Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

Artificial text detection in Russian language: a BERT-based Approach

Posokhov P. A.
ITMO University,
Saint Petersburg, Russia
paposokhov@itmo.ru

Skrylnikov S. S.
ITMO University,
Saint Petersburg, Russia
skrylnikovs@itmo.ru

Makhnytkina O. V.
ITMO University,
Saint Petersburg, Russia
makhnytkina@itmo.ru

Abstract

This paper describes our solution for the RuATD (Russian Artificial Text Detection) competition held within the Dialog 2022 conference. Our approach is based on the idea of transfer learning, using pre-trained RuRoBERTa, RuBERT, RuGPT3, RuGPT2 models. The final solution included Byte-level Byte-Pair Encoding tokenization, and a fine-tuned model RuRoBERTa model. The system got Accuracy metric value of 0.65 and took first place in the multi-class classification task.

Keywords: artificial text detection, transfer learning
DOI: 10.28995/2075-7182-2022-21-470-476

Распознавание сгенерированных русскоязычных текстов на основе моделей BERT

Посохов П. А.
Университет ИТМО,
Санкт-Петербург, Россия
paposokhov@itmo.ru

Скрыльников С. С.
Университет ИТМО,
Санкт-Петербург, Россия
skrylnikovs@itmo.ru

Махныткина О. В.
Университет ИТМО,
Санкт-Петербург, Россия
makhnytkina@itmo.ru

Аннотация

В данной статье описано наше решение для соревнования по распознаванию сгенерированных текстов RuATD (Russian Artificial Text Detection), проводящегося в рамках конференции Диалог 2021. Наш подход был основан на идее трансферного обучения, использовались предобученные модели RuRoBERTa, RuBERT, RuGPT3, RuGPT2. Итоговое решение включало токенизацию Byte-level Byte-Pair Encoding, и дообученную модель RuRoBERTa. Система получила значение метрики Accuracy 0,65 и заняла первое место в задаче мультиклассовой классификации соревнования.

Ключевые слова: распознавание сгенерированного текста; перенос обучения

1 Introduction

Artificial text detection systems are being developed for a long time now, first of those were based on the logical linguistic approach and were usually rule-based. The development of such systems was a time-consuming process, besides, the generated texts had the same type, because they used certain patterns, the generated texts though were meaningful and syntactically correct [9]. Recently, some researchers still use the rules as components of automatic text generation systems [17]. Later development of text generation systems was based on statistical approaches such as Markov chains [14]. However, the result for such models can be unpredictable, semantic connections can be lost and sentences can be grammatically incorrect. The active development of neural networks gave automatic text generation a new life. The generation of meaningful texts that are grammatically correct and close to human-written texts became possible after the creation of neural network architectures based on transformers [16]. These models show impressive results, as Clark et al. [4] suggest the ability of non-specialists to distinguish between human and machine text (GPT2 and GPT3) in three areas (stories, news articles and recipes), and found out that without training evaluators can distinguish GPT3 generated text from human-written text purely by chance.

However, such models can also be used with different aims, for example, to create fake news [15,18], product and service reviews [1, 2]. For example, [18] shows that people rate model-generated disinformation as credible, even more than human-written disinformation. That is why the artificial text detection task is very relevant nowadays. Researchers have already made attempts to develop detection systems for artificial texts. The main approaches are: 1) training models from scratch, using the bag of words model and classical machine learning methods, such as logistic regression [6, 15]; 2) the use of pre-trained models based on transformers [1, 13, 18]. The second approach shows the best results in the artificial text detection. At the same time, it is worth noting, that research in this field was mainly carried out on datasets in English and Chinese languages. In recent years, the automatic text generation in Russian language has also reached a high quality, especially due to the emergence of pre-trained models ruGPT3, ruT5 [8], but the task of artificial text detection has not been given due attention. The article proposes a solution to this task and determines the model used to generate the text. The code is publicly available at https://github.com/Anpopaicoconat/dialog2022.

