https://doi.org/10.15407/intechsys.2025.02.090
UDC 004.91

M.A. MUZYCHUK, Master's Student,
Computer Systems Software Department of the Applied Mathematics Faculty,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
37, Beresteyskyi ave., Kyiv, 03056, Ukraine
maryna.muzychuk@gmail.com
T.M. ZABOLOTNIA, PhD (Engineering), Asociate Professor,
Computer Systems Software Department of the Applied Mathematics Faculty,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute",
37, Beresteyskyi ave., Kyiv, 03056, Ukraine
https://orcid.org/0000-0001-8570-7571
tetiana.zabolotnia@gmail.com

AUTOMATIC CLASSIFICATION OF UKRAINIAN TEXTS BY FUNCTIONAL STYLES

The proposed multilevel method for classifying Ukrainian texts by functional style combines statistical analysis, keyword analysis, and contextual analysis based on the BERT model, which accounts for semantic and contextual dependencies in the text.

The results support the hypothesis that combining contextual features (generated by BERT) with statistical style parameters yields the highest classification accuracy. This highlights the advantage of the proposed model for tasks requiring high precision and stability in identifying functional text styles.

Keywords: *classification, functional style, stylometry, vectorization, machine learning*

Introduction and Problem Statement

Automatic classification of texts by functional styles is an important task in software engineering, as it enables the automation of text data processing for solving common problems such as information retrieval, document analysis etc. Determining a text's functional style requires analyzing its lexical, grammatical, and stylistic features while considering its context. The main challenge here lies in the significant stylistic diversity of the Ukrainian language. Moreover, within a single text, the boundaries between multiple styles may be blurred. Existing solutions

Cite: Muzychuk M.A., Zabolotnia T.M. Automatic Classification of Ukrainian Texts by Functional Styles. *Information Technologies and Systems*, Київ, 2025, Том 2 (2), 90–97. https://doi.org/10.15407/intechsys.2025.02.090

developed for other languages require substantial adaptation for Ukrainian, posing new challenges for engineers. Therefore, there is a need to develop specialized methods capable of identifying functional styles specifically in Ukrainian texts.

The goal of this study is to improve the accuracy of Ukrainian text classification by functional styles through the development of a method and software for automatic text style identification.

To achieve this goal, the study sets and addresses the following tasks: analyzing existing approaches to automatic text classification; identifying the advantages and disadvantages of current methods in the context of Ukrainian text classification by functional styles; describing the proposed method; outlining the architecture, implementation tools, and key modules of the software developed for the proposed method; evaluating the effectiveness of the developed model and comparing its results with alternative approaches.

Literature Review

Existing approaches to solving the problem of automatic text classification by functional styles can be broadly divided into two groups:

1. Vectorization methods — the purpose of these methods is to represent textual data in a vector format. They include *the keyword method*, *statistical methods, the n-gram method, TF-IDF,* and *BERT (as a vectorization method).* In the context of classifying Ukrainian texts by functional styles, vectorization methods offer **advantages** such as scalability and the ability to capture key textual features, enabling the analysis of stylistic differences. However, their **drawbacks** include sensitivity to the volume and quantity of processed data, as well as the need for adaptation to the Ukrainian language.

2. Machine learning methods — these methods aim to train a model that can later classify input data based on prior training. Among machine learning approaches, notable examples include *the support vector machine (SVM) method, probabilistic methods, ensemble methods and neural network-based techniques*. The **advantage** of these methods for text classification by style lies in their ability to account for complex stylistic structures and their scalability, allowing them to process large datasets without losing efficiency. However, their **disadvantages** include sensitivity to noise, which can lead to model overfitting, as well as the need for significant computational resources [1].

