ISSN 2007-9737

Bert for Classification of Russian Functional Styles

Valery D. Solovyev[1,*], Andrey M. Ten[2], Marina I. Solnyshkina[1], Mariia I. Andreeva[1]

[1] Kazan Federal University
Laboratory "Multidisciplinary Investigations of Texts", Kazan,
Russia

[2] Nobilis.Team,
Russia

maki.solovyev@mail.ru

Abstract. This paper tests the hypothesis that texts belonging to different functional styles possess distinct quantitative and linguistic parameters specific to each style. These parameters allow for quantitative classification using BERT. The research aims to develop a BERT classification model based on linguistic features of texts in five main functional styles: scientific, literary, official-business, journalistic, and colloquial. This approach addresses the problem of automatic classification of Russian functional styles based on statistical and morphological characteristics of texts. The selected hyperparameters for training the neural network include batch size, number of epochs, and initial learning rate. The study corpus comprises texts of the five abovementioned styles, totaling 163,421,783 tokens, sourced from the Russian National Corpus. The range of methods includes quantitative text analysis, morphological annotation, exhaustive analysis, and machine learning algorithms. The developed approach demonstrated high classification accuracy, indicating the promise of the proposed method. The results can be applied to tasks in automatic text processing, authorship attribution, and stylistic analysis. Future development includes classification models for various genres and domains, alternative transformer architectures (such as RoBERTa, GPT), larger datasets, and studying the impact of different fine-tuning strategies on classification quality.

Keywords. BERT, functional style, text classification, corpus linguistics, stylometry, automatic text analysis, statistical parameters, morphological annotation.

1 Introduction

The relevance of automatic profiling of texts, i.e., determining a text's style, genre, and type, is widely acknowledged and comprehensively presented in the published research (see for the overview Solnyshkina, Kupriyanov & Shoeva 2024). However, both classification algorithms and sets of discriminative features for different text types remain an open research niche (Melissourgou & Frantzi 2017).

The significance of this area of studies in general and research algorithm is obvious. Text profiling is important in computational linguistics (Kochetova & Popov 2019), and plays a crucial role across many areas, including linguistics, literary studies, marketing, and data analysis. Identifying a text's functional style allows for better understanding of its purpose, audience, and communicative features.

Automatic text classification is useful for tasks such as user review analysis, authorship verification, style adaptation, educational purposes, and even journalism. Classification models are vital for comparative literary and linguistic research (Murphy 2019), improving information extraction algorithms (Malhotra and Sharma 2017), and supporting machine translation tools (Dejica 2020).

Traditionally, functional style classification was performed manually by linguistic experts analyzing language features and structural elements (Lagutina, Lagutina, & Boychuk 2021). However, manual analysis becomes challenging and resource-intensive when dealing with massive datasets, making automated classification highly relevant (Solnyshkina, Solovyev& Ebzeeva 2024, Solovyev et al. 2024).

ISSN 2007-9737

Modern Natural Language Processing (NLP) technologies enable automation of this task with high accuracy. Among the most powerful tools in NLP is BERT (Bidirectional Encoder Representations from Transformers), developed by Google. It stands out due to its ability to consider extended of a context word, making it particularly effective for text classification tasks (Solovyev et al. 2024).

BERT and its best versions have already demonstrated high accuracy in tasks like sentiment analysis, semantic relationship detection, and thematic classification. In this study, we explore BERT's potential for identifying functional styles of Russian texts and compare its performance with traditional machine learning methods.

This paper specifically focuses on automating style classification, leveraging pre-trained models, adapting BERT for Russian-language text, comparing and contrasting BERT with classical classification methods.

Practical applications include but not limited to document analysis, content categorization, stylistic evaluation, and other NLP fields.

