THE EFFECTIVENESS OF FINE-TUNING LARGE LANGUAGE MODELS TO MEASURE TRUST IN THE CENTRAL BANK

Anastasia Matevosova
The Central Bank of the Russian Federation
Institute of Economics, Russian Academy of Sciences
Faculty of Economics, Lomonosov Moscow State University
Moscow, Russia
nastya.m.2003@yandex.ru

ABSTRACT

This article demonstrates the importance of fine-tuning large language models in economic research. The article tests a methodology for assessing trust in the central bank using a sentiment analysis of social network comments by building an indicator of trust in the Bank of Russia. Statistical approaches, classical machine learning algorithms, classical deep learning models, and approaches based on large language models were compared. Fine-tuning RuBERT model was recognized as the most effective approach to solving the task of sentiment analysis within the framework of building the indicator of trust in the Central Bank. The study demonstrates the advantages of domain-specific adaptation in which fine-tuning models can lead to higher quality than using larger generative models.

1 INTRODUCTION

Trust in the central bank is an important factor influencing the effectiveness of central bank monetary policy. The existing methods for measuring trust in the central bank are based on opinion polls Ehrmann M. (2013); Walti (2012). However, these methods are characterized by a number of disadvantages, making it difficult to quantify the dynamics of trust in the Central Bank.

Natural language processing methods have already demonstrated their advantages in other areas of economic and financial research Huang (2023); Kirtac K. (2024). But the effectiveness of applying certain approaches depends both on the specific task and on the specifics of the texts under consideration. Therefore, it is necessary to compare and choose the most effective approaches to solve the problem.

This paper suggests using modern NLP approaches to measure trust in the Central Bank based on a sentiment analysis of thematic comments on a social network. For this purpose, the effectiveness of various methods of sentiment analysis is compared within the framework of building an indicator of trust in the Bank of Russia according to the methodology Matevosova (2025). During the analysis, the following findings were obtained:

1. The method of measuring trust in the central bank by analyzing texts on social networks has shown its advantages in the framework of testing to build an indicator of trust in the Bank of Russia.
2. Despite the relative stability of the methodology, it is necessary to choose the most effective NLP approaches.
3. Large language models are significantly superior in quality to other approaches. The best results are demonstrated by the fine-tuned RuBERT, which surpassed not only the classical models, but also the generative large language models DeepSeek-R1-Distill-Qwen-7B and Yandex GPT.
4. The LoRA fine-tuning on Yandex GPT-4 Lite also demonstrates an effective way to improve the performance of large language models.

2 Related Work

Existing studies measure trust in the Central Bank using a direct question about trust in the framework of opinion polls Ehrmann M. (2013); Walti (2012). However, this method is characterized by a high sensitivity of the results to the formulation of the question and the overall structure of the survey. In addition, the survey-based method does not allow to measure trust in the central bank with a sufficiently high frequency, which significantly limits the possibilities of studying trust in the Central Bank.

Sentiment analysis of social media users' comments makes it possible to build various kinds of economic indicators based on big data, which can later be used to analyze the economic situation and model it. Thus, there are a number of studies on the construction of indicators of inflation expectations of the population based on data from social networks Goloshchapova I. (2017); Angelico (2022). But these works use rather primitive approaches based on rules and key expressions or classical machine learning algorithms.

However, classical deep learning models, and especially large language models, often demonstrate much better quality in the tasks of sentiment analysis of official economic texts Huang (2023); Pfeifer M. (2023).

Huang et al. (2023) used preliminary training of the BERT architecture model on large volumes of unlabeled financial texts and fine-tuning on an expert-marked sample of sentences to develop the FinBERT model for use in the field of financial texts. It has been demonstrated that this model solves the problem of sentiment analysis of financial texts better than Google's original BERT model, and its advantage grows with increasing financial vocabulary in texts. However, learning from scratch your own large language model requires significant computing resources and large amounts of prepared data. Therefore, when working with large language models, the transfer learning approach is common. As part of this approach, the pre-trained model is fine-tuned for a specific task using fine-tuning on a small amount of marked-up text data. The use of transfer learning increases the effectiveness of using large language models in solving specific tasks.