2 Task

The task set on RuATD (Russian Artificial Text Detection) is the multi-class classification of generated texts with generator model determination or assignment of Human class for cases when text is written by a person [11]. The list of response classes for this task contains the following:

- Human – text is written by a person;
- OPUS-MT – text is generated with machine translation model OPUS;
- ruGPT2-Large – text is generated with ruGPT2-Large model;
- ruGPT3-Large – text is generated with ruGPT3-Large model;
- ruGPT3-Medium – text is generated with ruGPT3-Medium model;
- ruGPT3-Small – text is generated with ruGPT3-Small model;
- M-BART – text is generated with Text2Text model M-BART;
- M-BART50 – text is generated with Text2Text model M-BART50;
- M2M-100 – text is generated with Text2Text model M2M-100;
- mT5-Large – text is generated with Text2Text model mT5-Large;
- mT5-Small – text is generated with Text2Text model mT5-Small;
- ruT5-Base – text is generated with Text2Text model ruT5-Base;
- ruT5-Base-Multitask – text is generated with Text2Text model ruT5-Base-Multitask;
- ruT5-Large – text is generated with Text2Text model ruT5-Large.

Initially, the task was to implement a multi-class classification. The input in this case is a text example, with an output being one of the 14 tags, containing the source of the text, being either title of the generation model or human.

Evaluation metric used for this task is accuracy, which is a standard metrics for classifier evaluation. It is the fraction of predictions the model got right.

3 Dataset

The dataset provided for the task contains 215,110 text «text»:«source», examples divided into training(129,066), test(64,533) and validation(21,511) sets. Training examples contain text, representing the statement from dialogue or chat (see Table 1).

Text	Class
Власти планируют закончить строительство аэропорта Сочи к 2018 году [Authorities plan to finish construction of Sochi airport by 2018].	ruGPT3-Large
Путин подписал указ об открытии музея Михаила Ивановича на Моховой улице в Москве[Putin signed a decree on the opening of the Mikhail Ivanovich Museum on Mokhovaya Street in Moscow].	mT5-Large
Мерелбеке — это муниципалитет, расположенный в бельгийской провинции Восточная Фландрия [Merelbeke is a municipality in the Belgian province of East Flanders]	Human
Вторая попытка привела к тому же результату [The second attempt had the same result].	OPUS-MT

Table 1. Examples from the training set are shown in the table

The data distribution by classes is shown in Figure 1. The classes in the data are unbalanced, however this split is explained with the ratio between model generated and human written texts. Therefore, 50% of the provided dataset contain examples of human written texts, the other half is accounted for text generation models, despite the dataset being unbalanced for 14 classes this split of samples is reasonable.

