**Authors:**

Aleksandr Epaneshnikov,

Tatyana Vakhrusheva,

Aleksandr Karzhanov

# **Author-based sentiment prediction**

## **Abstract**

The sentiment analysis of textual data has become an important area of research in natural language processing (NLP) due to the increasing amount of user-generated content on social media platforms. In this paper, we propose an author-based sentiment prediction model that takes into account the author's writing style and historical sentiment. The model uses a combination of machine learning and network analysis techniques to predict the sentiment of a given text.

Ultimately, we suggest using our model to predict the sentiment of the employees’ reviews on the employers in the Russian website CareerHabr. In order to evaluate the quality of prediction, the major measure we intend to use is F1 score.

## **Introduction**

Author-based sentiment prediction refers to the process of predicting the sentiment or emotional tone of a text based on the author's writing style, tone, and other characteristics. This approach assumes that the way an author writes can reveal their underlying emotions and attitudes towards a particular topic or subject.

Sentiment prediction as a whole has many practical applications, such as in marketing and customer service. By analyzing the sentiment of customer reviews, companies can identify areas where they might need to improve their products or services.[[1]](#footnote-0) Similarly, by analyzing the sentiment of social media posts, companies can gauge public opinion about their brand or products and consequently attempt to alter it. Furthermore, in politics, sentiment prediction can be used to analyze the public's reaction to political events and even predict election outcomes.

Despite its potential applications, sentiment prediction is still an active research area with several challenges. One of the main challenges is dealing with sarcasm and irony, which can often lead to misclassification of sentiment. Another challenge is dealing with domain-specific language and slang, which may not be present in standard sentiment lexicons.[[2]](#footnote-1) Apart from that, it is negations and non-standard word order in sentences that often complicate the process of extracting text’s sentiment.

## **Main part**

### **Literature Review**

Before the growth of ML and DL market and technology, to distinguish the sentiment of the given text, there have frequently been applied such methods as dictionary-based evaluation along with rule-based processing. To elaborate on it, sentiment evaluation envisaged matching each meaningful word of a text with the dictionary’s meaning and score from -1 (negative sentiment) to 1 (positive meaning), or, sometimes, 0 to 1. Each word’s score would then most frequently be added up and divided by the number of words, and the overall text’s score was supposed to give understanding of its sentiment.[[3]](#footnote-2) The dictionaries in use were generally thoroughly developed by domain researchers, as, for instance, was the case with SentiWordNet and its editions.[[4]](#footnote-3) The applied algorithm to calculate the sentiment of the whole piece of text would also evolve and get proposed to undergo multiple changes.

Owing to the multiple disadvantages of the approaches discussed above, such as, inability to work with structurally and semantically difficult text structures, negations, irony and sarcasm, as well as the sharp NLP development, over the recent years, to predict text sentiment analysis, researchers and developers have mainly been applying several fundamental Natural Language Processing (NLP) methods, which could be divided into the fields of machine learning (hereinafter referred to as “ML”), deep learning (hereinafter referred to as “DL”), as well as a mixture of ML and DL methods.

One of the most used ML models towards sentiment analysis is BERT versions – RoBERTa, AlBERT, DistilBERT – pre-trained deep learning models for natural language processing tasks, developed by Google. Among the greatest advantages of the BERT models is their ability to take into account the semantics of the sentences and word order in them.[[5]](#footnote-4) To be more precise, transformer, a fundamental part of these models, “process any given word in relation to all other words in a sentence, rather than [process] them one at a time”[[6]](#footnote-5). Fine-tuning BERT models involves changing the weights or adding extra stages of data processing.

In this way, the researchers of Stony Brook University released a dataset of collected news articles with the aim of encouraging the scientific community to find more efficient tools of classifying the articles by their sentiment. However, one of the researchers’ major conclusions was about the low efficiency of the fine-tuned BERT model for the given task.[[7]](#footnote-6)

In turn, DL methods to predict the sentiment essentially use neural networks to learn from large amounts of data and predict sentiment in new text. There are two main types of deep learning models that have been used for sentiment analysis: Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). In sentiment analysis, a CNN takes the text as input and applies a series of convolutional filters to extract features from the text. These features are then fed into a fully connected layer, which predicts the sentiment of the text.[[8]](#footnote-7) RNNs are designed to handle sequential data, such as text. In sentiment analysis, an RNN takes the text as input and processes it one word at a time, updating its internal state with each new word.[[9]](#footnote-8) The final state of the RNN is then used to predict the sentiment of the text.

