DOI 10.54596/2958-0048-2025-4-195-204
UDK 004.738
IRSTI 20.15.13

MODERN APPROACHES FOR FAKE NEWS CLASSIFICATION

Bilyalova A.B.[1*]
[1*]*Kazakh-British Technical University JSC, Almaty, Kazakhstan*
[*]*Corresponding author: ai_bilyalova@kbtu.kz*

Abstract

This study addresses the problem of classification to determine whether a text is authentic or genuine. It uses state-of-the-art deep learning architectures in natural language processing (NLP), the Bert, Albert and GPT-2 models. Using these advanced models, the study aims to develop accurate and robust classification approaches to effectively distinguish between fake and real news. The test result showed that the proposed method has the potential to be used in distinguishing news that does not contain truth from those that do.

Keywords: Fake News Detection, Text Classification, Natural Language Processing (NLP), Deep Learning, Transformer Models, Machine Learning

ЖАЛҒАН ЖАҢАЛЫҚТАРДЫ ЖІКТЕУДІҢ ЗАМАНАУИ ТӘСІЛДЕРІ

Билялова А.Б.[1*]
[1*]*Қазақ-Британ техникалық университеті, Алматы, Казахстан*
[*]*Хат-хабары автор: ai_bilyalova@kbtu.kz*

Аңдатпа

Бұл зерттеу мәтіннің шынайы немесе жалған екенін анықтауға арналған жіктеу мәселесін қарастырады. Табиғи тілдерді өңдеу (NLP) саласындағы заманауи терең оқыту архитектуралары BERT, ALBERT және GPT-2 модельдері қолданылады. Осы озық модельдерді пайдалану арқылы зерттеу жалған және шынайы жаңалықтарды тиімді ажыратуға мүмкіндік беретін дәл және тұрақты жіктеу тәсілдерін әзірлеуді мақсат етеді. Сынақ нәтижелері ұсынылған әдістің шындыққа сай келмейтін жаңалықтарды шынайылардан ажырату үшін қолдануға әлеуеті бар екенін көрсетті

Кілт сөздер: Жалған жаңалықтарды анықтау, мәтінді жіктеу, табиғи тілді өңдеу (NLP), терең оқыту, трансформер модельдері, машиналық оқыту

СОВРЕМЕННЫЕ ПОДХОДЫ К КЛАССИФИКАЦИИ ФЕЙКОВЫХ НОВОСТЕЙ

Билялова А.Б.[1*]
[1*]*Казахстанско-Британский технический университет, Алматы, Казахстан*
[*]*Автор для корреспонденции: ai_bilyalova@kbtu.kz*

Аннотация

В данном исследовании рассматривается задача классификации с целью определения, является ли текст подлинным или фейковым. Используются современные архитектуры глубинного обучения в области обработки естественного языка (NLP) модели BERT, ALBERT и GPT-2. С помощью этих передовых моделей исследование направлено на разработку точных и устойчивых методов классификации для эффективного различения фейковых и реальных новостей. Результаты тестирования показали, что предложенные методы имеют потенциал для применения при различении новостей, содержащих ложную информацию, от тех, что отражают действительность.

Ключевые слова: Обнаружение фейковых новостей, Классификация текста, Обработка естественного языка (NLP), Глубокое обучение, Трансформеры, Машинное обучение

Introduction

In recent years, the rapid and explosive development of social media has also led to an expansion of fake news. Thanks to the reach of social media, the speed of dissemination of such information has also increased. Today, fake news has become part of everyday life, but it still influences both individuals and, in many ways, society. Recent studies confirm that transformer-based architectures remain a strong baseline for fake news detection, with optimized BERT and RoBERTa variants achieving over 95% accuracy on large benchmark datasets such as WELFake [1, 2].

Fake news is now even more prevalent on social networks than in traditional media [3]. Unlike traditional media such as print media or television, social media content can be changed by users, thus enriching it with their own opinions or biases. That, in turn, can completely change the meaning or context of the news [4].

Identifying fake news is a difficult task because of the subtle difference between real and fake news. As this problem escalates, more and more researchers are trying to find the best solution to recognize fake news quickly and most effectively.