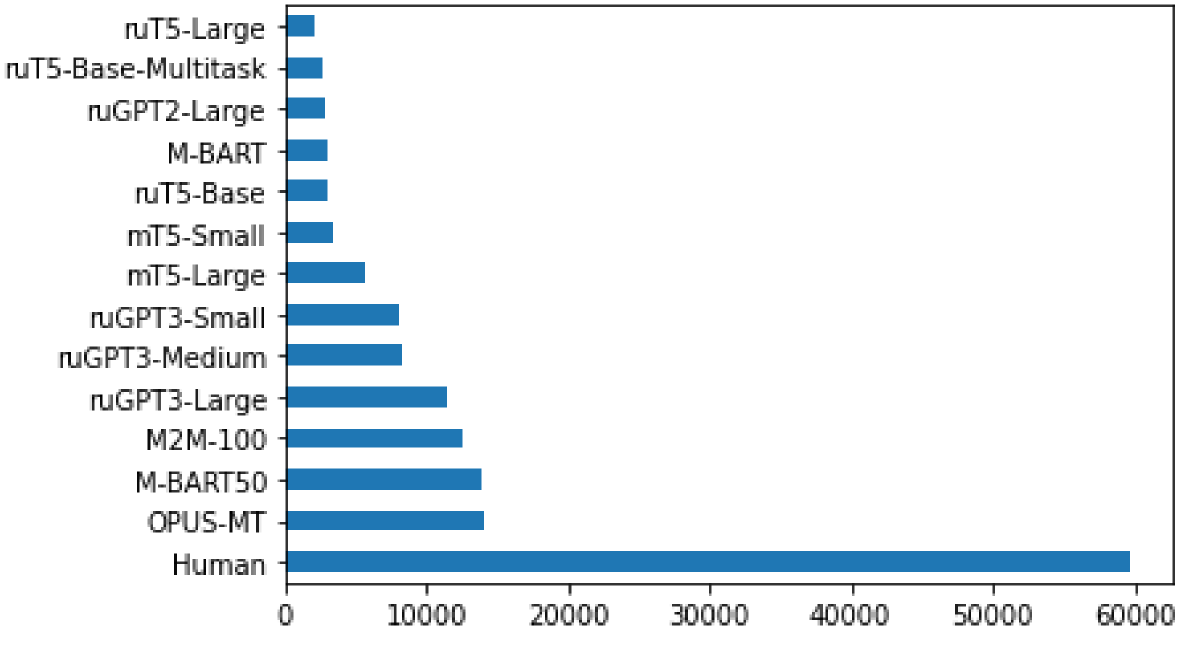

Figure 1. Data distribution by classes

4 Methods

Review of most recent works on this task allowed us to determine the most suitable models for its' solution in Russian language. The most commonly used technologies in natural language processing nowadays are transformer models. That is the reason why, the models chosen for solving this particular task also belong to this group of models. When implementing such models, usually the transfer learning is used. This is an approach in machine learning, when network knowledges from one task is transferred to solve another, related task. Text processing by such models is based on the process of tokenization. Word tokens are available from the network dictionary, and they represent words or their parts if full word is missing. The tokens themselves are encoded with embeddings, which are their vector representations, that are processed by the network in parallel, but they also save the information about the location of words in the sentence. Initially, such networks are trained on large datasets, after which they are pre-tuned for a specific task, which makes these models quite flexible. For solution of this task the following models were considered:

1. GPT2 - generative pre-trained transformer model often used in natural language processing tasks. It is based on the use of attention mechanisms, which allow the model to segment the input data and selectively focus on the most relevant one. This model surpasses the previous ones based on recurrent or convolutional neural networks, as it parallelizes computations much better. [10]. For this particular task, the rugpt2large model was chosen, pre-trained to work with the Russian language and available on the HuggingFace hub[1], with standard configuration and the number of un-freeze layers equal to 8. This model was trained on 170 gigabytes of data, representing 1024 long sequences. Dictionary size is 50257. The number of neurons in the output layer is 14, according to classes given.

2. GPT3. Unlike its' predecessor, it has more than 100 times more parameters [3]. The exact chosen model was rugpt3large_based_on_gpt2. This is the Russian language model pre-trained by SberDevices[2]. It has been trained on sequences of the same dimension. 80 billion tokens were used in training the first three epochs, after which the model was tuned to work with sequences of length 2048 for one more epoch. The output layer remained unchanged, however, the number of unfreeze layers was reduced to 4.

3. BERT (Bidirectional Encoder Representations from Transformers) this network was first introduced by Google and provided state-of-art results in many nlp tasks [5]. This model is primarily aimed at solving tasks that use the whole utterance such as sequence classification, token classification or question answering and requires fine-tuning for each specific use. For this task, the rubert-base-cased model was used, pre-trained by DeepPavlov[3] with 180 million parameters and a dictionary size of 119547. This model is the closest one to the model provided in the baseline solution of the task that is why it was chosen as a metric reference point for other models. The model standard parameters were chosen the output layer had 14 neurons, according to the number of possible response classes.