Both CNNs and RNNs can be combined with other deep learning techniques, such as word embeddings and attention mechanisms, to improve the accuracy of sentiment analysis. The biggest disadvantage of deep learning methods is their demand for large amounts of labeled data and computational resources.

### **Anticipated SNA Methods**

The objectives of the work are to study the IT companies of the Russian market, including the construction of graphs of companies based on the commonality of their characteristics, as well as the construction of a model capable of predicting the numerical assessment in the employee's review of the company, based on a variety of characteristics, including information about the company, text, and its sentiment, demographic and other employee data.

Collected data from the website <https://career.habr.com/journal> will allow us to build a model and graph based on various data, namely those shown in Figure 1.

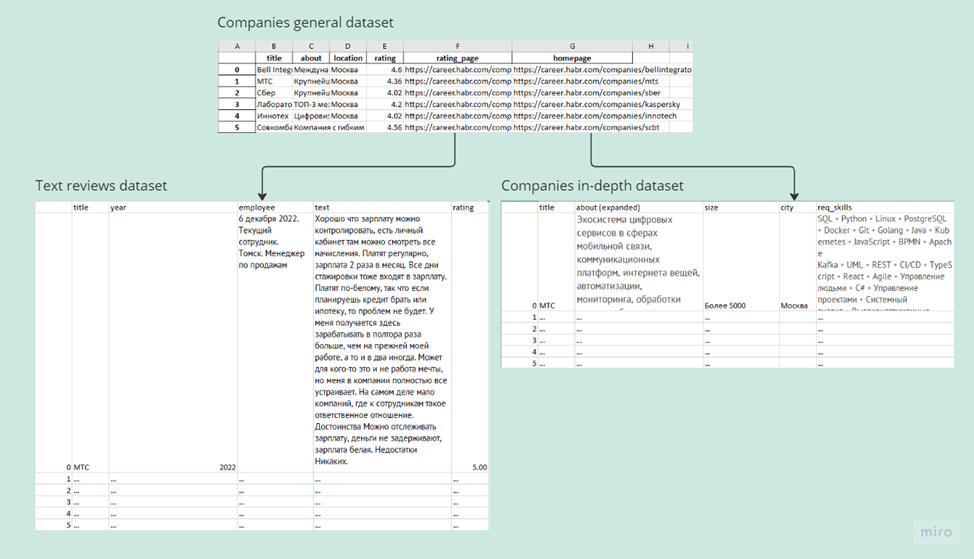


Fig 1.

After extracting and collecting the necessary data, all data from the ‘companies in-depth dataset’ will be transformed into a graph format, where the nodes will be skills and companies, and edges – the presence of this skill among the necessary ones in the company. An example of using a Bipartite graph, with which it will be possible to make clusters of companies by skill set in Figure 2:

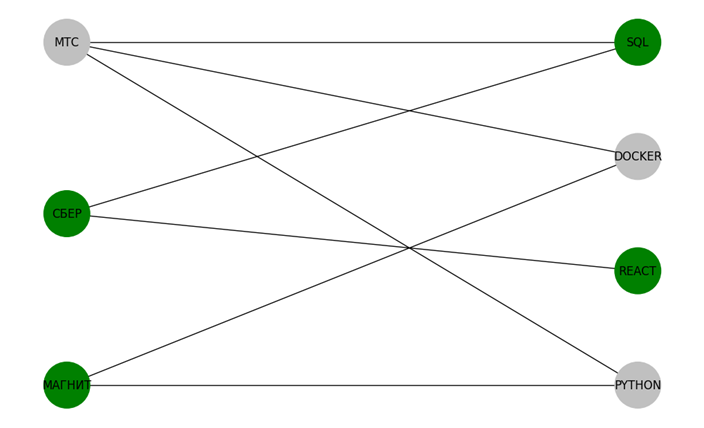


Fig. 2

The collected data from the ‘text reviews dataset’ will be sent to the pre-trained BERT model for the text classification task. The model for extracting embeddings can be mUSE, rubert-tiny 2, or Fastext.[[10]](#footnote-9). The target for this model will be the user's numerical assessment of the company (text\_reviews\_dataset[‘rating’]).