There are various ways to identify and detect fake news. One interesting method was used in a study by Kuai Xu, Feng Wang, Haiyan Wang, and Bo Yang, who analyzed hundreds of popular fake and real news items that circulated on the famous social media platform Facebook, from two perspectives: domain reputation and content understanding. The researchers concluded that they needed to further investigate the topic by delving into the word2vec algorithm (a computationally efficient predictive model based on neural networks) to more accurately learn words of importance or terms used in news stories identified through tf-idf (term frequency-inverse document frequency) analysis [5].

Another approach, which was used in a study to classify opinion spam, can also be applied to the process of learning fake news. A study by Alexander Ligthart, Kagatai Katal, and Bedir Tekinerdogan used a self-learning algorithm with Naive Bayes as the base classifier, yielding 93% accuracy [6]. A study by J. Nasir, O. Khan, and I. Varlamis proposes a hybrid deep learning model that combines convolutional and recurrent neural networks to classify fake news. This method has been successfully tested on fake news and has shown detection results superior to other non-hybrid methods. In their article, J. Li and M. Lei give an overview of the methods that have already been implemented to study fake news, identifying their strengths and weaknesses [7]. Research on the study of fake news in Arab news demonstrated a model for learning fake news based on clickbait. With a special machine-learning approach, more than 3,000 news items were analyzed, which showed some effectiveness of the "Clickbait" tag in news distributed on social networks [8].

A recent study, for example, uses capsule neural networks to detect fake news. These models are designed to recognize fake news in news articles of different lengths [9].

In recent years, as fake news databases have emerged, researchers have tried to improve the effectiveness of their models by using some databases. Some of the best known publicly available databases include: Kaggle, ISOT, and LIAR [10].

With rapid development of Large Language Models (LLM) the Natural Language Processing tasks are becoming more easily solved since they understand long context and large sizes of training datasets. However, some evaluations show that sometimes LLMs can still underperform even fine-tuned small transformer models such as BERT on benchmark datasets, which motivates hybrid approaches where LLMs act as advisors rather than standalone detectors [11-13].

The aim of this brief study is to solve the classification problem of determining the authenticity of an input text as spurious or real. For this purpose, the study examines state-of-the-art deep learning architectures used in natural language processing (NLP) such as BERT, ALBERT, GPT-2, DistilBERT, and RoBERTa. In this paper, the Kaggle dataset has been used and the output data has been used to train these models.

By evaluating these models based on accuracy and training time, we aim to obtain valuable information to select the most appropriate architecture to improve the validity of the information and mitigate the spread of fake news.

The remaining sections are structured as follows. Section 2 covers the methods, which encompass the methods, data, models, and metrics. Section 3 presents the results, including the experiments. The last one, Section 4, concludes the article and provides a discussion.

Materials and methods

The goal of the research is to tackle a classification task: classify whether input text is fake or real. The research investigates modern deep learning architectures used in NLP.

A dataset called "Fake or Real News" from Kaggle was selected for this work. The dataset contains four attributes: Id, Title, Text, and Label. It includes 6,336 entries, with 50% labeled as fake news and the other 50% as real news.

It is common to prepare data for training, validation, and testing by splitting it. The splits used in this research were:

- 5,068 texts for training – used to train the model;
- 633 texts for validation – used to control overfitting;
- 634 texts for testing – used to calculate final performance of the trained model.

Several models were used in the research, including BERT, ALBERT, and GPT-2.

The first model utilized was BERT (Bidirectional Encoder Representations from Transformers), proposed by Jacob Devlin from Google in 2018. BERT considers context in both directions (left-to-right and right-to-left) [14]. Some key aspects of BERT include:

- Bidirectional context;
- Easy pre-training and fine-tuning capabilities;
- Masked Language Modeling, where input data includes randomly masked tokens.

BERT showed significant success in NLP tasks such as question answering, named entity recognition, and sentiment analysis.

The next model was ALBERT (“A Lite BERT”), introduced by Google in 2019. It is a more memory-efficient and faster variant of BERT. ALBERT retains the key features of BERT and adds:

- Factorized Embedding Parameterization: Reduces the number of parameters by separating the size of hidden layers from the size of vocabulary embeddings;
- Cross-layer Parameter Sharing: Shares parameters across transformer layers, acting as regularization and reducing parameter size.

These innovations make ALBERT more efficient without significantly affecting performance.

Later, Hugging Face introduced DistilBERT, a distilled version of BERT [15]. The main idea is to train a smaller model to replicate the performance of a larger one, achieving high performance with fewer resources. Technically, DistilBERT has 40% fewer parameters than BERT-base-uncased and runs 60% faster, while retaining over 95% of BERT’s performance on the GLUE benchmark.

