Topical Text Classification of Russian News: a Comparison of BERT and Standard Models

Ksenia Lagutina
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
ksenia.lagutina@fruct.org

***Abstract*—The paper is devoted to the single-label topical classification of Russian news. The author compares the BERT features and standard character, word and structure-level features as text models. Experiments with OpenCorpora and eight news topics show that the BERT model is superior to standard ones, and achieves good classification quality for a small dataset of long news. Error analysis reveals the best classified topics: "economics", "culture", and "media". Comparison with the state-of-the-art research allows to consider BERT as a baseline for future investigations of analysis of texts in Russian.**

I. Introduction

Topical news classification is the task of news categorization into several thematic classes. Its subtask, single-label news classification, has many solutions based on standard text features and machine learning techniques [1].

Nevertheless, most of methods show significantly different results for different languages and datasets [1], [2], so they can not be acknowledged as universal approaches. However, in recent years the BERT language model has appeared, and it claims to be the method that can solve with high quality several text processing tasks, including news classification [3]. BERT has already proved its usefulness in English news analysis for different text datasets [4], [5].

The BERT model is also adapted for the Russian language [6]. But it remains under-researched for different tasks and datasets and almost does not accompanied by error analysis.

The goal of this research is to compare the BERT language model with standard statistical text models of character, word, and structure levels for the single-level topical classification of the Russian news. Standard models are based on the letters and punctuation marks occurrences, word and parts-of-speech n-grams. The text dataset consists of the Russian news from OpenCorpora. The subtask of this investigation is the error analysis in order to interpret classification results.

The paper is structured as follows. Section II describe the state-of-the-art in topical classification with BERT and other popular text features. Section III explains the methodology of the research: creation of the custom text dataset, extraction of features, and design of experiments. Section IV reveals results of experiments with different topic numbers and text models. Conclusion summarizes the paper and propose the directions of future studies.

II. State-of-the-art

In modern research the BERT is one of the most popular and efficient language model for topical classification of English texts [3]. It is based on the multi-layer bidirectional Transformer architecture and pre-trained on BookCorpus and Wikipedia. The BERT model outperforms other Transformer neural networks and standard methods in short news single-label classification [2].

Pappagari et al. [5] adapt the BERT model to fine-tune it for long texts. They split a text into fragments, compute BERT representations for each of them, stack representations into a sequence for the text, and apply Transformer layer to get a text embedding. This method called ToBERT, achieves 85 % in topical classification of long news from the 20-Newsgroups dataset.

Ye et al. [7] combine BERT with graph-based models. They get very high accuracy of 98 % for short news, but biomedical topic classification is performed significantly lower, 74 % of accuracy.

González-Carvajal and Garrido-Merchán [8] compare BERT with standard TF-IDF features in classification of English and Portuguese texts. They confirm high results of BERT model for English language (83–93 % of accuracy) and achieve good accuracy of 91 % for Portuguese news. The authors suggest to make BERT the default model for the natural language processing tasks.

The BERT model became popular not only for English, but also for national languages. Kuratov et al. [9] adapt it for Russian. They pre-train the BERT Transformer neural network on Russian-language Wikipedia and news and apply it to paraphrase identification, sentiment analysis, and question answering tasks. In all cases the new RuBERT model is better than standard machine learning approaches, neural network classifiers, and even multilingual BERT by 3–9 % of F-measure and accuracy.

Glazkova [10] classifies biographical text fragments by 10 topics. The model based on RuBERT embeddings, achieves highest results of 93 % F-measure and outperforms multilingual BERT, word2vec, and the standard TF-IDF approach combined with SVM. The author analyse errors and concludes that misclassified fragments in most cases are topically related to more than one category.

Vychegzhanin et al. [11] analyse different machine learning techniques in the Russian news classification. The TF-IDF

model is considered as a baseline. The authors classify the custom dataset of news from the Internet portals into six topics: accidents, culture, economics, politics, society, sports. RuBERT is compared with Logistic Regression, Light Gradient Boosted Machine, k-Nearest Neighbors, Random Forest, Naïve Bayes, and Support Vector Machine classifiers. It shows 88 % of F-measure, Support Vector Machine — 87 % of F-measure, while other methods achieve 78–85 %. RuBERT outperforms other approaches in quality, but not by high value.