In this study, the authors propose using a pre-trained BERT language model, which has proven to be an effective tool for text analysis due to its ability to consider context at both the word and sentence levels [2]. However, given the specifics of the Ukrainian language and the features of functional styles, standard available BERT models have significant limitations:

• bert-base-ukr-eng-rus-uncased [3], although a specialized Ukrainian model, is trained primarily on publicly available corpora that do not cover a wide variety of functional styles (e.g., literary, academic, journalistic). This limits its ability to accurately recognize stylistic features. Additionally, the model tends to mix stylistically similar styles, such as official-business and academic, due to the lack of clear stylistic segmentation in the data;

• SlavicBERT [4], despite its name, does not support the Ukrainian language. The model is trained mainly on Russian, Bulgarian, Polish, and Czech, making its application to Ukrainian texts not only inefficient but also methodologically incorrect;

• Multilingual BERT (mBERT) [5] — a multilingual model developed by Google to support over 100 languages, including Ukrainian. However, the model is trained on general corpora without accounting for stylistic nuances. As a result, it performs poorly in classification tasks.

Thus, given the identified limitations of existing models, there is a need to adapt BERT for the automatic classification of Ukrainian texts by style. To achieve this, it is necessary to fine-tune the model on specialized corpora covering various functional styles and to incorporate key distinguishing features—such as syntactic patterns, part-of-speech frequency, and specific linguistic expressions. This approach will enable deeper stylistic understanding, improve classification accuracy, and account for both contextual and structural features of each style. Ultimately, it will result in a model capable of effectively processing a wide range of Ukrainian texts while reflecting real-world language practices.

Method for Automatic Classification of Ukrainian Texts by Functional Styles

In the Ukrainian language, four main literary functional styles are distinguished: scientific, artistic, official-business, and journalistic. Currently, there is no universally defined set of features that would allow for an unambiguous determination of a text's style [3]. However, each of these styles has specific linguistic characteristics that can be generalized into three categories: lexicon, syntax, and morphology.

To achieve automatic classification of texts by functional styles, the authors propose using the following key parameters: contextual dependencies within the text, identification of words characteristic of a particular style, and stylometric aspects, i.e., statistical style parameters.

Based on the key aspects described earlier, this study proposes a multi-level method for automatic classification of texts by functional styles, which integrates statistical analysis, keyword analysis, and contextual analysis using neural networks.

Stage 1: Statistical analysis. The first step involves collecting statistical parameters of the text, including: frequency of different parts of speech (nouns, adjectives, verbs, etc.); usage of various verb forms (infinitives, imperative, impersonal forms); sentence structure (simple, complex, asyndetic).

The extracted statistical data is normalized: some features are analyzed relative to the text itself (e.g., the percentage of noun usage compared to the total word count). Others are analyzed in relation to the entire text corpus (e.g., the average sentence length or the number of paragraphs in a text).

Stage 2: Keyword analysis. At this stage, the text undergoes: tokenization, stop-word removal, lemmatization. After preprocessing, the frequency of lemmas characteristic of certain styles is calculated. For journalistic, scientific, and official-business styles, dictionaries of high-frequency words were compiled. For ar-

