ISSN 2710 - 1673 Artificial Intelligence 2024 № 2

UDC: 004.91 https://doi.org/10.15407/jai2024.02.085

M. Prytula
Ivan Franko National University of Lviv, Ukraine
50, Dragomanov Str., Lviv, 79005
marianna.prytula@lnu.edu.ua
https://orcid.org/0000-0003-3410-3350

FINE-TUNING BERT, DISTILBERT, XLM-ROBERTA AND UKR-ROBERTA MODELS FOR SENTIMENT ANALYSIS OF UKRAINIAN LANGUAGE REVIEWS

Abstract. Sentiment analysis is one of the crucial tasks of natural language processing, which includes recognizing emotions expressed in textual data from various fields of activity. Automated tonality detection impacts businesses and helps increase profits by analyzing customer sentiment and responding quickly to their level of satisfaction with products or services. Therefore, the development of tools that will allow qualitative classification of text sentiment is significant, considering that users leave many reviews on various social networks, platforms, and websites in today's world. The study examines the fine-tuning of BERT, DistilBERT, XLM-RoBERTa, and Ukr-RoBERTa models for sentiment analysis of reviews in the Ukrainian language, as transformer models demonstrate a better understanding of the context and show high efficiency in solving natural language processing tasks. The dataset used in this study comprised about 11,000 user comments in Ukrainian, covering a range of topics such as shops, restaurants, hotels, medical facilities, fitness clubs, and the provision of various services. The textual data was categorized into two classes: positive and negative. Following text preprocessing, the dataset was divided into training and test samples in an 80:20 ratio. The hyperparameters were selected to optimize the performance of the pre-trained models for comment sentiment classification, and their effectiveness was evaluated using metrics such as accuracy, recall, precision, and F1-score. The results show that DistilBERT requires significantly fewer computing resources and is faster than other models. The XLM-RoBERTa model achieved the highest accuracy of 91.32%. However, considering the time needed to train the model and all the classification metrics, Ukr-RoBERTa is the optimal choice.
Keywords: sentiment analysis, transformer models, BERT, DistilBERT, XLM-RoBERTa, Ukr-RoBERTa.

Introduction

Sentiment analysis in reviews involves examining product reviews to gauge the overall opinion or user's feeling about a product. Reviews, as a type of user-generated content, are becoming increasingly important and valuable for marketing teams, sociologists, psychologists, and others interested in understanding public mood, attitudes, and opinions.

BM Watson, a leading provider of AI solutions, reports that businesses utilizing AI for customer emotion analysis experience a 20% increase in customer satisfaction and a 15% increase in revenue [1].

Research conducted by Forrester showed that companies prioritizing understanding customer emotions are 85 % more likely to exceed revenue goals [2].

As the volume of online user reviews continuously grows, there is a greater demand for automated processing of this enormous amount of data. It is not feasible for humans to read and analyze all this textual information manually and independently. Therefore, sentiment classification is essential to analyzing and categorizing these documents.

Different types of sentiment analysis can be conducted based on specific needs, including binary and multi-class classification, emotion extraction, aspect-based, or fine-grained sentiment analysis. Binary sentiment analysis categorizes text into two distinct categories: usually positive and negative. This method offers a simple way to determine the overall sentiment of a given text, enables prompt evaluation of customer feedback, and helps recognize broad trends in consumer satisfaction or dissatisfaction. Multi-class sentiment analysis extends beyond binary classification, categorizing text into multiple sentiment classes, for example, positive, neutral, and negative. A more nuanced understanding of sentiment can be provided by allowing for differentiation between varying degrees of opinion. Extracting emotions involves identifying and categorizing specific emotions expressed in a text, such as joy, anger, sadness, surprise, and fear. This process enables an understanding of the emotional context and intensity behind textual data. It is widely used in fields like mental health analysis, where identifying specific emotions can offer vital insights into an individual's emotional state. Aspect-based sentiment analysis concentrates on recognizing sentiments toward specific aspects or features of a product or service mentioned in the text. Businesses can identify strengths and weaknesses in particular areas, such as product quality, customer service, or specific functionalities. Fine-grained sentiment analysis involves assigning a sentiment score to text, often on a numerical scale from 1 to 5, to indicate the strength of the sentiment. This approach offers a more detailed view of sentiment intensity, allowing organizations to quantify and monitor changes over time. Each sentiment analysis type provides unique insights and advantages according to organizations' requirements, enabling better understanding and addressing emotions expressed by customers, users, or stakeholders.

Various methods can be used for this task, starting with rule-based and lexicon-based approaches, machine learning models, and utilizing the most recent transfer learning techniques. A sentiment dictionary containing words or phrases along with their corresponding tonality values is utilized in a lexicon-based approach. Meanwhile, a rule-based approach uses rules to determine sentiment. These rules can utilize lists of positive and negative words, syntactic patterns, or more complex linguistic structures [3, 4].

Machine learning classifiers for real-time predictions, including Naive Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Decision Trees, etc., have repeatedly demonstrated evidence of efficacy in text classification scenarios [5-7].

Numerous researchers evaluate different deep learning techniques for sentiment analysis. Convolutional neural networks can extract local features from text, while recurrent neural networks can capture sequential dependencies and contextual information. The results of employing advanced neural network architectures provide evidence of their compelling effectiveness in categorizing sentiment [8, 9].

Recent advancements in natural language processing, especially with transformer language models, offer a promising opportunity for AI-driven businesses. Multilingual language models enhance the quality and accuracy of text analysis tasks when dealing with texts written in different languages. Studies have shown that transfer learning can greatly reduce the need for large, domain-specific datasets during training effective language models [10, 11].

Therefore, this paper aims to investigate the peculiarities of using large language models for the sentiment analysis of customer reviews in the Ukrainian language from different domains. Four pre-trained transformer models were used to classify binary sentiment.

Analysis of recent research and publications

A comprehensive research of NLP-based methods for sentiment analysis in finance compared different approaches, starting with lexicon-based methods and concluding with transformer models. The study found that NLP transformers performed better than other evaluated approaches. Despite using a relatively small dataset, the results suggest that these models are suitable for domains where extensive annotated data is unavailable [12].