In the article Ardekani A. M. (2024), based on fine-tuning of GPT-3 in relation to the classification of sentences containing references to inflation, the FinSentGPT model was obtained. It has been demonstrated that this model allows for a more flexible solution to the problem of determining financial sentiment, taking into account the context, and is suitable for various languages. In the work Pfeifer M. (2023), a specialized model for analyzing central bank CentralBankRoBERTa communications is also being developed based on fine-tuning of the initial large RoBERTa language model. However, the effectiveness of applying certain approaches depends both on the problem being solved and on the specifics of the texts under consideration.

Thus, Huang et al. (2023) demonstrate the advantages of the BERT model, while Ardekani 2024, Alonso-Robisco and Carbo 2023, Kirtac and Germano 2024 show the effectiveness of decoder-only models.

But NLP methods, and especially large language models, are still rarely used in economic research. There are examples in the literature of using large language models to study the tonality of financial texts Huang (2023), news sentiment Kirtac K. (2024), and analyzing central bank statements in communications Ardekani A. M. (2024); Alonso-Robisco A. (2023). Therefore, the use of NLP technologies seems promising in the framework of measuring trust in the Central Bank.

3 Approach

3.1 Methodology

The research is based on the methodology for constructing an indicator of trust in the central bank based on sentiment analysis, proposed in the article Matevosova (2025).

At the first stage of the methodology, a representative source for collecting comments is selected and a sample of media outlets publishing economic-related posts is formed.

At the second stage, a group of economic news posts and a group of posts mentioning the central bank are selected using a developed set of key expressions. Then comments related to the news posts of these two groups are collected.

As main part of the third stage of the methodology, it is necessary to calculate the tonality of comments through sentiment analysis and average the results obtained for the two groups within the time periods determined by the required frequency of the final indicator. As a result, two time series of the emotional tone of the comments to the formed groups of news posts are obtained.

At the fourth stage, the indicator of trust in the central bank is calculated using Equation 1, as the difference between the sentiment of comments on posts mentioning the Central Bank ($mood_{CB}$) and the sentiment of comments on economic posts ($mood_{EC}$).

$$trust_{CB} = \frac{mood_{CB} - mood_{EC} + 2}{4} \tag{1}$$

The resulting indicator of trust in the Central Bank is measured in the range from 0 to 1.

The methodology is based on the assumption that factors unrelated to confidence in the Central Bank equally shift the average tone of comments related to the discussion of the Central Bank and comments related to the discussion of the economic situation.

3.2 DATA

The methodology has been tested on Russian-language comments to build an indicator of trust in the Bank of Russia. The VK social network is the leader in monthly reach among the Russian population and at the same time does not have a pronounced shift in the audience profile. A program was written to collect data on posts and comments on the Russian social network using VK API methods. Based on the compiled set of key expressions, two groups of posts were selected: posts mentioning the central bank and posts on economic topics in general, to which comments were collected.

Next, a manual markup of 10,000 randomly selected comments is performed. The markup was carried out by a person with higher education based on reading the text of each comment from this sample and determining the emotional color. Then another person with a higher economic education checked the correctness of the markup, the controversial points were discussed and the annotators came to a consensus. The labeling task consisted in assigning the text of the comment to one of three classes: positive, neutral or negative. It is difficult to divide comments into more than 3 classes of emotional coloring within the markup. Usually, researchers consider two classes (negative and positive) Pfeifer M. (2023); Alonso-Robisco A. (2023) or three classes (negative, neutral, positive) Huang (2023); Ardekani A. M. (2024). However, not all comments have a pronounced emotional tone, so it is advisable to consider 3 classes, including neutral ones. As a result, the comment texts were classified into one of 3 classes: negative (50.32%), neutral (31%), and positive (18.68%).

The average comment length is 108.3, the median is 66 characters. The difficulty of analyzing this dataset is that the texts have an economic theme and a conversational communication style typical of social networks. Therefore, the texts have an economic and news context that needs to be taken into account, while at the same time the texts are emotional and relatively brief.