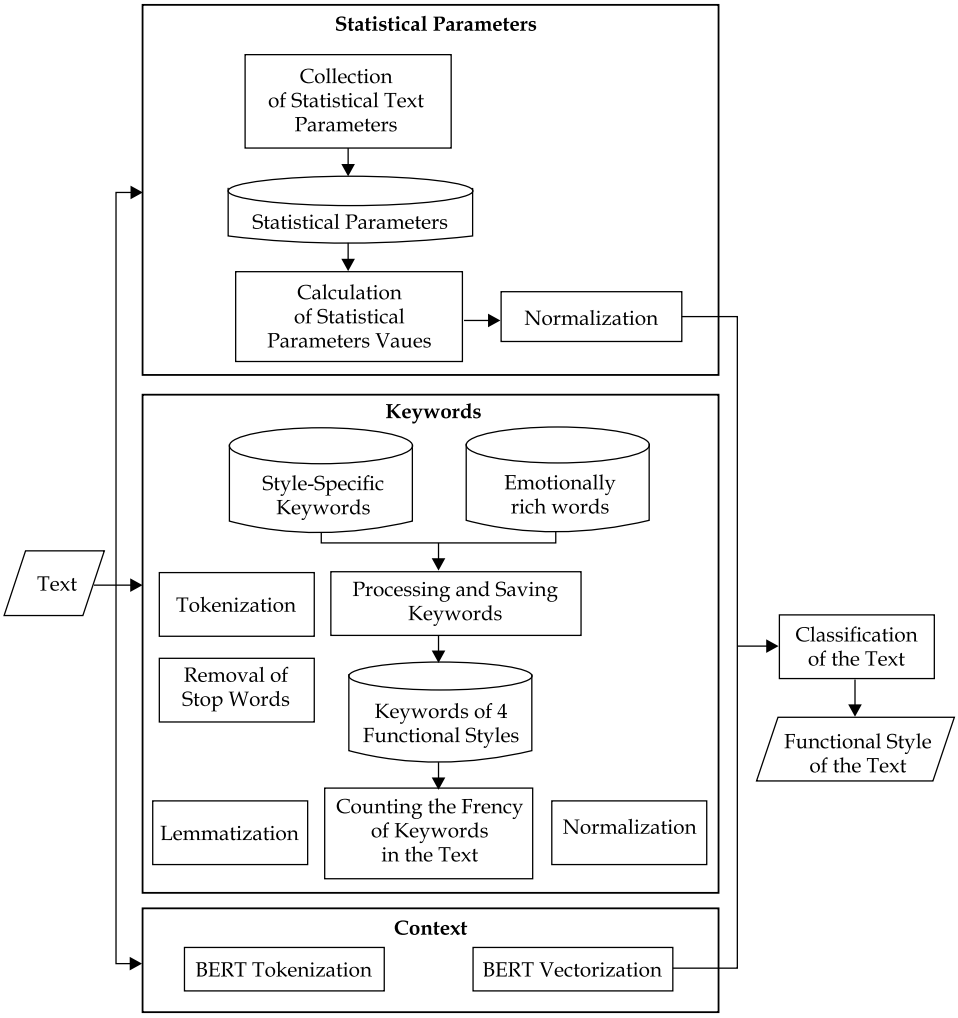

Fig. 1. Schematic representation of the proposed method for automatic text classification by functional styles

tistic style, since it primarily uses general vocabulary, analysis is performed using a set of emotionally charged words, which are distinctive for this style.

Stage 3: Contextual analysis with BERT. The third stage employs a BERT-based model to analyze not only individual words in a text but also relationships between sentences, enabling a deeper understanding of the text's semantics [4]. Through tokenization and vectorization using BERT, a high-dimensional vector representation of the text is obtained, incorporating both semantic and contextual dependencies.

The three main stages of the method can be executed in parallel since the output of each stage does not affect the operation of the other two.

A schematic representation of the proposed method is presented on Figure 1.

As a result of the proposed method, a model is formed that enables the classification of texts by functional style based on the analysis of statistical parameters, lexical features, and context. By combining three stages of analysis, which cover

different aspects of the text, the method ensures adaptability to the specifics of the Ukrainian language.

Implementation of Software and Results

Implementation. The software for automatic classification of Ukrainian texts by functional styles is implemented as a console application. The advantages of a console-based approach include: cross-platform compatibility, allowing it to run on different operating systems without additional configuration; integration capabilities, enabling incorporation into scripts or other systems for automated processing of large text volumes; flexibility in handling textual data.

The following tools were selected for developing the console application:

1. Python as the programming language.
2. Pandas for dataset handling.
3. TensorFlow as the machine learning framework.
4. Pymorphy3 for morphological analysis.
5. NLTK for natural language processing.
6. Redis as an in-memory database management system.
7. SpaCy for text processing and natural language understanding.

The software is designed using a modular architecture, which divides the system into independent modules, each responsible for a specific function. This approach enhances code maintainability and simplifies testing by allowing independent module validation.

The application consists of the following key modules:

1. Root module — the root module, which serves as the entry point for starting the program. This is the primary control module that ensures the coordinated operation of the entire system.

2. Command line processing — a module for processing command-line input, which handles the initial stage of interaction with the user.

3. Statistic parameters analysis — computes statistical characteristics of the text.

4. Key words usage analysis — analyzes the frequency of key terms in the text.

5. Context analysis — generates a vector representation of the text while considering contextual dependencies.

6. Values normalization — normalizes values obtained from statistical and keyword frequency analysis modules.

7. Cache interaction — a module for interacting with the Redis DBMS, which is used for data caching.

8. Filesystem interaction — a module that handles interactions with the file system.

Results. To evaluate the effectiveness of the proposed method for automatic classification of Ukrainian texts by functional styles, the following metrics were used: Accuracy, Precision, Recall, F1-score, and AUC. The choice of these metrics ensures a comprehensive evaluation of the model, as they allow for analysis of classification accuracy, the balance between accuracy and completeness of predictions, as well as the model's ability to reliably rank text styles, contributing to a thorough analysis of the results.