2 Related Works

2.1 Functional Styles as a Linguistic Category

The development of functional stylistics laid by Charles Bally (1909) were first catalyzed by the emergence of a functional approach to language and introduction of the term "functional style" in the early 20th century. The methodologies of functional stylistics were subsequently explored by members of the Prague Linguistic Circleand numerous publications in the mid-1950s. With the advent and progression of ethnolinguistics and sociolinguistics, the study of functional styles increasingly shifted toward a sociocultural framework, wherein styles were analyzed through the lens of the social functions of language (Gumperz 1982, Labov 2001).

O.Sirotinina (1993) delineates the concept of functional styles as distinct varieties of language use, each tailored to specific communicative purposes and social settings. She emphasizes that these styles are not rigid categories but dynamic systems influenced by various factors, including the speaker's intent, the audience, and the medium of communication. O. Sirotinina identifies five key functional styles, each characterized by unique linguistic features and serving particular societal functions. In recent decades, the computational modeling of texts across various genres—classified according to defined criteria and subjected to statistical analysis—has become closely aligned with the study of functional styles. Moreover, diachronic shifts in functional stylistic features, driven by temporal change, have prompted renewed scholarly interest in this domain (Moiseeva & Remizova, 2015).

According to O. Sirotinina, the primary functional styles in the Russian language include the following:

1. Scientific Style: Characterized by precise terminology, logical structure, and objective tone. It is used in academic and technical texts to convey information clearly and systematically.

2. Official-business Style: Marked by standardized expressions, formal tone, and impersonal language. Commonly found in legal documents, official correspondence, and administrative texts.

3. Journalistic Style: Combines informative content with expressive language to engage readers. It often includes rhetorical devices and is prevalent in newspapers, magazines, and broadcast media.

4. Colloquial Style: Reflects everyday spoken language, featuring informal vocabulary, idiomatic expressions, and a conversational tone. It is typical in personal communication and dialogues.

5. Artistic (Literary) Style: Utilizes figurative language, stylistic devices, and creative expression. Found in literature and creative writing, it aims to evoke emotions and aesthetic appreciation.

Each style serves a distinct function in society, facilitating effective communication within its respective domain.

Table 1. Comparison of methods

Method	Context-Aware	Requires Preprocessing	Accuracy
N-gram	No	Yes	Moderate
TF-IDF	No	Yes	Moderate
SVM	Partial	Yes	High
Random Forest	No	Yes	High
LSTM	Yes	No	High
BERT	Yes	No	Very High

2.2 Linguistic and Statistical Methods of Text Classification

Automatic assigning a text to a style/type or genre is a complex task, and over the years, various approaches have been used to address it. Although there are numerous approaches employed in this field, the main and high-performance are few and include the following: traditional machine learning methods, statistical models, and modern neural network solutions.

Early methods relied on manual annotation and frequency analysis. Common techniques include the following: (1) N-gram analysis: Measures the frequency of word/character sequences, useful for identifying style-specific patterns/ (2) TF-IDF: Assesses word importance within a document collection, helping to highlight stylistically significant features, (3) Part-of-speech analysis: Tracks how often different parts of speech appear, revealing stylistic trends; Syntactic analysis: Examines sentence structure and syntactic complexity.

These methods are still in use but their common limitations include disability to capture word context and semantic connections (Isaeva et al 2023).

2.3 Machine Learning Methods

With the rise of machine learning, new text classification algorithms became more accurate. Among the most are ubiquitously used are the following: (1) Logistic regression which uses probabilistic relationships between features; (2) Support Vector Machines (SVM) viewed as effective for sparse text data; (3) Random Forest / Gradient Boosting, the so-called ensemble tree-based methods enhancing accuracy; (4) Latent Dirichlet Allocation (LDA), i.e. topic modeling approach uncovering latent stylistic features.

However, these methods depend on pre-defined feature spaces, limiting their flexibility (Isaeva et al 2023).