The marked-up data are divided in the proportion (70 / 20 / 10) into training, validation, and test samples. Training and validation samples are used to train models from scratch or to fine-tune large language models.

Using a set of metrics: F1-macro, accuracy and Quadratic Weighted Kappa, the quality of the final models is compared on the test sample. The Quadratic Weighted Kappa was chosen as the main metric because it takes into account the order of classes, penalizes large differences more strongly, and also normalizes the degree of agreement relative to random. The code and data are provided in the repository[1].

[1] https://github.com/AMatevosova/Indicator-of-public-confidence-in-the-Central-Bank-based-on-sentiment-analysis

3.3 MODELS

The approaches under consideration include statistical methods (MultinomialNB) and classical machine learning algorithms (Logistic Regression, Random Forest, XGBoost) applied to embeddings based on TF-IDF with unigrams, classical neural network architectures (CNN, GRU, LSTM) with trainable embeddings. The modern approaches based on large language models were also considered: encoder-only model RuBERT, decoder-only models DeepSeek-R1-Distill-Qwen-7B and Yandex GPT.

BERT-based models (Devlin et al., 2019) are standard approaches among large language models for solving NLU tasks, in particular sentiment analysis. The pretrained RuBERT model (Zmitrovich et al., 2023) was fine-tuned on 1,000 and 7,000 comments.

The DeepSeek-R1 model specializes in logical inference, which can potentially lead to high-quality analysis of text tonality, supported by chains of logical reasoning. Therefore, it was decided to consider the DeepSeek-R1-Distill-Qwen-7B model. The zero-shot approach is used, in which the model is accessed according to the promt without providing examples in it: *"What is the sentiment expressed in the following comment?" Select sentiment value from positive, negative, or neutral. Return only the sentiment value in small letters"*. Despite the fact that the text of the comments was submitted in Russian, the text of the prompt itself was in English, as this significantly reduced the number of hallucinations.

Interaction with the classifier based on the Yandex GPT Pro model was carried out using a special prompt. The model itself is mainly focused on working with text in Russian, but it also supports English. A fine–tuning of the lightweight version of the YandexGPT 4 - YandexGPT 4 Lite model was performed using the Low-Rank Adaptation (LoRA) Hu et al. (2021) method for the multiclass classification task under consideration.

3.4 COMPARISON

The results of the comparison of models by classification quality are shown in Figure 1.

Neural networks are superior in quality to classical machine learning algorithms. Among the classical machine learning algorithms, XGBoost showed the best result, which is comparable to classical neural network architectures. GRU showed a slightly better result than LSTM and CNN, which showed approximately similar results. This is due to the peculiarities of the texts under consideration: their short length and insufficiently large training sample size for learning from scratch.

The pre-trained large language models demonstrate quality significantly better than models trained from scratch. The most effective approach is to use a fine-tuned on 7,000 comments RuBERT model[2]. It surpasses approaches based on generative large language models in terms of quality measured on the basis of QWK metric. Even using an incomplete training sample of 1,000 marked-up comments as part of fine-tuning RuBERT allows you to achieve good quality and significantly improve the result.

However, fine-tuning and applying this model takes little time, and even a small set of marked-up data is sufficient for high quality. Fine-tuning RuBERT on a full training sample (7,000 comments) using a single T4 GPU took about 6 minutes. It takes less than 20 seconds to apply the model on a test dataset of 2,000 comments to make predictions. While the time to apply the model to make predictions using the classifier based on YandexGPT Pro for 2,000 comments was 57 minutes. Thus, fine-tuning RuBERT provides a significant improvement in quality compared to models trained from scratch, and does not require significant computing resources and time.