For evaluating the effectiveness of the proposed method, its results were compared with two alternative approaches: the first based on the classical BERT model, and the second using SVM and statistical characteristics of the styles. This comparison allows for testing the hypothesis that integrating contextual features obtained with BERT with statistical style parameters improves the accuracy of text classification by functional styles.

To prepare the textual data required for training the BERT model, this study compiled a large corpus of Ukrainian texts with an even distribution across different functional styles. The source of the corpus was the open-access resource UberText 2.0 [9], which provided over 40,000 text samples — 10,000 per each of the four styles.

For a comprehensive evaluation of the model, an additional dataset with structurally and thematically diverse texts was used, downloaded from an open GitHub repository [10]. Stored in separate .txt files, this corpus was processed to create a validation dataset of 4,000 texts by iterating through the files and sequentially adding their content.

The analysis results indicate that the proposed model significantly outperforms both classical BERT and SVM, which is based on statistical features. The Accuracy metric reached **0.829**, while BERT and SVM showed 0.646 and 0.612, respectively, confirming the higher accuracy in text classification. Similarly, in terms of positive prediction precision (Precision), the proposed model demonstrated **0.780**, surpassing BERT (0.626) and SVM (0.541), indicating a reduction in false positive results. The Recall (0.709) and F1-score (0.729) metrics also exceeded the results of the alternatives, providing a balance between precision and recall.

The AUC score (**0.952**) confirms the model's ability to more accurately recognize text styles, outperforming BERT (0.908) and SVM (0.834). These results prove that the proposed model demonstrates not only high accuracy but also stability in working with different functional styles.

The obtained results confirm the hypothesis that combining contextual features formed with BERT with statistical style parameters provides the highest classification accuracy. The relative improvement in accuracy is **28.39%** compared to BERT and **35.49%** compared to SVM, emphasizing the advantage of the proposed model for tasks that require high accuracy and stability in identifying functional text styles.

Conclusions

The work presents an analysis of existing methods for automatic text classification by functional styles. An overview of the main vectorization methods and machine learning techniques is provided, along with an analysis of their advantages and disadvantages in the context of classifying Ukrainian texts by functional style. A multi-level method is proposed, which combines statistical analysis, keyword analysis, and contextual analysis based on the BERT model, allowing for the consideration of semantic and contextual dependencies within the text. The stages of the proposed method are presented: collecting statistical parameters of the text, identifying characteristic lemmas for styles, and applying contextual analysis to

improve text classification. A console application based on a modular architecture is implemented. The effectiveness of the method is evaluated using Accuracy, Precision, Recall, F1-score, and AUC metrics, demonstrating the superiority of the proposed model over alternative approaches.

REFERENCES

1. What are the advantages and disadvantages of Random Forest? URL: https://aiml.com/what-are-the-advantages-and-disadvantages-of-random-forest/ [Accessed 15 Nov. 2024]
2. Understanding searches better than ever before. URL: https://web.archive.org/web/20210127042834/https://www.blog.google/products/search/search-language-understanding-bert/ [Accessed 15 Nov. 2024]
3. mshamrai/bert-base-ukr-eng-rus-uncased. URL: https://huggingface.co/mshamrai/bert-base-ukr-eng-rus-uncased [Accessed 15 Nov. 2024]
4. Slavic BERT NER. URL: https://github.com/deeppavlov/Slavic-BERT-NER/blob/master/README.md [Accessed 15 Nov. 2024]
5. multilingual.md. URL: https://github.com/google-research/bert/blob/master/multilingual.md [Accessed 15 Nov. 2024]
6. Areshenkov Yu. O. *Stylistics of the Ukrainian language: lecture notes and lesson plans: teaching and methodological manual*. KrDPU, Kryvyy Rih, 2007, 3-th ed., 18p. [In Ukrainian: Арешенков Ю. О. Стилістика української мови: конспект лекцій та плани занять : навч.-метод. посіб.] https://doi.org/10.31812/0564/2140
7. Artistic style as a type of language. Substyles of artistic style. Genres of artistic style. Colors of artistic style. URL: https://studfile.net/preview/5721078/page:36 [Accessed 15 Nov. 2024] [In Ukrainian: Художній стиль як різновид мови. Підстилі художнього стилю. Жанри художнього стилю. Колорити художнього стилю]
8. BERT 101. State Of The Art NLP Model Explained. URL: https://huggingface.co/blog/bert-101 [Accessed 15 Nov. 2024]
9. UberText 2.0. URL: https://lang.org.ua/en/ubertext [Accessed 15 Nov. 2024]
10. Brown corpus of the Ukrainian language. [In Ukrainian: Браунський корпус української мови]