Another research study, which performed sentiment analysis on a dataset of Amazon reviews divided as positive or negative based on the number of stars included, was conducted in the thesis [13]. Based on the experiment results, DistilBERT was faster than BERT and maintained 99.6 % accuracy, performing 0.39 % worse than the BERT version of this model variation. DistilBERT maintained 99.1 % accuracy compared to RoBERTa, performing 0.68 % worse than the RoBERTa version. Although DistilBERT did not display the overall highest performance, it surpassed the other models in its ability to train large amounts of data efficiently. Finally, it was summarized that RoBERTa performs the best, followed by BERT and DistilBERT.

The paper [14] compares nine transfer learning models for classifying a COVID-19-related dataset in the English language. The models included BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, DistilBERT, XLM-RoBERTa-base, ALBERT-base-v2, Electra-small, and BART-large. The results showed that BART-large, BERT-base, and BERT-large achieved the highest accuracy.

A study [15] found that ukr-RoBERTa is more effective for short-length texts, while XLM-RoBERTa and ukr-ELECTRA are the better choices for longer texts in the Ukrainian language news classification dataset.

Overview of the used pre-trained models

Transformer-based models are a specific type of deep learning model architecture that has become very popular in natural language processing and other domains. At the core of these models is the self-attention mechanism, which allows the model to weigh the importance of different parts of the input sequence when processing each token. Transformers are built upon the encoder-decoder architecture. The encoder processes the input sequence to create a fixed-dimensional representation, while the decoder generates an output sequence. The model can focus on different input parts simultaneously by using multiple attention heads in parallel during both the encoder and decoder stages. Since transformers do not inherently understand the order of tokens in a sequence as recurrent or convolutional models do, positional encodings are added to provide this information to the model.

Transformers revolutionize natural language processing tasks by using attention mechanisms to efficiently learn contextual relationships within sequences, making them crucial in modern AI applications. These models are typically used in two stages: pre-training and fine-tuning.

The BERT pre-training process includes two unsupervised predictive tasks. The first task is the Masked Language Model, which hides certain words during training and tries to determine the missing word. The second task is Next Sentence Prediction, determining whether the second sentence comes after the first. The transformer encoder reads the entire sequence of words at once and learns the context of a word based on its surrounding words. The representation of each input sentence is created by combining positional embedding, which indicates the word's position in the sequence, segment embedding, which differentiates between sentences, and word embedding. At the beginning of each sentence, a special classification token [CLS] is added, and the final hidden state aggregate sequence representation of this token is used for classification tasks. Additionally, a special token, [SEP], is included to mark the end of a sentence or the separation between two sentences. The sum of all these embeddings forms the input layers for BERT [16].

The BERT is a deep bidirectional transformer architecture introduced by Google. It supports a multilingual universal language representation for 104 languages. BERT is pre-trained using 2,500 million words from unlabeled Wikipedia texts and 800 million words from the Book corpus to obtain contextual embeddings.

Based on the depth of the model architecture, there are two versions of the BERT language model:

- BERT base has 12 layers, 12 attention heads with 768 hidden dimensions, and a feed-forward network with 3072 dimensions, providing 110 million parameters in total.
- BERT large has 24 layers, 16 attention heads with 1024 dimensions, and 4096 feed-forward filters, resulting in 340 million parameters.

During fine-tuning, an untrained layer is added at the top of the output layer of the pre-trained transformer-based model. The pre-trained model weights already encode extensive information about the language, and this encoded information is used as a feature for the classification task. The fine-tuning tasks require less time to train on a much smaller dataset with these features, eliminating the need to learn the language from scratch. This approach allows BERT to accomplish state-of-the-art results on various downstream NLP tasks such as Named Entity Recognition, Question Answering, Sentiment Analysis, and Text Classification.

The architecture of the DistilBERT model is a simplified version of BERT, consisting of 6 layers, 6 attention heads with 768 dimensions, and a feed-forward network. The resulting model has 66 million parameters. This model was trained on the same dataset as BERT.

The DistilBERT is 60 % faster and 40 % smaller than the base BERT model while retaining 97 % of BERT’s performance [17]. It achieves this through distillation, a method that compresses a larger model into a smaller one, making the model more computationally efficient and faster.

Knowledge distillation is a technique in which a smaller model, known as the student, learns to mimic the behavior of a larger model, called the teacher, by reproducing its predicted probabilities. In supervised learning, models are usually trained to predict the correct class labels by minimizing the cross-entropy loss, which measures the difference between the predicted probabilities and the true labels, known as gold values. During knowledge distillation, the student model is trained using a distillation loss that leverages the teacher's full output distribution, providing a richer training signal. This process involves using a temperature parameter to smooth the output probabilities, making the probabilities less extreme. The final training objective combines the distillation loss with the supervised training loss to improve the student's performance.

RoBERTa, which stands for robustly optimized BERT approach, was developed by Facebook AI to improve BERT's performance through crucial optimizations in the training process. One of the main changes in RoBERTa is the removal of the Next Sentence Prediction objective, simplifying the training process and focusing solely on masked language modeling. RoBERTa uses dynamic masking, where the different tokens are masked in each epoch, leading to more robust training and improved generalization. The model is trained with larger mini-batches, higher learning rates, and significantly more data, covering 160 GB of text from various sources [18].

The XLM-RoBERTa model (Cross-lingual Language Model) is an extension of RoBERTa and an improvement over BERT. It is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. The architecture of the XLM-RoBERTa base model consists of 12 layers, 12 attention heads with 768 dimensions, and a feed-forward network [19]. This model is trained for more epochs than BERT and utilizes larger batch sizes during training, allowing the model to process more examples at once, thereby enhancing the stability and efficiency of the learning process. It also incorporates dynamic masking during the pre-training phase, which means that the masked tokens change between training epochs, providing the model with a more comprehensive understanding of context. These improvements enable XLM-RoBERTa to outperform models like BERT in multilingual tasks.