Thus, the BERT language model suits for Russian news classification. The best results are shown by the RuBERT adaptation of the original neural network.

Besides, the BERT embeddings can be successfully used not only for short fragments, but also for long texts, and achieve the same good results [5], [10].

Nevertheless, the standard features became the good model for the single-label topical classification.

Pittaras et al. [12] apply word2vec, WordNet, and semantic features to generate semantic vectors for words and combine them into text embeddings. Then they are used for classification of English news from the 20-Newsgroups and Reuters-21578 datasets. The classifier is DNN. The accuracy and F-measure are about 75–80 %. Error analysis shows that adding semantic information significantly improves classification quality.

Wang [13] compares different neural network models: CNN, DNN, LSTM, GRU, and disconnected RNN (DRNN) and GRU (DGRU) in English news classification. Experiments on large-scale datasets show low error rates of 1–5 % for AG news and Wikipedia articles. The authors conduct a small error analysis that provide examples where DGRU performs better.

Tellez et al. [14] create language-independent framework for different text categorization tasks and experiment with it for English, Spanish, Portuguese, and several other languages. Topical news classification is performed for English and Portuguese, text features include word and character n-grams, skip-grams, and TF-IDF, the classifier is SVM. For English results are significantly better than for Portuguese: 67–96 % against 57 % of F-measure. The classification quality is varied very much for different datasets even of the same language.

Romanov et al. [15] classify short Russian scientific texts by 15 topics. As the text model they apply word2vec, as classifiers — Logistic Regression, Random Forest, Support Vector Machine, and LSTM. F-measure is 60–70 %, the best results are achieved by the LSTM neural network. The accuracy is very high, 88–95 %. Such results are relatively good, since there is the quite big number of topics.

Zhang et al. [16] propose the TextING tool that builds individual graphs for each text and applies Graph Neural Networks. The authors use keywords and relationships between them to compute features. The accuracy for the topical classification of English news from the Reuters dataset achieves 95–98 %.

To sum up, standard features show very unstable results: quality of text classification significantly varies for different languages and different text datasets. The BERT model, on the contrary, shows quite good results in various cases.

Most of researchers experiment with different features, classifiers, hyperparameters of the method, and other technical details, but do not analyse errors. Although such analysis can help to understand the domain better and improve an approach.

Another lack in modern research lies in the fact that the BERT model is rarely combined with other features, especially for Russian news. So concatenations of BERT with standard features remain understudied.

Besides, investigations devoted to topical text classification are often experiment only with the same English-language text datasets: AG, Reuters, 20-Newsgroups. Analysis of Russian-language topics requires creation of the custom text corpora.

III. Methodology

A. Overview

The methodology of the research follows the rules of modern computational linguistics.

There is no dataset that has been unconditionally used by a large number of investigators to classify Russian-language news. Therefore, it was decided to collect a dataset based on OpenCorpora [17] that is popular in news analysis.

The text models include BERT and three standard levels of features: character, word, and structure-based ones. Standard features have often been used to classify texts into different categories over the years, so they can be considered as the baseline. The BERT language model is relatively new and needs the investigation under various conditions.

The experiment design is based on state-of-the-art techniques. It is organized as multiclass single-label classification for different number of classes and text features, including features combinations.

Let's discuss the methodology in more details.

B. The text dataset

The dataset for news categorization is based on the OpenCorpora text corpus [17]. OpenCorpora contains blogs, news from online media and Russian Wikipedia, fiction, legal texts, etc. It does not have certain categories that the corpus creators established for topical classification, but its news are marked with tags.

So I have chosen eight topics: "media", "sport", "science and technologies", "culture", "politics", "economics", "society", "health". These categories are based on tags of the same name. "Science and technologies" also includes tags "technologies", "space", and "Internet"; "culture" includes the tag "books"; "politics" includes the tag "in the world". Such additional tags correspond to the small number of texts and are used to expand big topics. When the text by tags can be matched with several topics, it is classified manually to the one category.