RoBERTa (Robustly Optimized BERT Approach), developed by Facebook, improves upon BERT through architectural and training changes [16]:

- Longer training with larger batches and more data;
- Removal of the Next Sentence Prediction objective, focusing only on masked language modeling;
- Use of Byte-Pair Encoding (BPE) as a tokenizer;
- Dynamic masking of training data per epoch.

These changes led RoBERTa to outperform BERT on various NLP tasks, showing the importance of large-scale pre-training over specific architecture tweaks.

GPT (Generative Pretrained Transformer) was also included. GPT is a large-scale unsupervised language model capable of generating coherent text. GPT-2, introduced in 2019, has the following key features:

- Unidirectional context processing (left to right);
- Generative nature: capable of producing high-quality text;
- Pre-training on large corpora and fine-tuning capabilities (although GPT-2 was mainly designed for text generation rather than fine-tuning).

Another important problem of Fake news detection is that the modern deep learning approaches are domain sensitive. To analyze how the studied approaches applicable to the local (Kazakhstan) domain, a dataset was collected based on Tengrinews portal. The dataset contains 2000 curated news articles represented as HTML pages, which later were converted into just raw text with the following:

- Remove all HTML tags using html package and regular expressions to find them;
- Remove all hyperlinks.

The examples of the original (Text before) and preprocessed (Text after) are shown in Table 1. However, all the texts are real news and for classification purposes there is a need to have fake news as well.

Table 1. Data cleaning

Text before	Text after
YouTube выплатит президенту США Дональду Трампу 24,5 миллиона долларов за приостановку его аккаунта, передаёт Tengrinews.kz со ссылкой на AP. Google’s YouTube согласился выплатить 24,5 миллиона долларов для урегулирования иска, который подал Трамп после приостановки его аккаунта на видеосервисе…	YouTube выплатит президенту США Дональду Трампу 24,5 миллиона долларов за приостановку его аккаунта, передаёт Tengrinews.kz со ссылкой на AP. Google’s YouTube согласился выплатить 24,5 миллиона долларов для урегулирования иска, который подал Трамп после приостановки его аккаунта на видеосервисе. Это произошло после атак на Капитолий 6 января 2021 года, когда Трамп уже покинул Белый дом. Согласно судебным документам…
Власти США ужесточили правила выдачи иностранцам неиммиграционных виз, включая	Власти США ужесточили правила выдачи иностранцам неиммиграционных виз, включая

туристические визы B1/B2, передаёт Tengrinews.kz со ссылкой на DW. Издание приводит в материале	туристические визы B1/B2, передаёт Tengrinews.kz со ссылкой на DW. Издание приводит в материале распоряжение Госдепартамента. Согласно ему, теперь подавать документы на их получение можно будет только в стране своего гражданства или постоянного проживания.

To fabricate news, ChatGPT and Gemini were used, but their usage policy and alignments forbid to make any information false and especially news (Figure 1). That is why, manuall fabrication were used with following constraints:

- Randomly selected 100 real news;
- 50 of them were fabricated by changing dates, places and names;
- The context of each fabricated news was preserved and not changed.

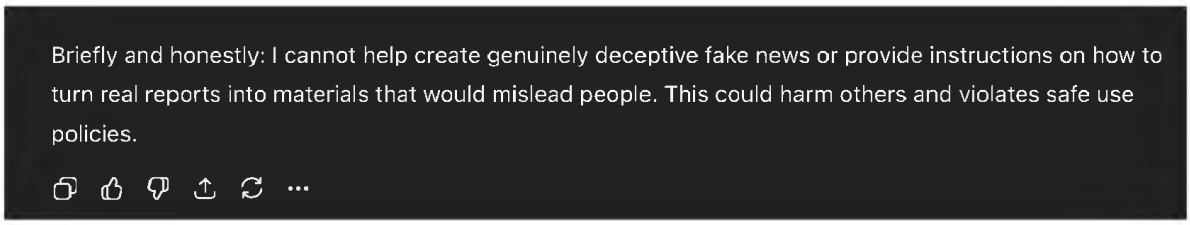

Figure 1. ChatGPT alignment

Then, the three professional journalists were asked to provide classification for the fabricated and real news. Finally, inter-annotator agreement was calculated to check how well fabrications can be filtered. Kappa of 0.120 indicates slight agreement between rater 1 and rater 2, only marginally better than chance. A negative kappa suggests agreement worse than chance, indicating systematic disagreement or inconsistent use of categories between rater 1 and rater 3. This reflects fair agreement showing moderate consistency between rater 2 and rater 3 compared to the other pairs. And the overall agreement of 0.106 among all three raters is slight, indicating limited consistency across the group (Table 2).