The Ukr-RoBERTa is a version of the RoBERTa model pre-trained specifically on a large-scale corpus consisting of Ukrainian Wikipedia, Ukrainian OSCAR deduplicated dataset, and internal dataset collected from social networks [20]. The model adheres to the architecture of the roberta-base-case model, which includes 12 layers, 768 hidden units, 12 attention heads, and 125 million parameters. It was mentioned that upon testing ukr-roberta-base on internal tasks, an average increase of 2 percent in the F1-score was achieved compared to multilingual BERT on multiclass and multilabel classification tasks.

Methodology

The BERT, DistilBERT and XLM-RoBERTa models for sentiment classification were implemented using the Transformers and TensorFlow libraries according to the diagram in Fig. 1.

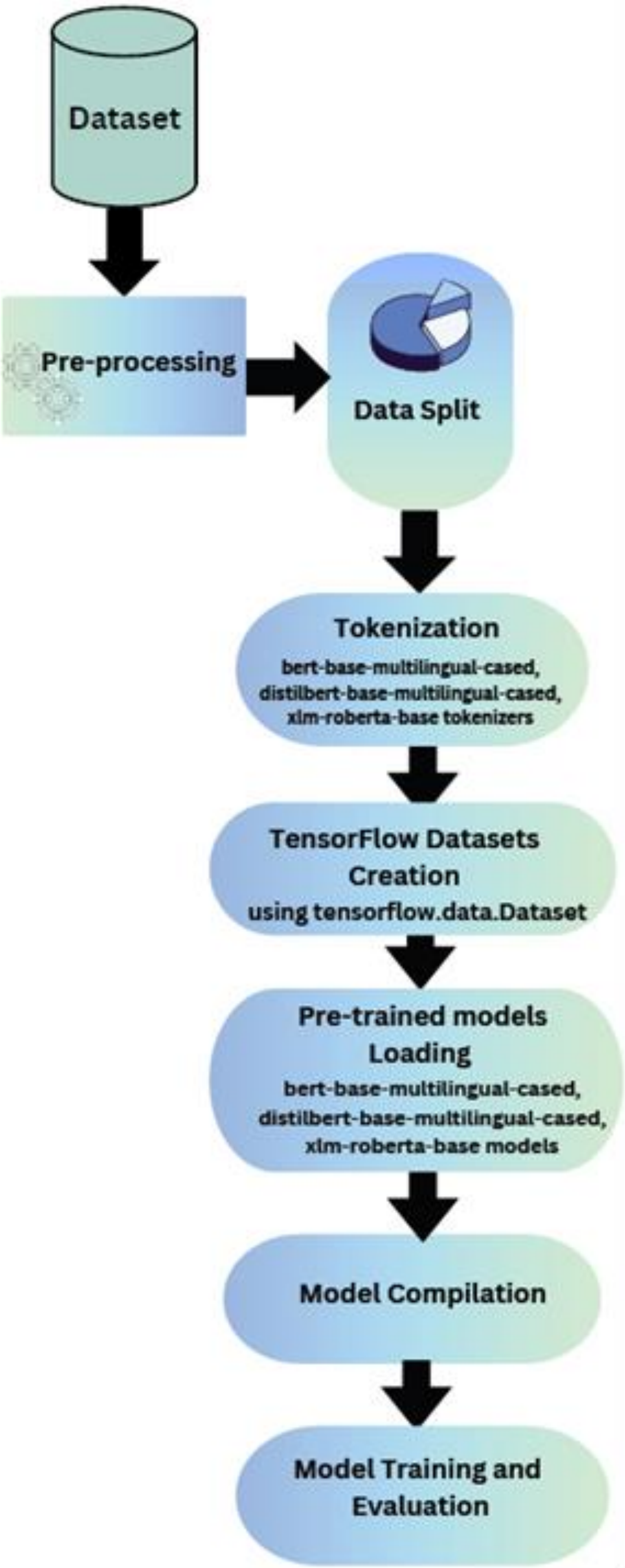


Fig. 1. Transformer-based model fine-tuning pipeline with TensorFlow library

This research used the custom multi-domain dataset with scraped user comments in the Ukrainian language for binary tonality classification: positive and negative. It contains textual information from customers about shops, restaurants, hotels, medical institutions, fitness clubs, and service providers. The general number of comments is 10,997 with the following distribution by classes 6136:4861, respectively, for negative and positive sentiment, as shown in Fig. 2. Class 0 relates to negative and class 1 to positive comments.