Statistics for the text topics is presented in the Table I. It contains the minimum, maximum, median and average number of words in the texts on the specified topic.

TABLE I. TOTAL WORD NUMBERS AND TOPIC SIZES IN THE TEXT CORPUS

Topic	Text number	Min	Median	Mean	Max
Media	155	78	190	345.4	6 720
Sport	63	59	211	354.1	4 932
Science and technologies	273	50	205	314.6	4 588
Culture	256	54	273	641.4	4341
Politics	321	37	216	396.3	21 417
Economics	187	67	191	262.2	1 572
Society	295	50	241	447.3	2 271
Health	92	78	194.5	418.2	7 174
All	1 642	37	216	409.6	21 417

Topics differ by size, six of them contain more than 150 texts. Most of texts are less than 300 words, but the dataset contain a significant number of long texts. The mean number of words in the text is 409.6. The dataset is small: it contains 1 642 texts in total.

C. Text models

Texts are presented as numerical features in four ways.

- As BERT features. Text embeddings are extracted using the RuBERT cased model from the DeepPavlov library [6]. BERT has a limitation that a text can not be more than 512 tokens, but many texts from the corpus are longer. To cope with this problem and take into account all text data, I divide texts into paragraphs, find mean RuBERT embeddings for each paragraph, and compute the mean of all paragraph embeddings. So each texts is represented as 768 BERT features.
- As character-based features. Character- and word-based features are taken from my previous research [18], where these text models show good results in authorship verification. Character-based features include average sentence length in character, frequencies of occurrences of each letter among all letters, and frequencies of occurrences of each punctuation mark among all punctuation in the text.
- As word-based features. They include average sentence length in words, average word length in characters, and frequencies of occurrences of unigrams, bigrams, and trigrams among top-40 n-grams (the 40 most frequent unigrams, the 40 most frequent bigrams, and the 40 most frequent trigrams are considered).
- As structure-based features. They include frequencies of occurrences of top-40 PoS-n-grams, $1 \leq n \leq 4, n \in \mathbb{N}$. The method of their computation is taken from the research [19].

BERT features are the state-of-the-art model in text classification, especially in topical classification. Other feature types describe character, word, and parts-of-speech levels of the text, and they have been the standard features in natural language processing for many years.

D. Design of experiments

The design of experiments with topical classification is showed in Fig. 1.

Firstly, texts (III-B) are presented as features (III-C). Feature vectors with different feature types for one text can be concatenated in order to evaluate combined text models.

Secondly, texts are classified into eight topics. The neural network classifier has the Bidirectional LSTM layer with 64 units and a dense output layer with the softmax activation function. The loss function is categorical cross-entropy, the optimization algorithm is Adam, the number of epochs is 100. The LSTM neural network is one of the best classifier in natural language processing tasks, including classification with BERT [3]. In order to estimate the stability of classification, I apply the five-fold cross-validation technique: 80 % of texts are the training samples, 20 % are the test ones.

Thirdly, classification results are evaluated with three standard measures: precision, recall, and F-measure, and also their standard deviations as stability measures.

Finally, I analyse misclassified texts and find the topic with the most number of errors. This topic is excluded from the experiment, and classification is repeated. In such a way there is determined topics that are better separated from each other.

The code for the feature extraction and topical classification is written in Python programming language and uses Stanza 1.3.0 NLP library for text representation and determination of parts of speech. For the BERT embeddings it uses DeepPavlov 0.17.2, for the classification — Keras 2.7.0.

IV. EXPERIMENTS WITH TOPICAL CLASSIFICATION

A. Classification into eight topics

The topical classification is organized as multiclass single-label classification into eight news categories: “media”, “sport”, “science and technologies”, “culture”, “politics”, “economics”, “society”, and “health”. The results of these experiments are presented in the Table II. The first column describes the applied text models or their combination: BERT, character, word, or structure-level features. The columns “Std. dev.” contain the standard deviations of the measure from the adjacent left column.