2.4 Deep Learning and Neural Networks

Neural models developed lately significantly improved text analysis. The key approaches, i.e. LSTM, CNN and Transformers (BERT, GPT, RoBERTa), significantly changed the modern paradigm of computer linguistics. LSTM (Long Short-Term Memory) are a type of recurrent neural network proved to be able to handle word sequences quite successfully but being computationally intensive. CNN, i.e. Convolutional Neural Networks are employed for text analysis predominantly in combination with pre-trained embeddings (Word2Vec, GloVe). Transformers, including BERT, GPT, RoBERTa, are modern architectures that to take into account the full context of the word in the sentence. Of all the above, BERT stands out as the one using bidirectional attention to analyze words in context, and as such being highly effective for text profiling in general and functional styles in particular.

BERT comprises numerous advantages. Firstly, it provides deep contextual understanding: BERT analyzes words in their context thus enabling identification of the stylistic devices in the text. Secondly, it is flexible as the model can be fine-tuned on various discourses and specific text corpora. Thirdly, BERT exhibits high accuracy: it outperforms traditional methods in text classification tasks. Thus, BERT shows great

Table 2. Corpus statistics

Style	Text Count	Sentence Count	Word Forms
Journalistic	46,011	5,036,978	68,320,489
Literary	3,974	5,319,869	60,915,483
Scientific	6,592	1,403,302	19,232,061
Official business	1,122	158,159	2,023,362
Colloquial	2,218	1,365,924	12,930,388
Total	59,917	–	163,421,783

potential for determining text style (Isaeva et al 2023).

3 Experiment

3.1 Software Used

The program was developed in Python, compatible across platforms. It requires Python 3.8 and libraries such as NumPy, TensorFlow, and Pandas.

3.2 Dataset Preparation

The Data source used for the current study was Russian National Corpus (RNC, https://ruscorpora.ru/) which comprises texts, annotated for functional styles. The five primary types of functional speech styles vary depending on the conditions and goals of the communication held in a specific area of social activity. The five functional styles traditionally classified in Russian linguistic tradition include texts of different genres: Journalistic style comprises news, articles, blogs; Literary – stories, novels, poetry; research articles and academic texts are written in Scientific style; collection of Official-business style contains documents, contracts, memos and, finally, everyday spoken language is viewed as Colloquial style (Sirotinina 1993).

3.3 Data Preprocessing

3.3.1 Data Preprocessing and Preparation for Model Training

Prior to training, the dataset underwent preprocessing to eliminate duplicate entries and empty strings, thereby enhancing the overall quality and reliability of the model:

1. Data Cleaning: All duplicate records and empty rows were systematically removed.
2. Data Splitting: The dataset was partitioned into two subsets—80% of the texts were allocated for training, while the remaining 20% were reserved for validation purposes.
3. Tokenization: The bert-base-multilingual-cased pre-trained tokenizer was employed to convert raw text into a numerical format compatible with the BERT architecture.

The tokenization phase comprised the following key steps:

- Texts were segmented into tokens (words or subwords) using the WordPiece algorithm.
- Each token was assigned a unique numerical identifier (input_ids) derived from the BERT vocabulary.
- An attention mask (attention_mask) was generated to enable the model to differentiate between meaningful input and padding tokens.
- If a text sequence was shorter than the maximum allowable length (256 tokens), it was padded with special [PAD] tokens to ensure uniform input dimensions.

This preprocessing pipeline was executed prior to the commencement of model training to ensure that the textual data was appropriately formatted and standardized for subsequent computational processing.

ISSN 2007-9737

3.4 Model Architecture

The model employed in this study is based on the pre-trained BERT architecture (bert-base-multilingual-cased), enhanced with an additional fully connected classification layer tailored for multi-class text classification tasks.

We utilized the standard implementation of Bert for sequence classification, configured to accommodate the number of output classes corresponding to the unique functional styles identified within the dataset. To optimize computational efficiency, we trained the model training using GPU.