### **Expected Results**

Graph representation of companies depending on their characteristics, a model capable of predicting an employee's assessment of the company based on his text (sentiment) and other characteristics.

## 

## **References**

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. “SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining,” n.d.

Bastan, Mohaddeseh, Mahnaz Koupaee, Youngseo Son, Richard Sicoli, and Niranjan Balasubramanian. “Author’s Sentiment Prediction.” arXiv, November 11, 2020.<http://arxiv.org/abs/2011.06128>.

Elia, Francesco. “Sentiment Analysis Dictionaries | Baeldung on Computer Science,” August 26, 2020.<https://www.baeldung.com/cs/sentiment-analysis-dictionaries>.

Enterprise AI. “What Is BERT (Language Model) and How Does It Work?” Accessed May 2, 2023. <https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model>.

Islam, Md Saiful. *A Deep Recurrent Neural Network with BiLSTM Model for Sentiment Classification*, 2018.

Larin, Vladimir. “Transformers-Ru.” Jupyter Notebook, April 11, 2023.<https://github.com/vlarine/transformers-ru>.

Mejova, Yelena. “Sentiment Analysis: An Overview,” 2009.

Rojas-Barahona, Lina. “Deep Learning for Sentiment Analysis.” Language and Linguistic Compass, December 29, 2016.

Zhang, Yangsen, Jia Zheng, Yuru Jiang, Gaijuan Huang, and Ruoyu Chen. “A Text Sentiment Classification Modeling Method Based on Coordinated CNN-LSTM-Attention Model.” *Chinese Journal of Electronics* 28, no. 1 (2019): 120–26.<https://doi.org/10.1049/cje.2018.11.004>.

Хабр. “Рейтинг русскоязычных энкодеров предложений,” June 5, 2022.<https://habr.com/ru/articles/669674/>.

Хабр. “Маленький и быстрый BERT для русского языка,” June 9, 2021.<https://habr.com/ru/articles/562064/>.

Хабр. “Fine-Tune модели основанной на трансформерах (Rubert) для классификации текстов,” August 15, 2022.<https://habr.com/ru/articles/682476/>.

1. Yelena Mejova, “Sentiment Analysis: An Overview,” 2009. [↑](#footnote-ref-0)
2. Lina Rojas-Barahona, “Deep Learning for Sentiment Analysis,” *Language and Linguistic Compass*, December 29, 2016, 3-4. [↑](#footnote-ref-1)
3. Francesco Elia, “Sentiment Analysis Dictionaries | Baeldung on Computer Science,” August 26, 2020,<https://www.baeldung.com/cs/sentiment-analysis-dictionaries>. [↑](#footnote-ref-2)
4. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, “SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining,” n.d. [↑](#footnote-ref-3)
5. “Fine-Tune модели основанной на трансформерах (Rubert) для классификации текстов,” Хабр, August 15, 2022,<https://habr.com/ru/articles/682476/>. [↑](#footnote-ref-4)
6. “What Is BERT (Language Model) and How Does It Work?,” Enterprise AI, accessed May 2, 2023,<https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model>. [↑](#footnote-ref-5)
7. Mohaddeseh Bastan et al., “Author’s Sentiment Prediction” (arXiv, November 11, 2020),<http://arxiv.org/abs/2011.06128>. [↑](#footnote-ref-6)
8. Yangsen Zhang et al., “A Text Sentiment Classification Modeling Method Based on Coordinated CNN-LSTM-Attention Model,” *Chinese Journal of Electronics* 28, no. 1 (2019): 120–26,<https://doi.org/10.1049/cje.2018.11.004>. [↑](#footnote-ref-7)
9. Md Saiful Islam, *A Deep Recurrent Neural Network with BiLSTM Model for Sentiment Classification*, 2018. [↑](#footnote-ref-8)
10. see “Рейтинг русскоязычных энкодеров предложений,” Хабр, June 5, 2022,<https://habr.com/ru/articles/669674/>; “Маленький и быстрый BERT для русского языка,” Хабр, June 9, 2021,<https://habr.com/ru/articles/562064/>; and Vladimir Larin, “Transformers-Ru,” Jupyter Notebook, April 11, 2023,<https://github.com/vlarine/transformers-ru> [↑](#footnote-ref-9)