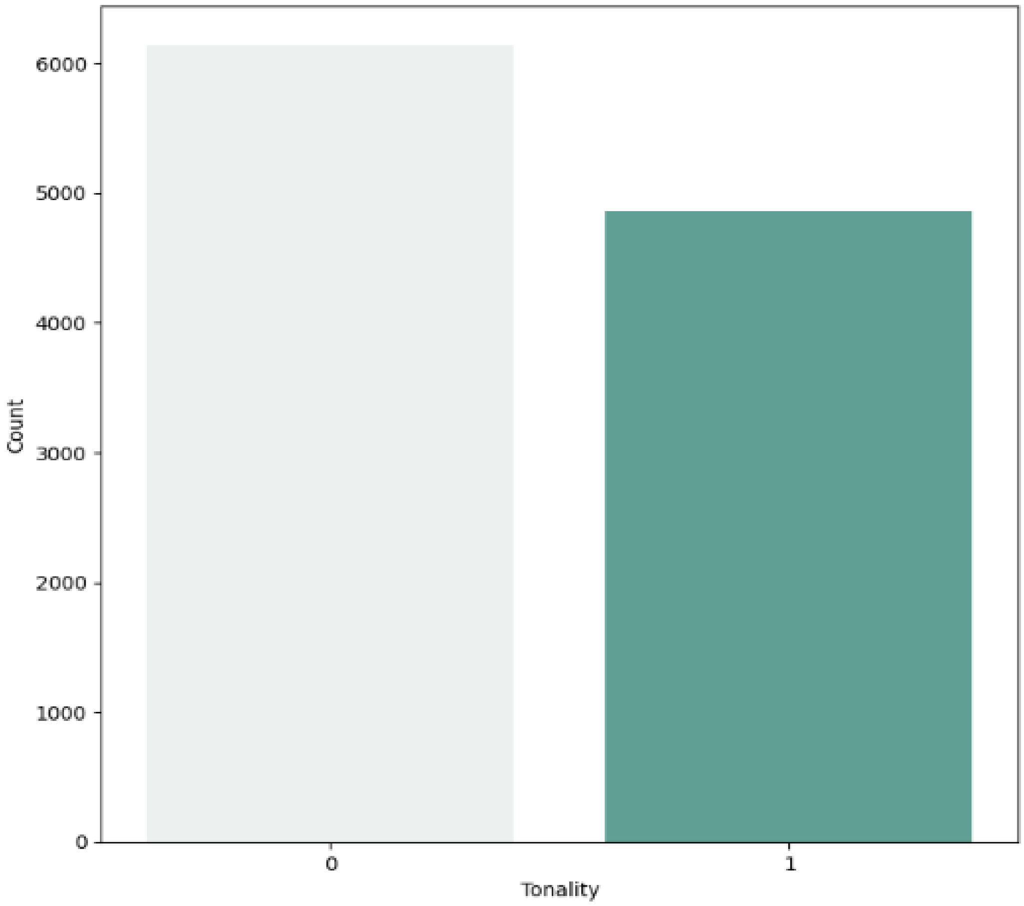


Fig. 2. Histogram with the tonal distribution of the dataset

Word cloud illustrates the prominent words in the corpus, with word size indicating frequency, as shown in Fig. 3.

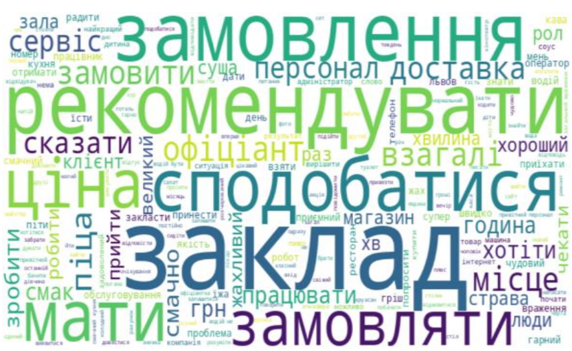


Fig. 3. Word cloud representation of the most frequent words

The following text preprocessing steps were conducted on the dataset:

1. Removal of URLs and HTML tags, punctuation, numbers, and special characters.
2. Tokenization.
3. Removal of commonly used words, which are insignificant and non-informative in natural language processing tasks, known as stop words.
4. Lemmatization was performed to normalize words by transforming them into a standard form so that words with similar meanings were grouped.

The data was divided into training and test sets: 20 % will be used for testing, and the remaining 80 % will be used for training. A random number generator was applied to ensure reproducibility of the data split for all models.

These sets were converted into TensorFlow datasets for further processing. The bert-base-multilingual-cased, and xlm-roberta-base, distilbert-base-multilingual-cased were loaded from the Hugging Face library [21–23].

A relevant tokenizer was applied to convert the text into a BERT format. This process includes encoding the text into token IDs, padding all sequences to the set maximum length (a value equal to 256 was chosen), and creating attention masks. Special tokens, such as [CLS] and [SEP], indicate the beginning and end of a sequence, respectively, in transformer-based models. An attention mechanism focuses on relevant parts of the input sequence. The attention mask represents an array that denotes which tokens are actual and which are padding. The model ignores padding tokens with an attention mask value of 0 during training and inference. The BERT model and its variants in need inputs in a specific dictionary format. So, each input is converted to a dictionary with keys corresponding to the model's expected input_ids, attention_mask feature names and combined with the tonality label.

The samples within the dataset were encoded and shuffled to prevent learning unintended sequence patterns during training. Then, they were combined into batches of a specified size, which allowed the model to process multiple examples simultaneously during each training step.

The next step was configuring the model's training, utilizing the optimizer to minimize the specified loss function with monitoring accuracy as a metric to assess model performance.

As the paper [24] suggests, the optimal hyperparameter values for pre-training deep bidirectional transformers are task-specific. However, a range of values has been identified that work well across all tasks: a batch size of 16 or 32, a learning rate (Adam) of $5e^{-5}$, $3e^{-5}$ or $2e^{-5}$, and the number of epochs of 2, 3 or 4. It was found that large data sets are less sensitive to hyperparameter selection than small data sets. This highlights the crucial role of the exhaustive search over these parameters and selecting the best model for the development set.

After performing a selection, the subsequent hyperparameter values were implemented for the BERT and DistilBERT models:

- Adam optimizer with a learning rate of $2e^{-5}$
- Batch Size: 16
- Epochs: 3.

XLM-RoBERTa performed better according to the accuracy metric with a batch size of 32, while other parameters remained the same.

The ukr-roberta-base model [25] was fine-tuned using the PyTorch library according to the pipeline in Fig. 4, as TensorFlow is incompatible according to Hugging Face's guidelines.

PyTorch datasets were created utilizing a custom dataset class inherited from torch.utils.data.Dataset and Data Loaders that handle batching and shuffling, providing iterable access to datasets during model training and evaluation using torch.utils.data.DataLoader. Hyperparameter values and a random number generator were the same as for the BERT and DistilBERT models.

Results and discussion

All proposed models were trained and evaluated using the same computing environment, Google Collab Pro, with a selected Tesla T4 GPU.

After the training, the models' performance was measured using different metrics. Confusion matrices were used to monitor the number of true positives, true negatives, false positives, and false negatives. Also, the test set's accuracy, precision, recall, and F1-score were evaluated to analyze the performance of the sentiment classification models. The obtained reports on the effectiveness of the models are shown in Fig. 5 – Fig. 8.