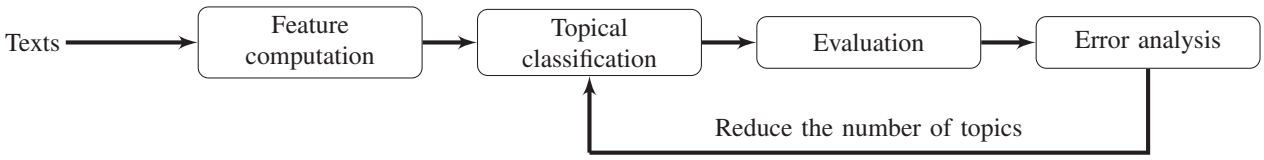

Fig. 1. Topical classification experiments

TABLE II. MULTICLASS TOPICAL CLASSIFICATION OF NEWS: CROSS-VALIDATION WITH LSTM

Text model	Precision	Std. dev.	Recall	Std. dev.	F-measure	Std. dev.
Char	37.9	0.1	38.4	1.6	37.8	0.6
Struct	30.5	2.0	29.7	2.0	29.7	2.0
Word	31.3	2.6	30.0	1.2	30.1	1.7
Char+Word+Struct	47.2	1.9	43.5	1.7	44.2	1.0
BERT	**78.9**	2.6	**78.2**	1.9	**78.3**	2.4
BERT+Word	**79.3**	0.5	**78.1**	0.2	**78.6**	0.2
BERT+Char	**79.0**	0.8	**78.5**	0.9	**78.4**	0.1
BERT+Struct	**78.9**	0.6	77.3	1.6	77.6	0.6
BERT+Char+Word+Struct	77.4	0.3	76.7	1.7	76.6	0.9

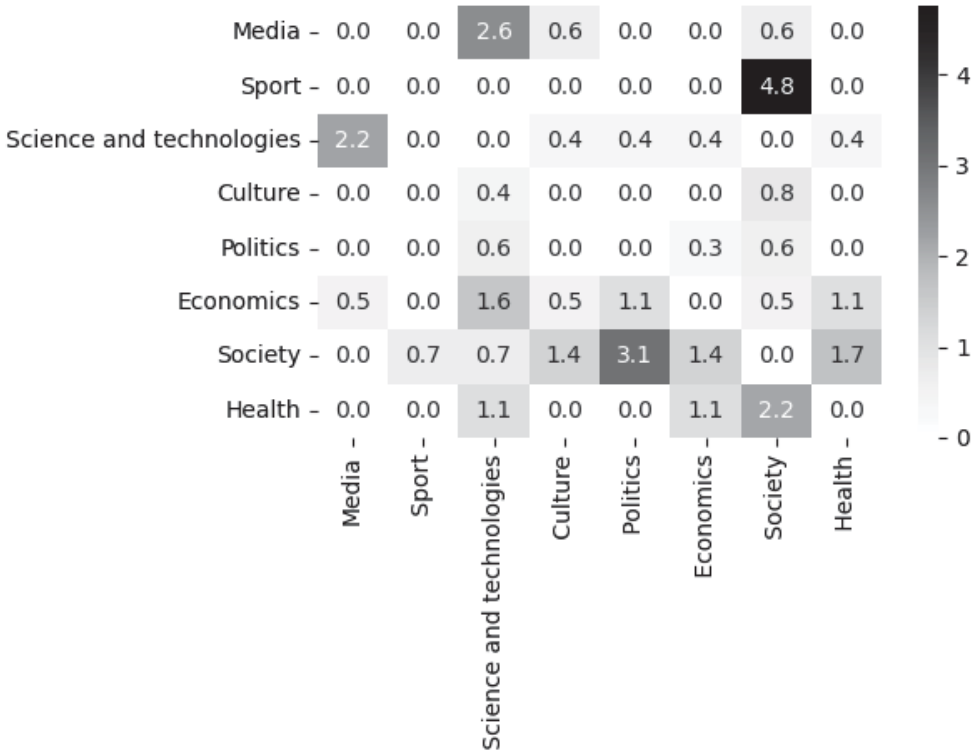

Fig. 2. The confusion matrix in percentages for the classification into eight news topic with BERT features and the LSTM classifier

