittermayer, "Forecasting Intraday stock price trends with text mining techniques," 2004, p. 10 pp.

[6] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, pp. 1–19, Feb. 2009.

[7] X. Zhou and Australasian Database Conference, Eds., *Database technologies 2002: proceedings of the Thirteenth Australasian Database Conference ; Monash University, Melbourne, January/February 2002*. Sydney: Australian Computer Society, 2002.

[8] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, vol. 1398, C. Nédellec and C. Rouveirol, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 137–142.

[9] S. Henrard, N. Speybroeck, and C. Hermans, "Classification and regression tree analysis vs. multivariable linear and logistic regression methods as statistical tools for studying haemophilia," *Haemophilia*, vol. 21, no. 6, pp. 715–722, Nov. 2015.

[10] G. M. Di Nunzio, "Using scatterplots to understand and improve probabilistic models for text categorization and retrieval," *Int. J. Approx. Reason.*, vol. 50, no. 7, pp. 945–956, Jul. 2009.

[11] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[12] Association for Computing Machinery, W. B. Croft, International Conference on Research and Development in Information Retrieval, and Trinity College Dublin, Eds., *SIGIR '94: proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 3 - 6 July 1994, Dublin, Ireland*. London: Springer, 1994.

[13] L. Mylnikov, B. Krause, M. Kuetz, K. Bade, and I. Shmidt, *Intelligent data analysis in the management of production systems (approaches and methods)*. Moscow: BIBLIO-GLOBUS, 2017.

[14] L. A. Mylnikov, A. V. Seledkova, and B. Krause, "Forecasting characteristics of time series to support managerial decision making process in production-And-economic systems," *Proc. 2017 20th IEEE Int. Conf. Soft Comput. Meas. SCM 2017 6 July 2017*, pp. 853–855.