4. RoBERTa. Is a transformer model pre-trained on a large corpus of the raw texts only, without any labels with the MLM (Masked Language Modeling) objective. This model trains on masked sentences which is rather different from is different from traditional recurrent neural network (RNN) approaches that usually see the words in set order, or from other transformer models like GPT, which internally mask the future tokens. Roberta's approach allows the model to learn a bidirectional representation of the sentence. For this task we used RuRoBERTa-large model with the following parameters: number of epochs equals 4, batch size equals 1, learning rate equals 2e-5. The used optimizer was AdamW, based on it a linear scheduler with a warmup period was also used during which the learning rate increases linearly from zero to the initial one, and after that linearly decreases from the initial one set in the optimizer to 0. The warmup process is used in models with attention mechanisms to avoid the loss of weights the model learned during pre-training. Used model is an encoder, which was trained for the Russian language by the SberDevices team and available on the HuggingFace hub[4]. Its initial task is mask filling. In tokenization, Byte-level Byte-Pair Encoding is used [12]. This method allows model to have a smaller dictionary, with a larger number of options. In this case, the dictionary consists of 50,265 examples, the number of parameters is 335 million. When text is processed by the model, the first cls token is used for aggregation, which is followed by a dropout layer, with a probability of 0.1, this parameter is required to avoid model overfitting. The train data is fed to the input layer being fully connected with a dimension of 1024 neurons. Here the hyperbolic tangent (tanh) activation function is used, followed by a dropout again. The output layer contains 14 neurons, corresponding to the number of classes in the task. The use of RuRoBERTa-large, pre-trained by sberbank-ai, for this task is justified by the fact that it shows higher accuracy metrics when working with Russian text data in comparison with other models.

[1] https://huggingface.co/sberbank-ai/rugpt2large

[2] https://huggingface.co/sberbank-ai/rugpt3small_based_on_gpt2

[3] https://huggingface.co/DeepPavlov/rubert-base-cased

[4] https://huggingface.co/sberbank-ai/ruRoberta-large

5 Experiments

The competition rules included providing the baseline solutions for the artificial text detection task. The first one is based on the use of the "bag of words" method with the tf-idf measure and logistic regression, the second is based on the application of the BERT model after fine-tuning, more specifically, it uses pre-trained DeepPavlov rubert-base-cased available on the HuggingFace hub. The approaches used in the study were also based on the concept of transfer learning, such pre-trained models as RuRoBERTa, RuBERT, RuGPT3, RuGPT2 were used. The Table 2 below shows the results of artificial text detection.

Model	**Accuracy**
RuRoBERTa	0.65035
RuBERT	0.59817
Baseline BERT	0.59813
RuGPT3	0.54574
RuGPT2	0.47258
Baseline tf-idf	0.44280

Table 2. Models results on artificial text detection

RuGPT2 did not prove to be better than the basic solution in this task, therefore the decision was made to change the model. The RuGPT3 model showed higher accuracy than its predecessor, however, still insufficient in terms of the model applicability for the task. The RuBERT model predictably turned out to be on par with the fine-tuned model provided in the baseline. Still BERT based model showed better results according to GPT ones, as it considers both contexts of the word, whilst GPT models are based on the use of left context only. The best results were obtained with use of the RuRoBERTa model, it resulted in accuracy equal to 0.65035, on the test data of the competition, thus taking first place in the multi-class classification. This model's architecture and hyperparameters are optimized for best efficiency and further modifications would lower the efficiency of the model, without proper pre-training. For comparison, the accuracy of the baseline of the BERT and tf-idf solutions is 0.59813 and 0.44280, respectively.

The figure 2 shows confusion matrix heatmap normalized by number of examples in every class. The model classifies human-generated texts best of all classes, being accurate at 89% of examples. The confusion in classification of artificially generated text messages was mostly noticed among those generated by models designed primarily for machine translation, them being OPUS-MT, M-BART50 and M2M-100. Another notable remark is that the model also confuses the messages generated by the same architecture of different sizes. For example, 28% of the messages generated by mT5-Small are recognized as generated by mT5-Large; the confusion between the ruGPT3-Large, ruGPT3-Medium, ruGPT3-Small models varies from 7% to 15%. Significant error in classification occurs due to recognition of artificial texts as human class. The largest percent of wrongly classified messages were generated by mT5-Large, ruT5-Large models. This can be explained by the quality of text generation for these models.