Received 06.03.2025

М.А. Музичук, студентка,
Національний технічний університет України
«Київський політехнічний інститут ім. Ігоря Сікорського»,
Берестейський просп. 37, м. Київ, 03056, Україна
maryna.muzychuk@gmail.com

Т.М. Заболотна, канд. техн. наук, доцент,
Національний технічний університет України
«Київський політехнічний інститут ім. Ігоря Сікорського»,
Берестейський просп. 37, м. Київ, 03056, Україна
https://orcid.org/0000-0001-8570-7571
tetiana.zabolotnia@gmail.com

АВТОМАТИЧНА КЛАСИФІКАЦІЯ ТЕКСТІВ УКРАЇНСЬКОЮ МОВОЮ ЗА ФУНКЦІЙНИМИ СТИЛЯМИ

Вступ. Автоматична класифікація текстів за функційними стилями є важливим завданням в інженерії програмного забезпечення, оскільки вона дозволяє автоматизувати оброблення текстових даних для ефективного вирішення таких розповсюджених задач, як пошук інформації, аналіз документів тощо. Процес визначення функційного стилю вимагає аналізу лексичних, граматичних та стилістичних особливостей тексту з урахуванням його контексту. Основною складністю тут є те, що українська мова характеризується значною різноманітністю стилістичних варіацій. Наявні рішення,

розроблені для інших мов, потребують суттєвої адаптації для української. У зв'язку з цим постає необхідність розроблення спеціалізованих методів, здатних ідентифікувати функційні стилі саме в українськомовних текстах.

Мета статті. Метою даної роботи є підвищення точності класифікації текстів українською мовою за функційними стилями шляхом розроблення методу та програмного забезпечення для автоматичного визначення стилю тексту.

Методи. Запропонований в статті багаторівневий метод класифікації текстів українською мовою за функційними стилями поєднує статистичний аналіз, аналіз ключових слів та контекстний аналіз на основі моделі *BERT*, що дозволяє враховувати семантичні та контекстуальні залежності в тексті. Етапами запропонованого методу є: збір статистичних параметрів тексту, визначення характерних лем для стилів, а також застосування контекстного аналізу для покращення класифікації текстів. Метод реалізовано в межах консольного застосунку, що базується на модульній архітектурі.

Результат. Запропонована модель значно перевершує як класичну *BERT*, так і *SVM* за метрикою Accuracy: вона досягла 0,829, тоді як *BERT* і *SVM* показали 0,646 і 0,612 відповідно. За точністю позитивних передбачень (Precision) запропонована модель продемонструвала 0,780, випередивши *BERT* (0,626) і *SVM* (0,541), що свідчить про зменшення кількості хибно-позитивних результатів. Показники *Recall* (0,709) і *F*1-score (0,729) також перевищують результати альтернатив, забезпечуючи збалансованість між точністю і повнотою. Показник *AUC* (0,952) підтверджує здатність моделі точніше розпізнавати стилі текстів, перевершуючи значення *BERT* (0,908) і *SVM* (0,834).

Висновки. Отримані результати підтверджують гіпотезу, що поєднання контекстних ознак, сформованих з допомогою *BERT*, із статистичними параметрами стилю забезпечує найвищу точність класифікації. Це підкреслює перевагу запропонованої моделі для задач, які вимагають високої точності та стабільності у визначенні функційних стилів тексту.

Ключові слова: *класифікація, функційний стиль, стилеметрія, векторизація, машинне навчання.*