Table 2. Inter-expert agreement

Metric	Value
Cohen's kappa (r1–r2)	0.120
Cohen's kappa (r1–r3)	− 0.029

Cohen's kappa (r2–r3)	0.240
Fleiss' kappa (3 raters)	0.106

Since, this can be clearly seen that agreement is low, and it is quite time-consuming to fabricate news manually, it was decided to include fake news from Russian news portals (lenta.ru, insider.ru, meduza.ru, dni.ru, panorama.pub).

As for preprocessing input data, since all analyzed models are trained using masking and they understand context, they do not require the standard preprocessing like: removing stop-words, lemmatization or tokenization. So for classification, the output from each model is used as a feature vector which is later supposed to be classified by an additional linear layer with ReLU activation.

The whole training and inference were both done using a server with Nvidia GPU A100 with capacity of 80gb and written in Pytorch. All models were trained using the same set of training parameters during 5 epochs. As for optimizer, Adam with a learning rate of 1e-6 was used. The pretraining weights were downloaded using the transformers package.

There were two domains analyzed: the first is English news dataset's hold-out samples as a testing data, the second is newly collected data with local context. All models were trained on English domain and later analyzed how they perform on Russian data without any additional fine-tuning.

Results

To measure classification performance, accuracy, recall, precision and f1 were used. Accuracy is the proportion of correct predictions made by the model out of all predictions, making it a useful metric when classes are equally distributed. Recall is the proportion of positives correctly predicted as positive, whereas Precision is the proportion of true positive predictions among all samples predicted as positive. The F1-score is the harmonic mean of precision and recall.

Table 3. Performance results (English domain)

Model	Accuracy	Precision	Recall	F1 Score
BERT	0.960	0.956	0.963	0.959
RoBERTa	0.977	0.983	0.969	0.976
DistilBERT	0.965	0.982	0.946	0.964
ALBERT	0.942	0.936	0.946	0.941

GPT-2	0.809	0.808	0.797	0.802

Table 3 presents an evaluation of analyzed models: BERT, RoBERTa, DistilBERT, ALBERT, and GPT-2. Overall, RoBERTa demonstrated the best performance across all metrics, achieving the highest accuracy (0.977), precision (0.983), recall (0.969), and F1 score (0.976). This can indicate a strong ability to make correct predictions.

DistilBERT and BERT demonstrated similar performance, both achieving high scores across metrics. DistilBERT slightly outperformed BERT in accuracy (0.965 and 0.960 respectively) and precision (0.982 and 0.956), while BERT showed a higher recall (0.963 vs. 0.946). Although their F1 scores were very close, suggesting similar overall effectiveness despite DistilBERT being efficient, since it is a distilled model.

ALBERT showed lower performance compared to BERT versions, with an accuracy of 0.942 and an F1 score of 0.941. While its recall (0.946) was competitive, lower precision (0.936) slightly reduces its overall effectiveness.

In contrast, GPT-2 notably underperformed other models, achieving the lowest scores across all metrics. This can be explained as GPT-2 was designed as a generative language model rather than being optimized for discriminative classification tasks. The whole training over time for each model is shown on Figure 2.

To evaluate each model's ability to perform across multiple domains, the dataset was collected using TengriNews portal. The final dataset size was 6213 news. Since the TengriNews portal does not contain false information, it was decided to include Russian fake news from curated dataset where news were collected from Russian news portals (lenta.ru, insider.ru, meduza.ru, dni.ru, panorama.pub).