The model architecture comprises the following core components:

a) Tokenization and Input Representation

Following data preprocessing, the tokenized text inputs were converted into tensor format and supplied to the model through three essential input components:

- input_ids: sequences of token identifiers representing the textual input.
- attention_mask: binary masks indicating which tokens should be attended to (real tokens) and which are merely padding.
- labels: ground-truth class labels (representing text styles), used as target values during supervised learning.

These components enable BERT to generate context-sensitive embeddings via its attention mechanism, effectively ignoring padded elements while learning from the labeled data.

b) Embedding Layer

Each input token is transformed into a high-dimensional vector embedding that reflects its semantic and syntactic context, as derived from the surrounding textual environment.

c) Classification Layer

The contextual embeddings are subsequently passed through a fully connected (dense) layer, which outputs a probability distribution over the four target classes, corresponding to the predefined functional styles of the texts.

3.4.1 Training Parameters

The model was fine-tuned on the functional style classification task using the following hyperparameters:

a) Maximum Sequence Length: 256 Tokens

- Input sequences were truncated or padded to a maximum length of 256 tokens to balance memory efficiency and contextual relevance.
- Empirically, this length was found sufficient to capture the key stylistic and functional features typically present within sentence- or paragraph-level segments.
- Although longer sequences (e.g., 512 tokens) could potentially capture more information, they would also result in significantly higher memory consumption and training time, without a commensurate improvement in classification accuracy (Devlin et al. 2019).

b) Batch Size: 8

- The batch size determines the number of training samples processed simultaneously during each forward and backward pass.
- A batch size of 8 was selected as a compromise between stable gradient updates and manageable GPU memory consumption.
- Smaller batch sizes reduce memory overhead but may introduce greater variance in gradient estimates, while larger sizes demand more computational resources.

c) Number of Training Epochs: 4

- An epoch is defined as one complete pass through the entire training dataset (Devlin et al. 2019).
- The model was trained for four epochs, which was deemed sufficient to achieve convergence based on validation loss and performance metrics.

The choice of four training epochs was empirically validated: a lower number of epochs did not allow the model to reach optimal convergence, while an excessive number increased the risk of overfitting [5]. In this context, convergence refers to the stage in model training at which further changes in the learning rate become negligible, and the prediction error—i.e., the discrepancy between the predicted and actual values—is minimized. Thus, convergence indicates the model's proximity to an optimal solution, beyond which the probability of error decreases, and predictive accuracy improves (Ott et al. 2018):

- After four epochs, the model exhibited strong generalization capabilities without any significant degradation in performance on the validation set [Devlinet al 2019).

5. Optimizer: AdamW

Training was conducted using the AdamW optimizer (Adaptive Moment Estimation with Weight Decay), an enhanced variant of the standard Adam algorithm incorporating L2 regularization through weight decay:

- AdamW adaptively adjusts learning rates for individual model parameters, thereby accelerating convergence and improving generalization performance (Devlin et al. 2019).
- Unlike the conventional Adam optimizer, AdamW properly decouples weight decay from gradient updates, thereby yielding more stable training dynamics and reducing the risk of overfitting (Devlit et al. 2019).

3.4.2 Initial Learning Rate: 5e-5

- The learning rate controls the magnitude of updates during gradient descent. An initial learning rate of 5e-5 (0.00005) was chosen in accordance with established best practices for fine-tuning BERT models (Devlin et al 2019).
- Learning rates exceeding 1e-4 tend to destabilize training and hinder convergence, whereas excessively low rates (e.g., below 1e-6) result in prohibitively slow optimization.
- In conjunction with the AdamW optimizer, the selected learning rate facilitated effective weight updates while avoiding erratic fluctuations in the loss function (Devlin et al 2019).

3.5 Training Process

The model was trained over the course of four epochs. Parameter optimization was performed using the AdamW algorithm, a modified version of the standard Adam optimizer, which incorporates an L2 regularization term (weight decay). This regularization mechanism significantly mitigates the risk of model overfitting.