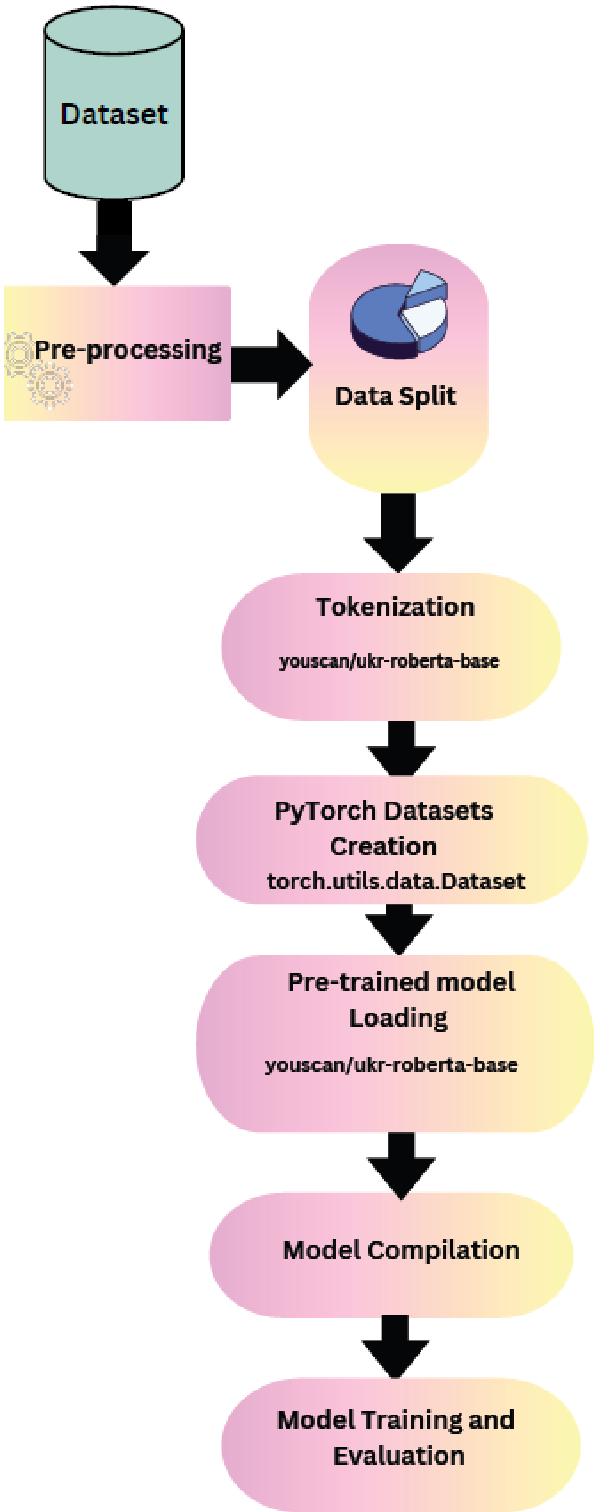
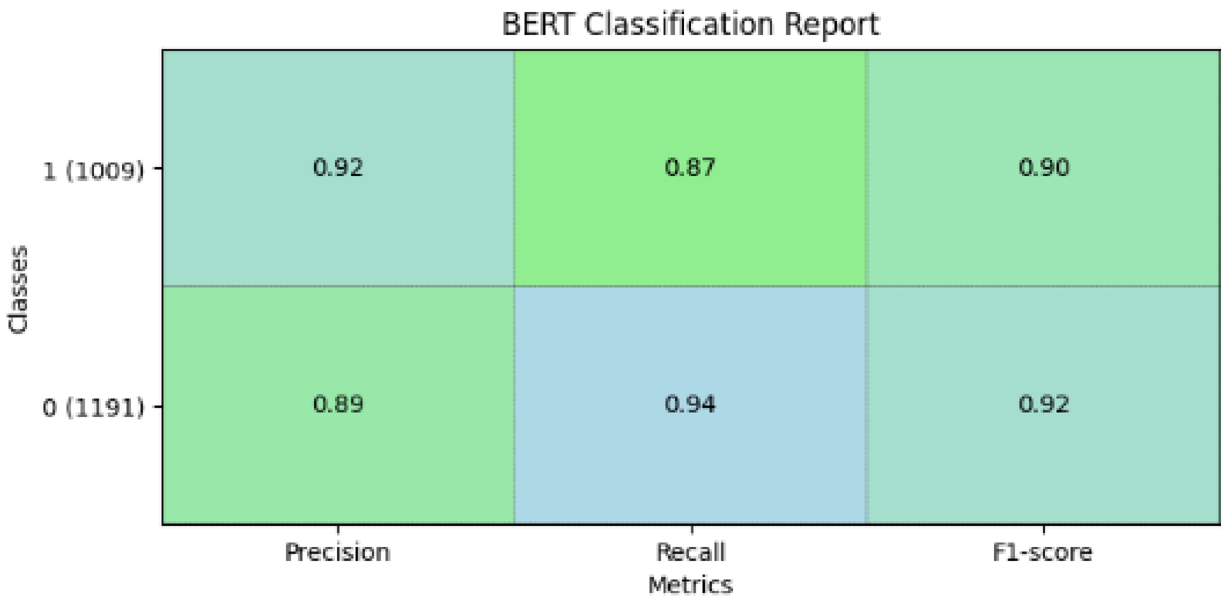
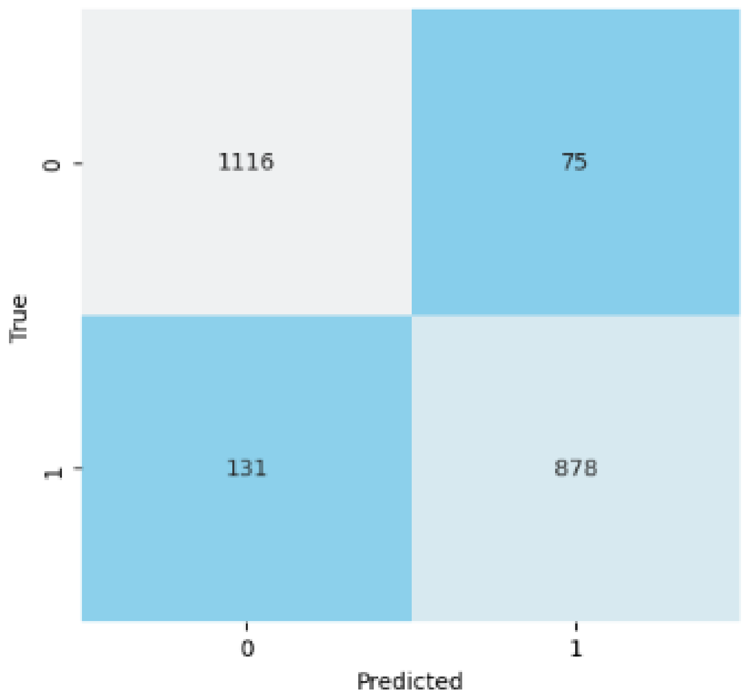