TABLE III. MULTICLASS SINGLE-LABEL TOPICAL CLASSIFICATION OF NEWS FOR DIFFERENT TOPIC NUMBERS

Topic number	Text model	Precision	Std. dev.	Recall	Std. dev.	F-measure	Std. dev.
8	BERT	78.9	2.6	78.2	1.9	**78.3**	2.4
8	Char+Word+Struct	47.2	1.9	43.5	1.7	44.2	1.0
8	BERT+Char+Word+Struct	77.4	0.3	76.7	1.7	76.6	0.9
7	BERT	85.9	1.0	84.3	0.7	**84.7**	0.8
7	Char+Word+Struct	50.6	4.1	47.4	3.1	48.3	3.4
7	BERT+Char+Word+Struct	84.8	0.6	82.8	0.9	83.4	0.8
6	BERT	90.1	1.3	90.0	2.5	89.9	1.9
6	Char+Word+Struct	56.7	4.1	53.1	1.4	53.9	2.2
6	BERT+Char+Word+Struct	90.7	2.1	89.7	1.6	**90.1**	1.8
5	BERT	91.4	1.5	91.9	2.2	**91.5**	1.9
5	Char+Word+Struct	64.0	2.0	60.4	0.8	61.5	1.1
5	BERT+Char+Word+Struct	91.6	3.0	90.6	2.4	91.0	2.7
4	BERT	94.4	2.7	95.1	2.6	**94.7**	2.6
4	Char+Word+Struct	66.1	3.2	63.5	2.8	64.3	2.8
4	BERT+Char+Word+Struct	93.9	2.8	94.3	2.1	94.0	2.4
3	BERT	97.0	0.9	97.8	0.4	**97.4**	0.7
3	Char+Word+Struct	73.0	5.1	72.1	3.3	71.5	3.6
3	BERT+Char+Word+Struct	95.4	1.7	96.1	0.9	95.7	1.3
2	BERT	97.7	0.5	97.0	1.4	97.3	1.0
2	Char+Word+Struct	84.7	6.7	85.0	6.4	84.6	6.7
2	BERT+Char+Word+Struct	98.5	1.0	98.9	0.8	**98.7**	0.9

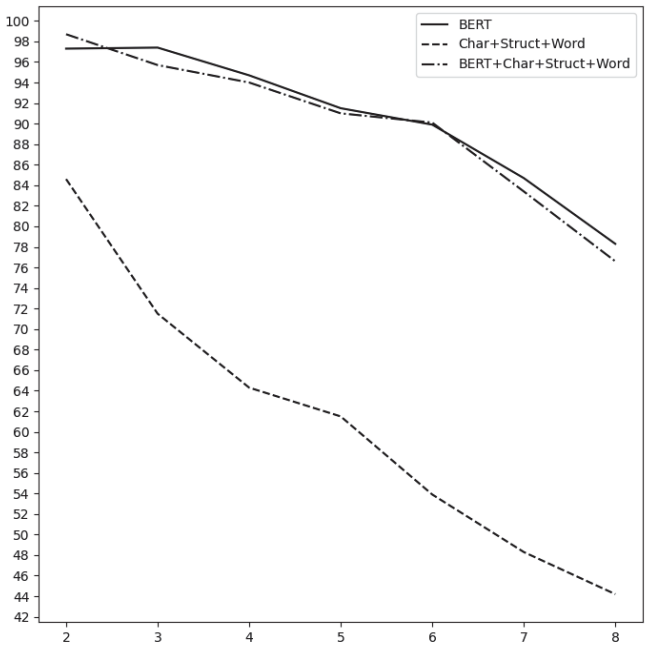

Fig. 3. Multiclass topical classification of news: F-measure dependence on the number of topics

We can see that BERT features significantly outperform all models: 78–79 % against less than 48 % of precision, recall, and F-measure. Combinations of three levels of standard features improves classification quality by 7–15 % of F-measure in comparison with single feature types. Combinations of standard features with the BERT model do not improve any results.

It is noteworthy that all text models are stable in classification quality. In every experiment the standard deviations of all measures are less than 3 percentage points.