During the training process, a gradual decrease in the loss function was observed. However, certain epochs exhibited fluctuations, attributable to the optimizer's operational dynamics and the inherent characteristics of the dataset. The graph below illustrates the progression of the loss function along with the corresponding training time.

The following observations can be made:

- During the second epoch, the loss function temporarily increased, which may be attributed to internal fluctuations of the AdamW optimizer.
- The third epoch yielded the best results, with the lowest observed loss value.
- The average batch processing time ranged from 8.74 to 21.48 seconds per iteration.

3.6 Model Evaluation

To assess the performance of the proposed model, the following metrics were calculated: Accuracy = 94.7%; F1-score (macro) = 92.3%; Precision (macro) = 93.1%; Recall (macro) = 91.5%. As observed, these metrics indicate the high quality of the classification model.

Thus, it can be concluded that training the model on 59,917 texts resulted in high accuracy (94.7%). However, there remain unexplored options, such as (a) increasing the number of epochs and (b) adjusting the optimizer parameters. The time spent on classification tasks can be classified as significant; therefore, the prospects

ISSN 2007-9737

Table 3. Training process

Epoch	Loss	Time	Speed (sec/iter)
0	0.098	6h 43m	10.79
1	0.0234	13h 22m	21.48
2	0.106	5h 26m	8.74
3	0.00738	6h 21m	10.22

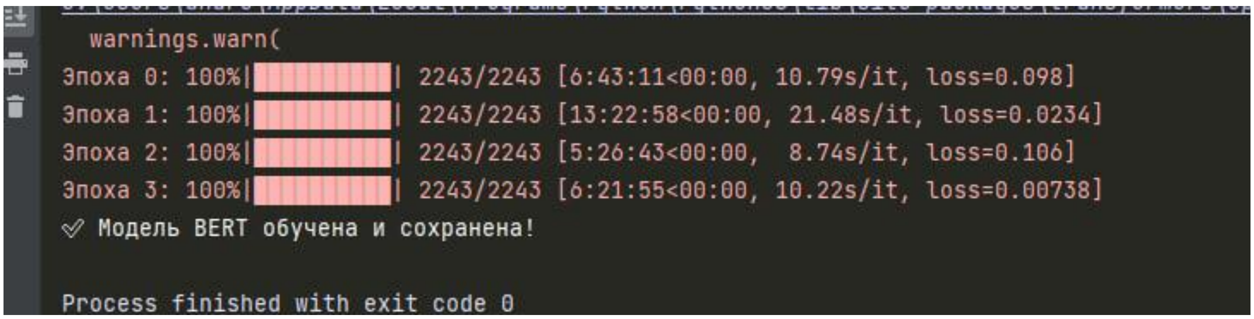

Fig. 1. Visualization of the training process

for further research involve using distilled versions of BERT (e.g., DistilBERT). Several spikes in the loss function during the 2nd epoch are likely linked to abrupt changes in weight parameters, and it seems that their elimination could be achieved by implementing a learning rate scheduling plan. An additional optimization measure to balance learning speed and accuracy may involve adjusting the batch size.

4 Discussion

Numerous attempts have been made to identify the functional character of the relationship between different registers/discourses and confirm the fact that “the linguistic features that make up a register are motivated by the needs and constraints of the communicative situation” (Li, Dunn, & Nini 2023, p.789). All the developed classification models are built on the notion of some kind of linguistic co-occurrence of linguistic features in the texts of specific registers/discourses or functional styles, i.e. that a register/functional style or a discourse could be determined by sets of linguistic features.

The classification results we achieved in this study are consistent with (Adhikari et al 2019) who utilized four datasets. i.e. Reuters-21578, arXiv Academic Paper dataset, IMDB and Yelp 2014 reviews, for document classification tasks and confirmed BERT’s effectiveness in document classification tasks. BERT achieved the state of the art across the four datasets under study. However, compared to the robust performance of BERT in capturing functional style-specific features in our study, their model’s F score is lower: 90.7 for Reuters, 75.2 for AAPD, 55.6 for IMDB and 72.1 for Yelp ’14.