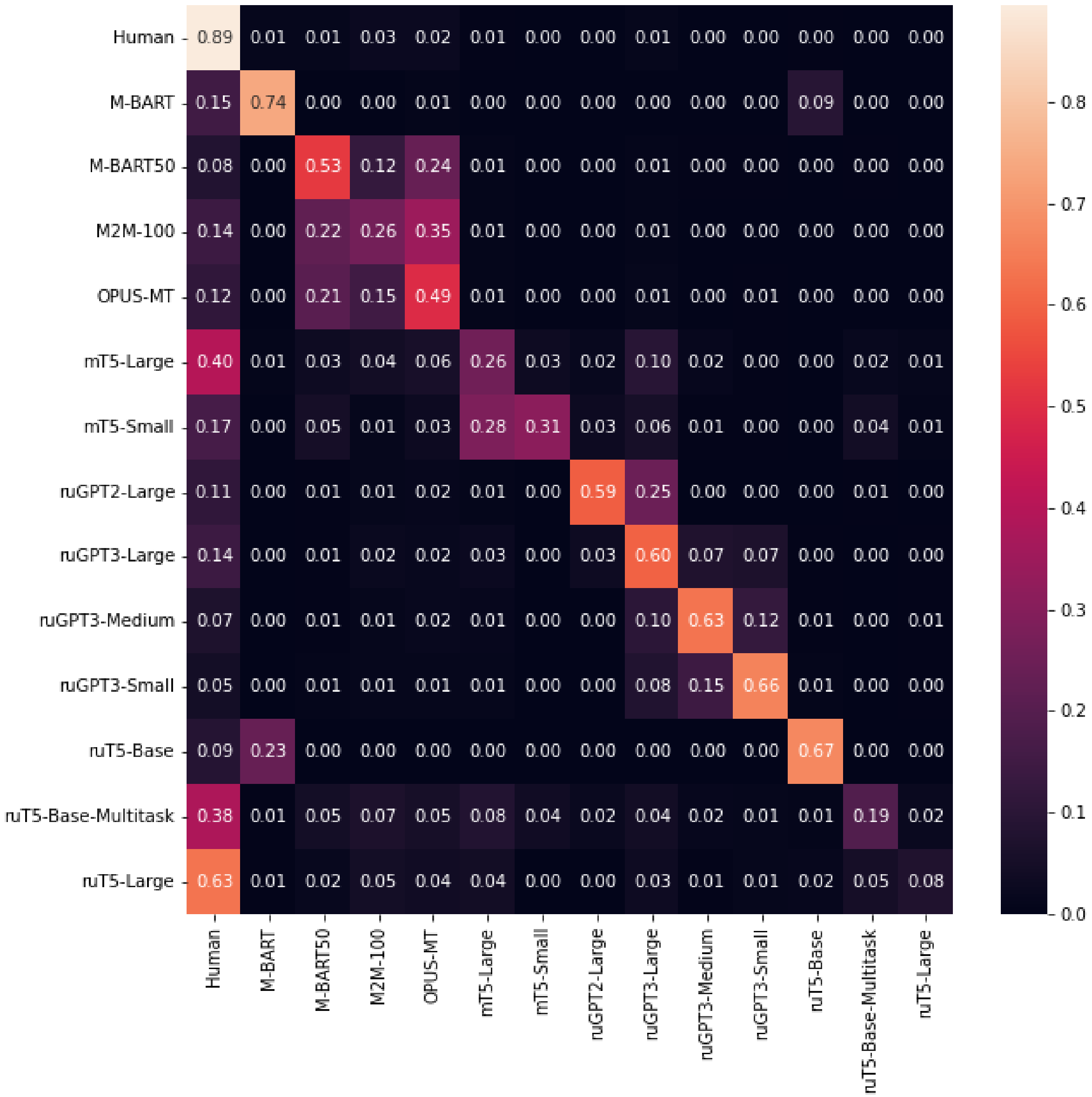

Figure 2. Heatmap of confusion matrix by classes

6 Conclusion

This paper presents the study of transfer learning usage for the artificial text detection task. Based on the analysis of modern research, it was concluded that for datasets in English and Chinese, this approach shows the best results this paper validates this statement also being true for Russian language datasets. A comparative analysis of the pre-trained Russian language models' usage showed the advantages of the RuRoBERTa model. The model presented in the article has several ways to be used, on the one hand, it can be used for its intended purpose to determine artificially generated messages, and to determine the exact model. On the other hand, it can be used to improve the quality of text generation, for example, as a filter for generated messages based on similarity between generated text and text written by a person.