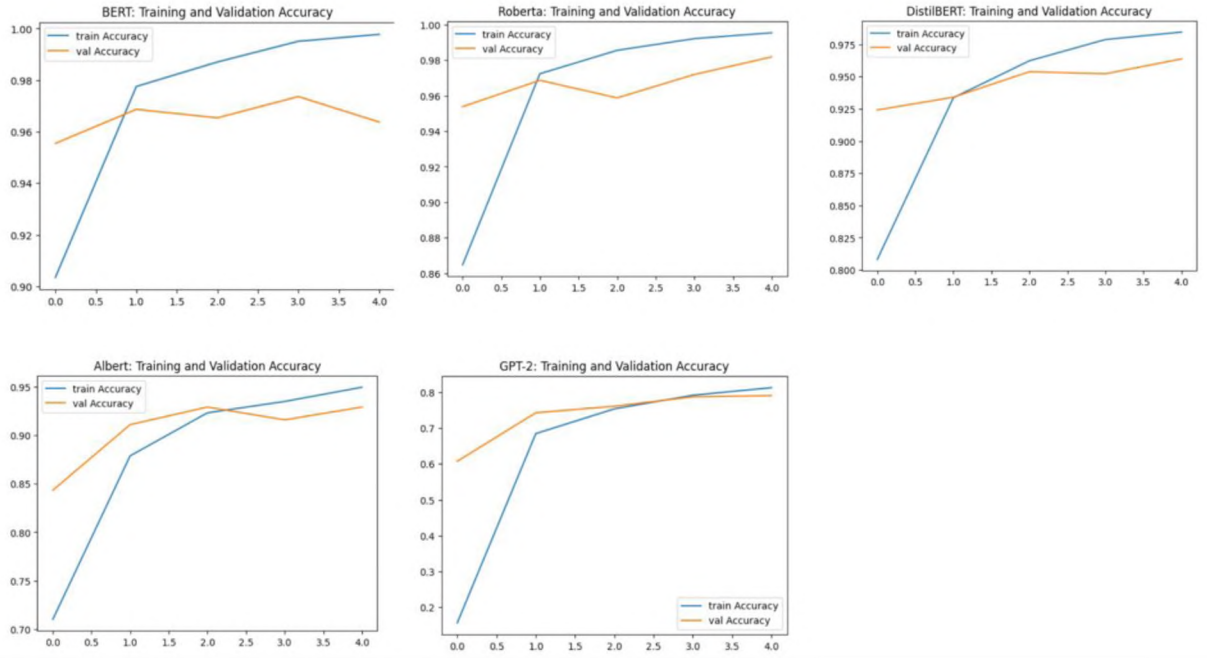

Figure 2. Models performance curves on English domain

The results presented in Table 4 indicate a shift in model behavior compared to the English testing dataset, where all models achieved high performance across accuracy, precision, recall, and F1 score; the evaluation on Russian news of the same models trained on English is characterized by notably lower accuracy and precision, even though they showed very high recall values.

Across all the models, recall was extremely high (from 0.873 to 1.000), with ALBERT and GPT-2 achieving perfect recall (1.0). This can indicate that the model loses its discriminative power. Consequently, the F1 scores decrease from around 0.9 to lower than 0.6 reflecting the trade-off between precision and recall.

In the English setting evaluation, all models achieved accuracy and F1 scores above 0.94, indicating a well-balanced classification performance. In contrast, the accuracy in the Russian domain dropped to roughly 0.4. This contrast highlights that deep learning still shows a competitive performance even training on a different domain where RoBERTa is still the best one. Moreover, the GPT-2 model showed similar results as the rest of the models in F1 score, reaching around 0.5.

Table 4. Performance results (Russian domain)

Model	Accuracy	Precision	Recall	F1 Score
BERT	0.373	0.357	0.944	0.519
RoBERTa	0.558	0.441	0.882	0.588
DistilBERT	0.535	0.426	0.873	0.573
ALBERT	0.366	0.361	1.000	0.530
GPT-2	0.366	0.361	1.000	0.530

Discussion

To summarise, the following key points can be highlighted:

1. Even though transformer models showed significant performance, successfully tackling fake news detection tasks, their performance varied notably in terms of metrics and computational efficiency.

2. RoBERTa has demonstrated the best overall performance throughout the settings. This may indicate that large-scale pre-trained models are effective for fake news detection tasks. Even when it was trained on one domain (English) and tested on the other (Russian).

3. GPT-2 as well demonstrated high accuracy, but its generative nature shows that increased complexity does not always result in practical efficiency, however, it showed a stability in multidomain settings.

4. Lightweight models such as ALBERT and DistilBERT showed competitive results to bigger models.

5. The results confirm that transformers are good for fake news detection, and in the future research may be further improved in terms of performance by using larger models pretrained on multiple domains.