Most probably, the low results of standard features mean that texts in the OpenCorpora dataset are very close by style. They are written in official informative language and have the similar lexicon and structure of phrases and word combinations. BERT features, evidently, grasp the finer differences between topics, because there is a large number of features — 768, but this language model is difficult to interpret.

B. Error analysis

If we extract misclassified texts from the experiment with BERT features, we compute the following confusion matrix:

Fig 2. Rows show false-positive errors in percentages of all texts from the row topics. This experiment is chosen for error analysis because of the best classification quality.

From the matrix we see that the couple of topics that are confused with each other most often in relative values, is "sport"–"society", 4.8 % of "sport" texts are classified as "society" ones. The second by value is "society"–"politics", 3.1 %. "Society" is the topic with the greatest percentage of false-positive and false-negative errors. "Sport" and "politics" are the best by false-positives in percentage. "Sport" is also good by false-negatives.

The topic "society" is the most scattered inside. It contains texts that describe many areas of human life and frequently mention health, economics, ets.

The topic "science and technologies" is also a set ot texts about very different areas of knowledge. Several texts are confused with the "media" because of the term "Internet" and close words. "Media" contains texts about the journalistic sphere, and in many cases they describe web portals.

It is interesting that topics "culture" and "politics" are also aggregators of different subtopics, but they are classified with relatively small number of errors.

If we sum the numbers of false-positive error percentages for each topic, the most misclassified is "society". Then we exclude it from the experiments, reclassify the dataset, and recompute the confusion matrix. Again, we can find the next topic, the worst by errors number, and repeat the experiment with less topics number. In such a way the topics are excluded in the following order:

1) "society";
2) "science and technologies";
3) "politics";
4) "economics";
5) "culture";
6) "media".

The last two topics for binary classification are "sport" and "health".

We can notice that the topic "science and technologies" remains the one with the big number of errors after removal of "society". "Media", on the contrary, become the one of the best classified after elimination of "science and technologies", and is the last excluded.

The classification quality in experiments with the change of the topic number is shown in Table III. There are experiments with the best results: BERT features and the combination of standard features, and also the tuple of all feature types in order to estimate the combination with BERT.

The change in the F-measure is more clearly depicted in the graph in Fig. 3. The x-axis shows the number of topics, the y-axis shows the value of the F-measure.

After removal of each topic the classification quality grows significantly by 3–6 percentage points for BERT. The exception is the last step with moving from 3 to 2 topics, when the F-measure value remains close to 97 %. In absolute values precision, recall, and F-measure become very high 90–97 % already with 6 topics or less. The standard deviations are very low in all cases.

The quality of classification with standard features improves better than with BERT: by 4–13 percentage points at each experimental step. When the BERT provides the fast growth in the early steps, the standard features achieves high results and good quality improvement only for 4–3–2 topics: 64–84 %. The last case with 2 topics is better by 13 percentage points of the F-measure then the previous with 3 topics, but it also has a significant standard deviation of 6.4–6.7 percentage points for all measures.

The combination of standard features with the BERT model does not provide quality improvement for any number of topics. The BERT features remain self-sufficient for the news topical classification.

To sum up, for the small number of narrow topics both BERT and standard models show high results. When the task is to classify several common topics, the BERT features significantly outperform character, word, and structure-level ones. So the BERT language model can be accepted as the baseline in Russian news classification because of its stability and great quality.

V. Conclusion

In this paper I compare the BERT language model and character, word, and structure-level features for single-label classification of Russian news from OpenCorpora. BERT features show the best quality of F-measure 78–98 % for 8–2 topics numbers. This model significantly outperforms standard ones on a small dataset.

Comparison of these results with the state-of-the-art allow to conclude that the BERT model shows very good classification quality for long and short texts, large and small datasets of Russian news. So it is stable in different cases, therefore, it can be acknowledged as a high-quality baseline for future research for Russian language.

Combinations of standard features with BERT do not show results improvement even for the small number of topics, although all feature types provide high F-measure for 2–3 news topics. Probably, BERT features can be more successfully combined with high-level features based on deeper aspects of linguistics and the subject area.