Fig. 4. Transformer-based model fine-tuning pipeline with PyTorch library



a)

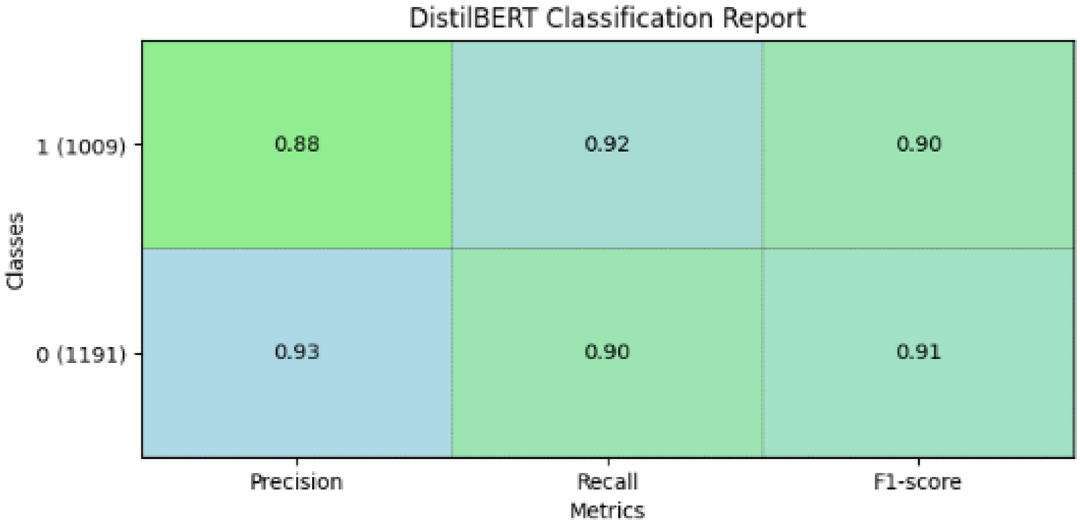


b)

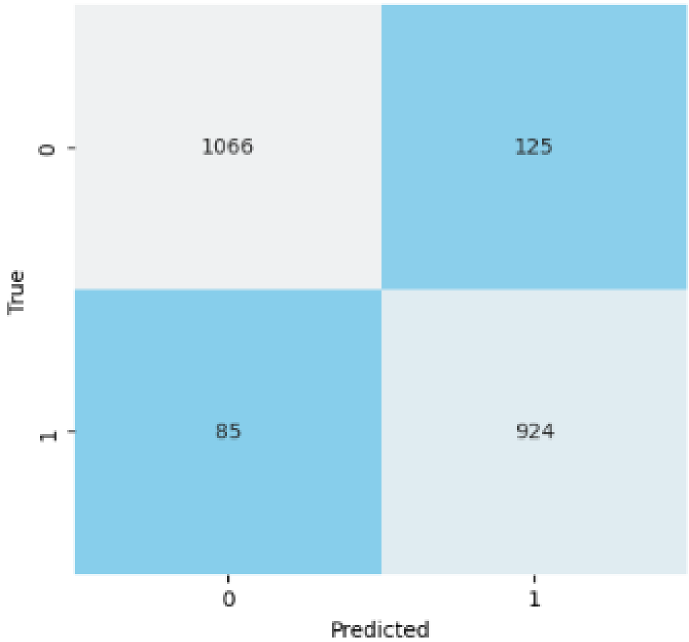
Fig. 5. Classification report (*a*) and Confusion matrix (*b*) using the BERT model

The testing set for all four models contained 1191 negative and 1009 positive samples.

The BERT classification report shows that the model achieved precision (0.92) for the positive class and recall (0.87), showing that most positive instances were correctly identified. The F1-score of 0.90 effectively balances these metrics. Conversely, it demonstrated higher recall (namely 0.94) for the negative class, while precision was slightly lower (0.89). As a result, the F1-score value was 0.92.

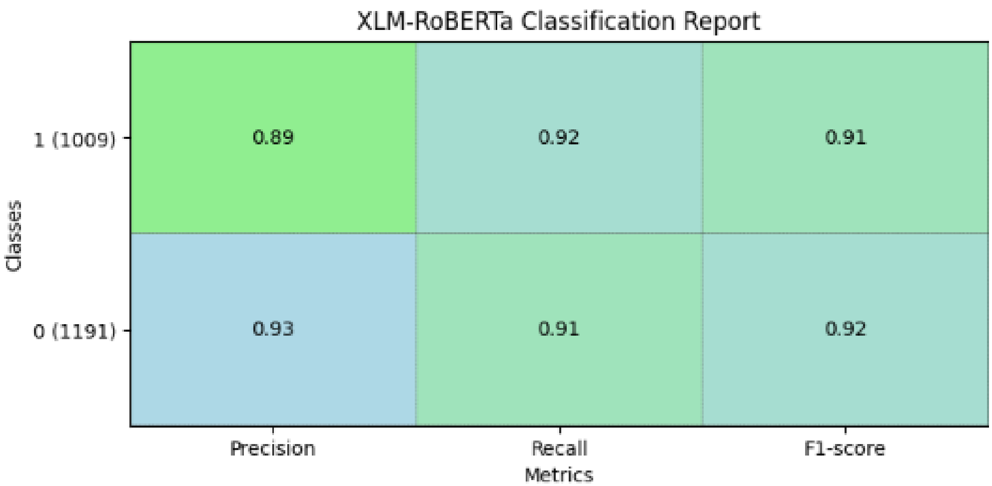


a)