This performance also exceeds the previous studies with Russian datasets. Lagutina K.V. and co-authors (2021) examine a dataset containing the classes of scientific articles, advertisements, tweets, novels, reviews, and political articles. They achieved the highest classification accuracy (F1 = 98%) for fiction only. In a subsequent study (Lagutina, 2023), they attained even higher accuracy, i.e. F1 = 99% while classifying novels, articles, reviews, social media posts, and news texts from the OpenCorpora corpus. Increasing the number of classification groups makes the task more complex.

A taxonomy of ten genres – including science fiction, fantasy, detective stories, prose, history, information technology, natural sciences, historical sciences, medicine and health, cooking, culture, and art – was analyzed in (Nikolaev, 2022). The best achieved accuracy in this case was only F1 of 71.11%, and notably, this result was obtained after just three training epochs.

5 Conclusion

This article presents a classification of functional styles using a pre-trained BERT model. The model training algorithm includes the preparation of a text corpus, data annotation, setting of training parameters, and subsequent validation of the results.

The experiments confirmed the hypothesis that texts of different functional styles possess distinctive linguistic parameters, which can be used for accurate automatic classification. The model achieved high accuracy (94.7%), demonstrating the effectiveness of using BERT for this task. The obtained scores are comparable to those reported in previous studies.

The study also identified certain limitations, including the time required for training and fluctuations in the loss function during the second epoch, which are likely related to abrupt changes in weight parameters.

Promising directions for further research include:

- The use of distilled versions of BERT (such as DistilBERT) to reduce training time,
- Implementation of learning rate scheduling to stabilize training, and
- Adjustments to batch size to optimize the balance between training speed and classification accuracy.

Overall, the proposed approach shows high potential for application in automatic text processing tasks, authorship attribution, and stylistic analysis of Russian-language texts.

Acknowledgments

The research was supported by the RSF grant № 24-78-10129[1].

[1] http://rscf.ru/project/24-78-10129/

References

1. **Adhikari, A., Ram, A., Tang, R., Lin, J. (2019).** DocBERT: BERT for Document Classification. arXiv: 1904.08398. DOI: 10.48550/arXiv.1904.08398.
2. **Bally, C. (1909).** Traité de stylistique française [Treatise on French stylistics]. Heidelberg: Carl Winter.
3. **Chowdhary, K.R. (2020).** Natural language processing. Fundamentals of artificial intelligence. pp. 603–649.
4. **Dejica, D. (2020).** Understanding Technical and Scientific Translation: A Genre-based Approach. Scientific Bulletin of the Politehnica University of Timisoara. Transactions on Modern Languages/ Buletinul Stiintific al Universitatii Politehnica din Timisoara. Seria Limbi Moderne, Vol. 19, No. 1. pp. 56–66.
5. **Devlin, J.**, **Chang, M.W.**, **Lee, K.**, **Toutanova, K. (2019).** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint 1810.04805. DOI: 10.48550/arXiv. 1810.04805.
6. **Gumperz, J.J. (1982).** Discourse Strategies. Cambridge: Cambridge University Press.
7. **Isaeva, M.Z.**, **Alisultanova, A.V.**, **Dasaev, M.-A.**, **Tsuev, M.M. (2023).** Methods and technologies of natural language processing. Journal Name. Vol. 49, pp. 99–014. DOI: 10.26200/GSTOU.2023.49. 99.014
8. **Lagutina, K.V.**, **Lagutina, N.S.**, **Boychuk, E.I. (2021).** Text classification by genre based on rhythm features. Model. Anal. Inform. Sist., Vol. 28, No. 3. pp. 280–291. DOI:10.18255/1818-1015-2021-3-280-291.
9. **Li, H., Dunn, J., Nini, A. (2023).** Register variation remains stable across 60 languages. Journal of Computational Linguistics, Vol. 49, No. 4. pp. 789–822. DOI: 10.1162/coli_ a_00489.
10. **Kenton, J.D.M.W.C.**, **Toutanova, L. K. (2019).** Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, Vol. 1, No. 2.