DeepSeek-R1-Distil-Qwen-7B showed quality lower than the classifier based on Yandex GPT Pro. However, unlike Yandex GPT Pro, Russian is not a priority for the DeepSeek-R1 model, and a distilled version of the model was considered, rather than the full model, as is the case with YandexGPT Pro. Taking these observations into account, DeepSeek-R1-Distil-Qwen-7B showed good results compared to Yandex GPT Pro. The quality of the sentiment analysis solution using the fine-tuned lightweight version of Yandex GPT 4 Lite (LoRA YandexGPT 4 Lite) is better than the quality obtained using YandexGPT Pro, which confirms the effectiveness of fine-tuning decoder-only mod-

[2] `https://huggingface.co/AnastasiaMML/EconComments-RuBERT-sentiment`

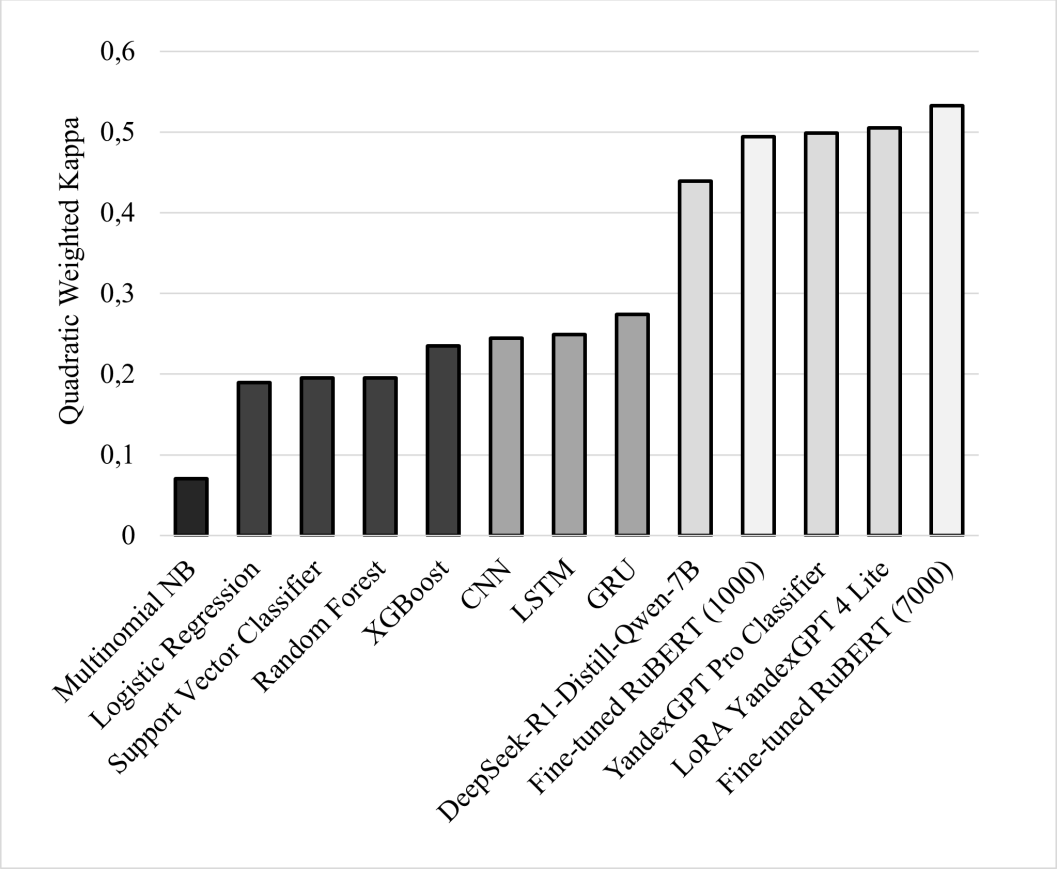

Figure 1: Comparison of different quality models based on the metric of Quadratic Weighted Kappa.

els. The time spent fine-tuning YandexGPT 4 Lite using rsLoRA was 22 minutes, which confirms the effectiveness of this method in terms of significantly improving quality while saving resources.

4 Results of the sentiment analysis

The task of classifying by tonality is quite difficult, as it requires an understanding of the news and economic context typical of Russia.