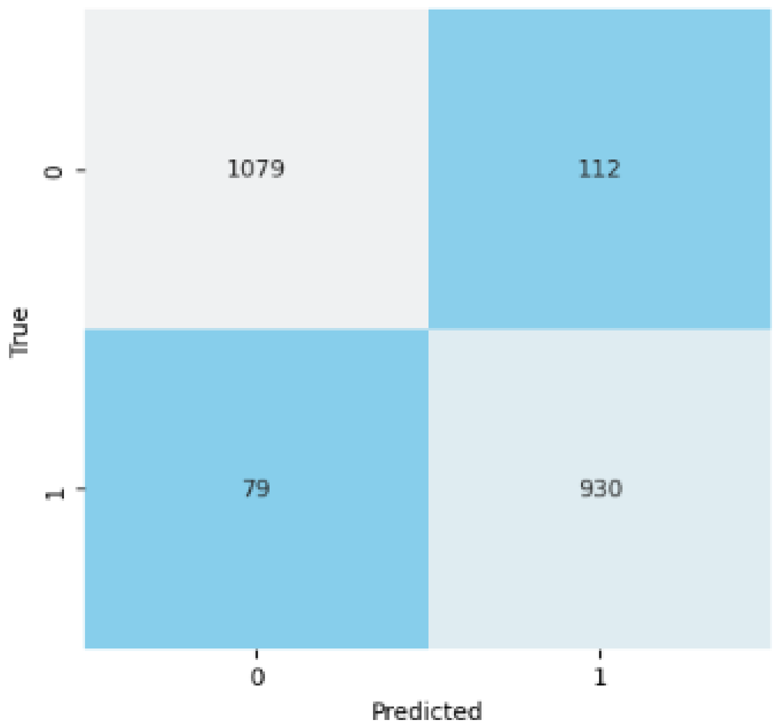


b)

Fig. 6. Classification report (*a*) and Confusion matrix (*b*) using the DistilBERT model



a)

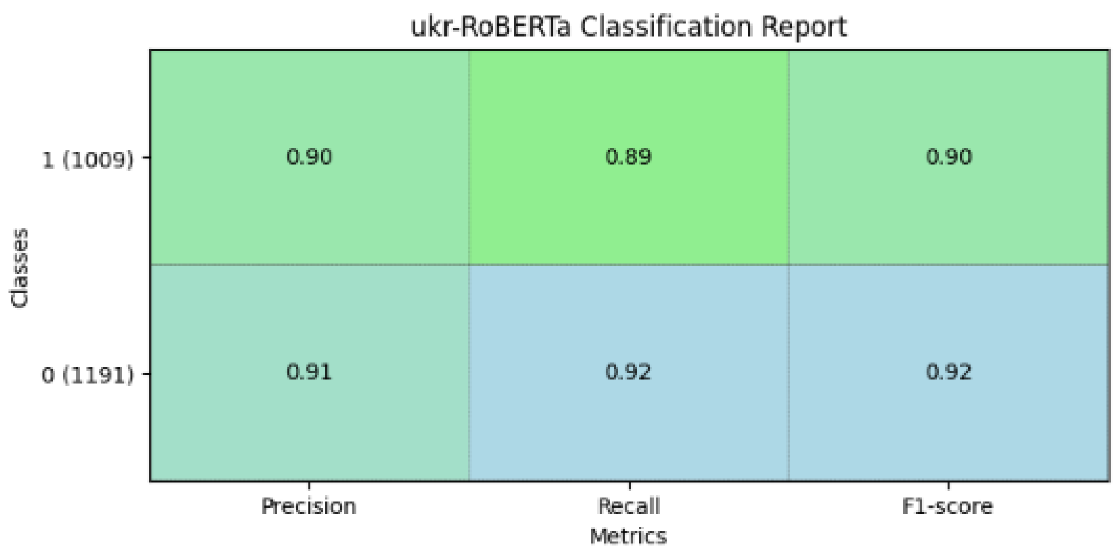


b)

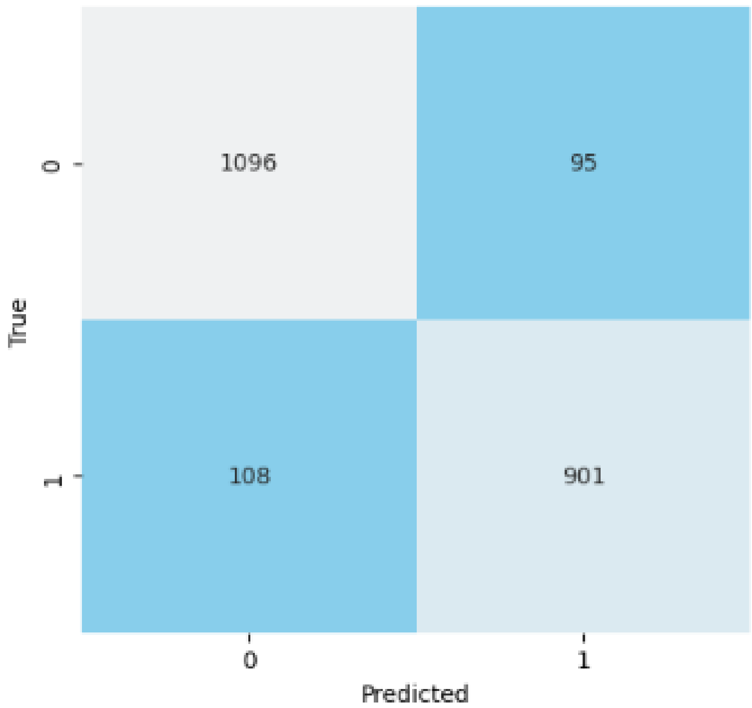
Fig. 7. Classification report (*a*) and Confusion matrix (*b*) using the XLM-RoBERTa model

Fig. 5b illustrates that the model makes more errors in predicting the positive sentiment, as indicated by the higher false positive rate for the negative class.

The DistilBERT and XLM-RoBERTa models have similar results with better prediction of negative class, as displayed in Fig. 6 – Fig. 7. However, the XLM-RoBERTa excels for 13 and 6 samples in classifying negative and positive sentiment, respectively.



a)



b)

Fig.8. Classification report (*a*) and Confusion matrix (*b*) using the Ukr-RoBERTa model

Analyzing the obtained results allows us to conclude that the binary sentiment is classified highly effectively in Ukrainian texts after fine-tuning the transformer-based models.