The error analysis shows the topics that are frequently confused with others: "society" and "science and technologies". The more deep linguistic investigation of these misclassified texts can be goal of the next research.

Another direction of the future investigations is comparison of BERT with various stylometric features in different natural language processing tasks: authorship verification, genre and style classification, etc.

Acknowledgment

The reported study was funded by RFBR, project number 20-37-90045.

REFERENCES

[1] R. Katari and M. B. Myneni, "A survey on news classification techniques," in *2020 International Conference on Computer Science, Engineering and Applications*. IEEE, 2020, pp. 1–5.

[2] K. Lagutina and N. Lagutina, "A survey of models for constructing text features to classify texts in natural language," in *29th Conference of Open Innovations Association FRUCT*. IEEE, 2021, pp. 222–233.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *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)*, 2019, pp. 4171–4186.

[4] K. S. Nugroho, A. Y. Sukmadewa, and N. Yudistira, "Large-scale news classification using bert language model: Spark nlp approach," in *6th International Conference on Sustainable Information Engineering and Technology 2021*, 2021, pp. 240–246.

[5] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak, "Hierarchical transformers for long document classification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 838–844.

[6] M. Burtsev, A. Seliverstov, R. Airapetyan, M. Arkhipov, D. Baymurzina, N. Bushkov, O. Gureenkova, T. Khakhulin, Y. Kuratov, D. Kuznetsov *et al.*, "Deeppavlov: Open-source library for dialogue systems," in *Proceedings of ACL 2018, System Demonstrations*, 2018, pp. 122–127.

[7] Z. Ye, G. Jiang, Y. Liu, Z. Li, and J. Yuan, "Document and word representations generated by graph convolutional network and bert for short text classification," in *ECAI 2020*. IOS Press, 2020, pp. 2275–2281.

[8] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," *arXiv preprint arXiv:2005.13012*, 2020.

[9] Y. Kuratov and M. Arkhipov, "Adaptation of deep bidirectional multilingual transformers for Russian language," *arXiv preprint arXiv:1905.07213*, 2019.

[10] A. V. Glazkova, "Topical classification of text fragments accounting for their nearest context," *Automation and Remote Control*, vol. 81, no. 12, pp. 2262–2276, 2020.

[11] S. Vychegzhanin, E. Kotelnikov, and V. Milov, "Comparative analysis of machine learning methods for news categorization in Russian," in *CEUR Workshop Proceedings*, 2021, pp. 100–108.

[12] N. Pittaras, G. Giannakopoulos, G. Papadakis, and V. Karkaletsis, "Text classification with semantically enriched word embeddings," *Natural Language Engineering*, vol. 27, no. 4, pp. 391–425, 2021.

[13] B. Wang, "Disconnected recurrent neural networks for text categorization," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2311–2320.

[14] E. S. Tellez, D. Moctezuma, S. Miranda-Jiménez, and M. Graff, "An automated text categorization framework based on hyperparameter optimization," *Knowledge-Based Systems*, vol. 149, pp. 110–123, 2018.

[15] A. Romanov, K. Lomotin, and E. Kozlova, "Application of natural language processing algorithms to the task of automatic classification of Russian scientific texts," *Data Science Journal*, vol. 18, no. 1, 2019.

[16] Y. Zhang, X. Yu, Z. Cui, S. Wu, Z. Wen, and L. Wang, "Every document owns its structure: Inductive text classification via graph neural networks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 334–339.

[17] V. Bocharov, S. Alexeeva, D. Granovsky, E. Protopopova, M. Stepanova, and A. Surikov, "Crowdsourcing morphological annotation," in *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Volume 1*, 2013, pp. 109–114.

[18] K. V. Lagutina, "Comparison of style features for the authorship verification of literary texts," *Modeling and Analysis of Information Systems*, vol. 28, no. 3, pp. 250–259, 2021.

[19] A. M. Manakhova and N. S. Lagutina, "Analysis of the impact of the stylometric characteristics of different levels for the verification of authors of the prose," *Modeling and Analysis of Information Systems*, vol. 28, no. 3, pp. 260–279, 2021, in Russian.