Further development of the model, could possibly include training of separate classifiers corresponding to each applied model and use of output vectors built-in aggregation, thereby combining the models into an ensemble.

References

[1] Adelani, D., Mai H., Fang F., Nguyen H., Yamagishi J., Echizen I. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection// The 34-th In-ternational Conference on Advanced Information Networking and Applications. — Caserta, Italy, 2020. — 1341–1354.

[2] Bekmanova G., Sharipbay A., Omarbekova A., Yelibayeva G., Yergesh B. Adequate assessment of the customers actual reviews through comparing them with the fake ones // Proceedings of 2017 International Conference on Engineering and Technology — Antalya, Turkey, 2018. — P. 1-4

[3] Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T.J., Child R., Ramesh A., Ziegler D.M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei, D. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. — 2020 — Vol. 33. — P. 1877—1901.

[4] Clark E., August T., Serrano S., Haduong N., Gururangan S., Smith N.A. All that's 'human' is not gold: Evaluating human evaluation of generated text. // 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference — 2021. — P. 7282-7296.

[5] Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Minneapolis, MN, 2019. — Vol. arXiv:1810.04805. — version 2. Access mode: https://arxiv.org/abs/1810.04805.

[6] Ippolito D., Duckworth D., Callison-Burch C., Eck D. Automatic Detection of Generated Text is Easiest when Humans are Fooled // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — P. 1808–1822.

[7] Jawahar, G,. Abdul-Mageed, M., Lakshmanan, L. Automatic detection of machine generated text: A critical survey// Proceedings of the 28th International Conference on Computational Linguistics. — Barcelona, Spain, 2020. — P.2296–2309.

[8] Matveev A., Makhnytkina O., Matveev Y., Svischev A., Korobova P., Rybin A., Akulov A. Virtual Dialogue Assistant for Remote Exams // Mathematics — 2021. — Vol. 9, No. 18, 2229. — Access mode: https://doi.org/10.3390/math9182229.

[9] Mauldin M.L. Semantic rule based text generation//10th International Conference on Computational Linguistics, COLING 1984 and 22nd Annual Meeting of the Association for Computational Linguistics. — Stanford, CA , 1984. — P. 376–380.

[10] Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. Language Models are Unsupervised Multitask Learners // OpenAI Blog — 2019. — Access mode: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[11] Shamardina T., Mikhailov V., Chernianskii D., Fenogenova A., Saidov M., Valeeva A., Shavrina T., Smurov I., Tutubalina E., Artemova E. Findings of the The RuATD Shared Task 2022 on Artificial Text Detection in Russian//Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue. – 2022. – Vol.21.

[12] Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. — Berlin, Germany, 2016. — Vol. 1. — P. 1715–1725.

[13] Solaiman I., Brundage M., Clark J., Askell A., Herbert-Voss A., Wu A., Radford A., Krueger G., Kim J.W., Kreps S., McCain M., Newhouse A., Blazakis J., McGuffie K., Wang J. Release Strategies and the Social Impacts of Language Models — 2019. — arXiv:1908.09203. — version 2. Access mode: https://arxiv.org/abs/1908.09203

[14] Szymanski G., Ciota Z. On-line text generation using Markov Models// Proceedings of the International Conference Modern Problems of Radio Engineering, Telecommunications and Computer Science. — Lviv-Slavsko, Ukraine, 2004. — P. 339–341.

[15] Uchendu, A., Le, T., Shu, K., Lee D. Authorship attribution for neural text generation//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2020. — P. 8384–8395.

[16] Vaswani A., Shazeer N.M., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is All you Need // Advances in Neural Information Processing Systems. — Long Beach, CA, 2017. — P. 5999–6009.

[17] Vodolazova T., Lloret E. The impact of rule-based text generation on the quality of abstractive summaries // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — Varna, Bulgaria, 2019. — P. 1275–1284.

[18] Zellers R., Holtzman A., Rashkin H., Yonatan Bisk Y., Farhadi A., Roesner F., Choi Y. Defending Against Neural Fake News// Proceedings of the 33rd International Conference on Neural Information Processing Systems. — Vancouver, Canada, 2019. — P. 9054–9065.