Table 1 summarizes the time to complete training for three epochs and the accuracy obtained on the testing set. BERT had the longest training phase duration, followed by XLM-RoBERTa, with a 0.68% difference in accuracy. The DistilBERT was the fastest, taking only 12 minutes and 51 seconds to train. However, its performance was the lowest. The ukr-RoBERTa was slightly slower than DistilBERT and achieved second place according to accuracy. To sum up, the Ukr-RoBERTa model is the best choice considering both the classification metrics (see Fig. 8) and the time for the training phase.

Conclusions

This study highlights the effectiveness of transformer models such as BERT, DistilBERT, XLM-RoBERTa and ukr-RoBERTa. The investigation involved using a customized Ukrainian language dataset to search for hyperparameters and fine-tune models to achieve notable results in accuracy and efficiency.

Table 1. Training time and accuracy of proposed models for sentiment analysis

Model	Training stage completion time	Accuracy on the testing set
BERT	00:26:17	90.64%
DistilBERT	00:12:51	90.45%
XLM-RoBERTa	00:25:11	91.32%
Ukr-RoBERTa	00:17:09	91.18%

The accuracies of all the proposed models ranged from 90.45% to 91.32%, with XLM-RoBERTa achieving the highest value. The results show that fine-tuning provides notable performance not only in high-resource languages but also in low-resource languages, such as Ukrainian.

Further research should focus on experimenting with different pre-trained models and architectures, as well as increasing the size and quality of the training dataset.

References

1. The importance of using AI-driven sentiment analysis in customer feedback [Electronic resource]. - Mode of access: https://moldstud.com/articles/p-the-importance-of-using-ai-driven-sentiment-analysis-in-customer-feedback
2. Designing for Emotional Resonance in Software Interactions [Electronic resource]. - Mode of access: https://moldstud.com/articles/p-designing-for emotional-resonance-in-software-interactions
3. Bonta, V., Kumaresh, N., Janardhan, N. (2018) A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology,* 8(S2), 1-6.
4. Nandwani, P., Verma, R. (2021) A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11, 81. https://doi.org/10.1007/s13278-021-00776-6
5. Pavitha, N., Pungliya, V., Raut, A., Bhonsle, R., Purohit, A., Patel, A., Shashidhar, R. (2022) Movie recommendation and sentiment analysis using machine learning. *Global Transitions Proceedings,* 3, 279-284. https://doi.org/10.1016/j.gltp.2022.03.012
6. Haque, T., Saber, N., Shah, F. (2018) Sentiment analysis on large scale Amazon product reviews. *IEEE International Conference on Innovative Research and Development*. 10.1109/ICIRD.2018.8376299
7. Olenych, I., Prytula, M., Boyko, Ya., Sinkevych, O., Khamar, O. (2023) Using feature engineering in machine learning models for fake news detection. *Electronics and information technologies*, 24, 46-56.
8. Bharadwaj, L. (2023) Sentiment Analysis in Online Product Reviews: Mining Customer Opinions for Sentiment Classification. *International Journal For Multidisciplinary Research,* 5(5), https://doi.org/10.36948/ijfmr.2023.v05i05.6090
9. Colón-Ruiz, C., Segura-Bedmar, I. (2020) Comparing deep learning architectures for sentiment analysis on drug reviews. *J. Biomed. Informatics,* 110, 103539, https://doi.org/10.1016/j.jbi.2020.103539
10. Mathew, L., Bindu, V. (2020) A review of natural language processing techniques for sentiment analysis using pre-trained models. *2020 Fourth international conference on computing methodologies and communication (ICCMC). IEEE*. 340345.
11. Sahar, A., Ayoub, M., Hussain, S., Yu, Y., Khan, A. (2022) Transfer learning-based framework for sentiment classification of cosmetics products reviews. *Pakistan Journal of Engineering and Technology*, 5(3), 38-43.
12. Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L., Trajanov, D. (2020) Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access*. 8, 131662–131682.
13. Wang, M.L. (2024) Fine-Tuning BERT for Sentiment Analysis Doctoral dissertation, UCLA. [Electronic resource]. - Mode of access: https://escholarship.org/uc/item/5kr0p5m
14. Qasim, R., Bangyal, W., Alqarni, M., Almazroi, A. (2022). A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. *Journal of Healthcare Engineering*, 2022(2), 1-17.
15. Panchenko, D., Maksymenko, D., Turuta, O., Luzan, M., Tytarenko, S., Turuta, O. (2022) Ukrainian News Corpus as Text Classification Benchmark. *International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications*, 1635, 550–559. https://doi.org/10.1007/978-3-031-14841-5_37
16. Deepa, M.D. (2021) Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task. *Turkish Journal of Computer and Mathematics Education*, 12(7), 1708-1721.
17. Sanh, V., Debut, L., Chaumond, J. Wolf, T. (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108
18. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019) Roberta: A robustly optimized bert pretraining approach. https://doi.org/10.48550/arXiv.1907.11692
19. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2019) Unsupervised cross-lingual representation learning at scale. https://doi.org/10.48550/arXiv.1911.02116
20. Ukrainian Roberta was trained by YouScan data science team. [Electronic resource]. - Mode of access: https://github.com/youscan/language-models
21. XLM-RoBERTa (base-sized model) [Electronic resource]. - Mode of access: https://huggingface.co/FacebookAI/xlm-roberta-base
22. Model Card for DistilBERT base multilingual (cased) [Electronic resource]. - Mode of access: https://huggingface.co/distilbert/distilbert-base-multilingual-cased#citation
23. BERT multilingual base model (cased) [Electronic resource]. - Mode of access: https://huggingface.co/google-bert/bert-base multilingual-cased.
24. Devlin, J., Chang, M., Lee, K., Toutanova, K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805
25. Ukrainian Roberta [Electronic resource]. - Mode of access: https://huggingface.co/google-bert/bert-base-multilingual-cased.

The article has been sent to the editors 10.06.24.
After processing 25.06.24.
Submitted for printing 28.06.24.