The results of the sentiment analysis for all collected comments are used to build an indicator of trust in the central bank according to the methodology. In order to verify the relevance of the constructed indicator of trust in the Bank of Russia, its dynamics are compared with the dynamics of the net trust indicator calculated based on the results of the response to the question of trust in the Bank of Russia in the framework of the FOM survey. Net trust is calculated using the method proposed in Ehrmann M. (2013) as the proportion of respondents who answered ”trust” minus the proportion of respondents who answered ”do not trust”. On an annual basis, the correlation of the trust indicator based on sentiment analysis using the fine-tuned RuBERT model with the survey-based indicator of net trust is 0.71.

The relationship between the indicator of confidence in the Central Bank under consideration and the inflation expectations of the population is analyzed on a weekly basis in Matevosova (2025) by constructing an autoregressive moving average model with generalized autoregressive conditional heteroscedasticity in residuals (ARMA-GARCH). In addition to the lags of the confidence indicator, the following exogenous regressors are added to the model: lags in the consumer price index and the exchange rate. The dependent variable is inflation expectations, measured using the inflation expectations indicator based on big data Goloshchapova I. (2017). It follows from the results of

the model that an increase in public confidence in the Central Bank, all other things being equal, significantly reduces inflation expectations with a two-week lag. This conclusion is another evidence in favor of the relevance of the indicator of confidence in the Central Bank at a high-frequency level.

5 Conclusion

The methodology for constructing an indicator of trust in the Central Bank was tested by building an indicator of public trust in the Bank of Russia based on comments from users of the VK social network. The relevance of the indicator of trust in the Central Bank based on sentiment analysis and its advantages over classical survey-based methods were demonstrated. An analysis of the dynamics of the indicator of trust in the Central Bank can potentially improve the effectiveness of monetary policy. However, to achieve this effect, it is necessary to use the most effective methods of sentiment analysis based on this comparison. The approach based on fine–tuning the encoder-only model RuBERT was recognized as the most effective approach.

According to the results of a study of the use of various models to solve the problem of sentiment analysis, fine-tuning the RuBERT model on 7000 comments demonstrated the best quality. It can be seen that the RuBERT encoder language model does a really good job of understanding natural language, such as classifying texts by emotional color, surpassing generative models. The RuBERT fine-tuning approach not only allows you to achieve the best quality among the models under consideration, but also does not require large computing resources to perform fine-tuning, and also provides fast final application, therefore it is the most effective.

References

Carbo J. M. Alonso-Robisco A. Analysis of cbdc narrative by central banks using large language models. *Finance Research Letters*, 58(104643), 2023.

Miccoli J. Quarta M. Filippo Angelico, Marcucci C. Can we measure inflation expectations using twitter? *Journal of Econometrics*, 228(2):259–277, 2022.

Bryce C. Dowling M. Long S.C. Ardekani A. M., Bertz J. Finsentgpt: A universal financial sentiment engine? *International Review of Financial Analysis*, 94(103291), 2024. doi: 10.1016/j.irfa.2024.103291.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Stracca L. Ehrmann M., Soudan M. Explaining european union citizens' trust in the european central bank in normal and crisis times. *Scandinavian Journal of Economics*, 115(3):781–807, 2013.

Andreev M. Goloshchapova I. Measuring inflation expectations of the russian population with the help of machine learning. *Voprosy Ekonomiki*, (6):71–93, 2017. doi: 10.32609/0042-8736-2017-6-71-93.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Wang H. Yang Y. Huang, A.H. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023. doi: 10.1111/1911-3846.12832.

Germano G. Kirtac K. Sentiment trading with large language models. *Finance Research Letters*, 62 (105227), 2024. doi: 10.1016/j.frl.2024.105227.

A. Matevosova. Modelling trust in the central bank using sentiment analysis. *Russian Journal of Money and Finance*, 84(1):3–25, 2025.

Marohl V.P. Pfeifer M. Centralbankroberta: A fine-tuned large language model for central bank communications. *The Journal of Finance and Data Science*, 9(100114), 2023. doi: 10.1016/j.jfds.2023.100114.

Walti. Trust no more? the impact of the crisis on citizens' trust in central banks. *Journal of International Money and Finance*, 31(3):593–605, 2012.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. A family of pretrained transformer language models for russian, 2023.