**Application of Natural Language Processing (NLP) Methods to Predict the Realized Volatility of Shares of the Largest Russian Companies**

Shaidullin Ansel Ilgizovich

1-year graduate student, Department of Business Informatics, Graduate School of Management, National Research University Higher School of Economics.

Contacts: [ansel.shajdullin@yandex.ru](mailto:ansel.shajdullin@yandex.ru), [aishajdullin@hse.ru](mailto:aishajdullin@hse.ru), 89264960767

Thematic direction - "Digital Economy"

**ANNOTATION**

The relevance of the chosen topic lies in the fact that in the modern world the role of using various methods is growing, with the help of which it is possible to analyze risks and make forecasts. The purpose of the work is to build an adequate model for assessing and predicting the volatility of the stock market using machine learning methods based on data from social networks.

The first mention of the practical possibility of using behavioral factors in research devoted to the analysis of investor sentiment is found in articles from the early 2000s. Various researchers, including [Daniel, Hirshleifer & Teoh, 2002] and [Tseng, 2006], have reported an increasing amount of empirical evidence showing that the stock market is driven by investor psychology.

The article [Audrino, Sigrist & Ballinari, 2020] uses a methodology to assess the impact of various news on the realized volatility of a number of US companies. The article [Ulyankin, 2020] analyzes the influence of media news sentiment on various macroeconomic indicators. The methodology for researching news sentiment proposed by [Audrino F., Sigrist F., Ballinari D., 2020] has been expanded with the addition of new, more modern methods of researching the sentiment of text (methods of the Random Forest, XGBoost, Light GBM).

The realized volatility of shares of a number of Russian companies that belong to different sectors of the economy is analyzed: from commodity companies - oil, gas (Lukoil, Rosneft, Gazprom, Novatek) to companies engaged in the service sector - the banking sector (Sberbank), IT (Yandex), telecommunications (MTS). As it turned out, different companies (with different focus) react differently to moods and / or news on social networks.

The study measured sentiment and attention signals associated with individual stocks and the stock market index, using text data from the social network Twitter; tested the hypothesis that companies in different sectors may react differently to news and comments on social networks. Investor attention and sentiment have a significant impact on realized volatility, as do comments on social media. For service companies, the impact of social media is more significant. When attention and mood factors are included in the economic model, the forecast quality of the model increases slightly. For Yandex, information about mood and attention in social networks has great explanatory power. And the factors of attention and mood improve the forecast quality on average for all methods by 7% relative to the economic model. Then there are Sberbank and MTS, for which, when choosing the optimal method, the sentiment model improves the forecast quality by 3% in comparison with the economic model. Companies related to oil and gas production react worst to social media factors, and even with the best method, they give an increase in forecast quality of 2% on average. The results of the study allow us to expand our understanding of the possibilities of machine learning for analyzing the influence of the sentiment of messages in social networks on some financial indicators.

Research methodology:

- collection of data on realized volatility, economic and financial indicators, as well as data from the social network Twitter;

- training of the ULMFiT classifier of the sentiment of the text on the automatically marked data;

- construction of five linear and non-linear models - linear regression, lasso regression, random forest, XGBoost model (gradient boosting) and LightGBM (extended gradient boosting);

- training on an in sample (training) sample by the above methods, three models:

• base model (included variables of realized volatility and its derivatives);

• economic model (economic and financial indicators are also included);

• mood model (attention variables are also included).

- forecasting the realized volatility for the day ahead on the out-of-sample (test) sample, analyzing the degree of influence of economic and financial variables, mood and attention variables.

The novelty of the research is as follows: 1) new models have been built, adapted for the Russian financial market; 2) applied new machine learning methods to predict stock volatility based on the sentiment analysis of text on social networks. The work also offered practical recommendations on the choice of the field of application of the models and the results obtained.

Key words: news sentiment, realized volatility, stock market, neural networks.

JEL: G17, G32

**Bibliography**

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. The Journal of Finance, 59(3), 1259–1294.

Audrino F., Sigrist F., Ballinari D. The impact of sentiment and attention measures on stock market volatility //International Journal of Forecasting. – 2020. – Т. 36. – №. 2. – С. 334-357

Corsi, F., Audrino, F. and Reno, R. (2012). HAR Modeling for Realized Volatility Forecasting. In: Handbook of Volatility Models and Their Applications. (pp. 363-382). New Jersey, USA: John Wiley & Sons, Inc. ISBN 9780470872512

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. Review of Financial Studies, 28(1), 1–32.

Daniel, K., Hirshleifer, D., & Teoh, S. H. (2002). Investor psychology in capital markets: evidence and policy implications. Journal of Monetary Economics, 49(1), 139–209.

Tseng, K. C. (2006). Behavioral finance, bounded rationality, neurofinance, and traditional finance. Investment Management and Financial Innovations, 3(4), 7–18.

Ulyankin, F. (2020). Forecasting Russian Macroeconomic Indicators Based on Information from News and Search Queries. Russian Journal of Money and Finance, 79(4), pp. 75–97.