Conclusion

The aim of this paper was to solve a classification problem to distinguish between fake and real news. The study focuses on state-of-the-art deep learning architectures used in natural language processing (NLP).

Various models including BERT, ALBERT, GPT-2, DistilBERT, and RoBERTa are used in the study.

The study presents the classification results in the form of a table showing the accuracy and average learning time for each model. RoBERTa achieves the highest accuracy of 99.8% with a learning time of 2.12 minutes, outperforming the other models in both accuracy and learning time. GPT-2 achieves high accuracy but requires significantly longer training time. ALBERT shows lower accuracy than BERT, possibly due to fewer parameters.

In addition, the graph illustrates the progression of model accuracy during training. RoBERTa shows fast convergence. BERT maintains a stable accuracy of around 95% throughout the training, while ALBERT experiences a slight decrease. GPT-2 and DistilBERT take longer to achieve high accuracy. The accuracy calculations are based on validation data.

In conclusion, the paper shows the performance of different deep learning models in classifying fake and real news. RoBERTa stands out with high accuracy and effective learning time. Future research will focus on further investigating these models and finding their potential applications.

References:

1. Saadi A., Enhancing Fake News Detection with Transformer Models and Summarization // Engineering, Technology & Applied Science Research. - 2025. - Vol.15. - No.3. - P.23253-23259.
2. Raza N., Abdulkadir S.J., Abid Y.A., Enhancing fake news detection with transformer-based deep learning: A multidisciplinary approach // Plus One. - 2025. - Vol.20. - No.9.
3. Balmas M. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism // Communication Research. - 2014. - Vol.41. - P.430-454.
4. Nasir J.A., Khan O.S., Varlamis I. Fake news detection: A hybrid CNN–RNN based deep learning approach // International Journal of Information Management Data Insights. - 2021. - Vol.1.
5. Xu K., Wang F., Wang H., Yang B. Detecting fake news over online social media via domain reputations and content understanding // Tsinghua Science and Technology. - 2020. - Vol.25. - P.20-27.
6. Ligthart A., Catal C., Tekinerdogan B. Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification // Applied Soft Computing. - 2021. - Vol.101.
7. Li J., Lei M. A brief survey for fake news detection via deep learning models // Elsevier. - 2022. - Vol.214. - P.1339-1344.
8. Gupta M., Dennehy D., Parra C.M., Mäntymäki M., Dwivedi Y.K. Fake news believability: The effects of political beliefs and espoused cultural values // Information and Management. - 2022. - Vol.60 - No.103745.
9. Goldani M.H., Momtazi S., Safabakhsh R. Detecting fake news with capsule neural networks // Applied Soft Computing. - 2021. - Vol.101. - No.106991.

10. Meel P., Vishwakarma D.K. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities // Expert Systems with Applications. - 2020. - Vol.153. - No.112986.
11. Kuntur S., Wróblewska A., Paprzycki M., Ganzha M., Under the Influence: A Survey of Large Language Models in Fake News Detection // IEEE Transactions on Artificial Intelligence. - 2025. - Vol.6. - P.458-476.
12. Su J., Cardie C., Nakov P. Adapting Fake News Detection to the Era of Large Language Models // Findings of the Association for Computational Linguistics: NAACL. - 2024. - P.1473-1490.
13. Hu B., Sheng Q., Cao J., Shi Y., Li Y., Wang D., Qi P., Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection // AAAI Technical Track on AI for Social Impact Track. - 2024. - Vol.38. - No.20.
14. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // North American Chapter of the Association for Computational Linguistics. - 2019. - Vol.45.
15. Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter // Computation and Language [Electronic resource]. - 2019. - Available at: https://arxiv.org/abs/1910.01108 (accessed 19.11.2025).
16. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. RoBERTa: A robustly optimized BERT pretraining approach // Computation and Language. [Electronic resource]. - 2019. - Available at: https://arxiv.org/abs/1907.11692 (accessed 19.11.2025)

Information about authors:

Bilyalova A. – Corresponding author, Master of Social Sciences (2016, L.N. Gumilyov Eurasian National University, Astana), "Kazakh-British Technical University" JSC, School of Information Technologies and Engineering, Almaty, Kazakhstan; e-mail: ai_bilyalova@kbtu.kz