11. **Kochetova, L.A., Popov, V.V. (2019).** Research of axiological dominants in press release genre based on automatic extraction of key words from corpus. Nauchnyi Dialog, Vol. 6.
12. **Labov, W. (2001).** The anatomy of style-shifting. In **Eckert, P., Rickford, J.R. (Eds.),** Style and Sociolinguistic Variation pp. 85–108. Cambridge: Cambridge University Press.
13. **Lagutina, K.V., Lagutina, N.S., Boychuk, E.I. (2021).** Text classification by genre based on rhythm features. Modeling and Analysis of Information Systems, Vol. 28, No. 3. pp. 280–291.
14. **Lagutina, K.V. (2023).** Genre classification of Russian texts based on modern embeddings and rhythm. Automatic Control and Computer Sciences, Vol. 57, No. 7, pp. 817–827.
15. **Malhotra, R., Sharma, A. (2017).** Quantitative evaluation of web metrics for automatic genre classification of web pages. International Journal of System Assurance Engineering and Management, Vol. 8, No. 2, pp. 1567–1579.
16. **Melissourgou, M.N.**, **Frantzi, K.T. (2017).** Genre identification based on SFL principles: The representation of text types and genres in English language teaching material. Corpus Pragmatics, Vol. 1, No. 4. pp. 373–392.
17. **Moiseeva, I., Remizova, V.F. (2015).** Intensifikatory v angliyskom yazyke: Funktsional’no-stilisticheskiy i grammaticheskiy aspekty [Intensifiers in English: Functional-stylistic and grammatical aspects [in Russian]]. Sovremennye Problemy Nauki i Obrazovaniya, Vol. 1.
18. **Murphy, S.E. (2019).** Shakespeare and his contemporaries: Designing a genre classification scheme for Early English Books Online 1560-1640. ICAME Journal, pp. 59–82.
19. **Nikolaev, P.L. (2022)**. Classification of books into genres based on text descriptions via deep learning. International Journal of Open Information Technologies, Vol. 10, No. 1. pp. 36–40.
20. **Ott, M.**, **Edunov, S.**, **Grangier, D.**, **Auli, M. (2018).** Scaling Neural Machine Translation. Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 1–9. Brussels, Belgium: Association for Computational Linguistics.
21. **Rothman, D. (2021).** Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python. PyTorch, TensorFlow, BERT, RoBERTa, and more. Packt Publishing Ltd.
22. **Sirotinina, O.B. (1993).** Functional Styles of the Russian Language. Moscow: Nauka.
23. **Solnyshkina, M.**, **Kupriyanov, R.**, **Shoeva, G. (2024).** Linguistic Profiling of Text Genres: Adventure stories vs. textbooks. Research Result. Theoretical and Applied Linguistics, Vol. 10. pp. 115–132. DOI: 10.18413/2313-8912-2024-10-1-0-7.
24. **Solnyshkina, M.I., Solovyev, V.D., Ebzeeva, Y.N. (2024).** Approaches and tools for Russian text linguistic profiling. Russian Language Studies, Vol. 22, No. 4, pp. 501–517. DOI: 10.22363/2618-8163-2024-22-4-501-517
25. **Solovyev, V., Solnyshkina, M., & Ten, A. (2024).** A BERT-based classification model: The case of Russian fairy tales. Journal of Language and Education, Vol. 10, No. 4. pp. 98–111.
26. **Turchin, A.**, **Masharsky, S., Zitnik, M. (2023).** Comparison of BERT implementations for natural language processing of narrative medical documents. Informatics in Medicine Unlocked, Vol. 36, No. 101139.

Article received on 25/08/2024; accepted on 10/01/2025.
**Corresponding author is Valery D. Solovyev.*