NATALIIA DARCHUK
Taras Shevchenko National University of Kyiv
ORCID id: orcid.org/0000-0001-8932-9301

Fields of research: computer linguistics, corpus linguistics, quantitative linguistics, grammar and semantics of the Ukrainian language.

OKSANA ZUBAN
Taras Shevchenko National University of Kyiv
ORCID id: orcid.org/0000-0002-2644-3892

Fields of research: computer linguistics, linguistic expertise, quantitative linguistics, grammar and semantics of the Ukrainian language.

VALENTYNA ROBEIKO
Taras Shevchenko National University of Kyiv
ORCID id: orcid.org/0000-0003-2266-7650

Fields of research: speech analysis, recognition and synthesis, phonetics, natural language processing.

YULIIA TSYHVINTSEVA
Taras Shevchenko National University of Kyiv,
Institute of the Ukrainian Language of the National Academy of Sciences of Ukraine
ORCID id: orcid.org/0000-0002-9684-3840

Fields of research: Ukrainian neology, lexicology and lexicography.

VICTOR SOROKIN
Taras Shevchenko National University of Kyiv
ORCID id: orcid.org/0000-0002-3637-0535

Fields of research: automatic syntax analysis, automatic semantic analysis, natural language processing.

NATALIIA DARCHUK, OKSANA ZUBAN, VALENTYNA ROBEIKO,
YULIIA TSYHVINTSEVA, VICTOR SOROKIN, MYKOLA SAZHOK

MYKOLA SAZHOK
Institute for information technologies and systems
of the National Academy of Sciences of Ukraine
ORCID id: orcid.org/0000-0003-1169-6851

Fields of research: speech analysis, recognition and synthesis, natural language processing.

DOI: doi.org/10.35321/all91-09

THE SYSTEM FOR AUTOMATIC STYLOMETRIC ANALYSIS OF UKRAINIAN MEDIA TEXTS TEXTATTRIBUTOR 1.0 (TECHNIQUES, MEANS, FUNCTIONALITY)

Ukrainos žiniasklaidos tekstų automatinės stilometrinės analizės sistema „TextAttributor 1.0“ (metodai, priemonės, funkcionalumas)

ANNOTATION

This paper presents the structure, algorithms, implementation, and experimental results of the automatic TextAttributor system developed by the authors of the paper for statistical Ukrainian-language text parameterisation using a multiparametric set of statistical indices, characterising the author’s text style and applicable to authorship attribution tasks. Based on the created linguistic resources and software, the system generates a linguistic analysis based on the calculated statistical indices and performs a comparative study of two texts. An additional criterion for statistical indexing is the text toxicity index, calculated through the method of verbal identification of toxic sentiment. Authorship and toxicity detection tasks are addressed using two methods: dictionary- and rule-based statistical calculations and machine learning. The current findings implemented in the beta version of TextAttributor are thoroughly examined.

KEYWORDS: Computational linguistics, Ukrainian language, sentiment analysis, authorship attribution, stylometry, text classification.

The System for Automatic Stylometric Analysis of Ukrainian Media Texts TextAttributor 1.0 (Techniques, Means, Functionality)

ANOTACIJA

Šiame straipsnyje pristatoma straipsnio autorių sukurtos automatinės sistemos „TextAttributor“, skirtos statistiniam ukrainiečių kalbos tekstų parametrizavimui naudojant daugiaparametrinį statistinių rodiklių rinkinį, apibūdinantį autoriaus teksto stilių ir taikomą autorystės atribucijos uždaviniams, struktūra, algoritmai, įdiegimas ir eksperimentiniai rezultatai. Sukurtų lingvistinių išteklių ir programinės įrangos pagrindu sistema generuoja lingvistinę analizę pagal apskaičiuotus statistinius indeksus ir atlieka dviejų tekstų lyginamąją analizę. Papildomas statistinio indeksavimo kriterijus yra neigiamas nuotaikas sukeliančio teksto indeksas, apskaičiuojamas taikant žodinio neigiamų nuotaikų identifikavimo metodą. Autorystės ir pagiežos nustatymo užduotys sprendžiamos dviem metodais: žodynu ir taisyklėmis pagrįstais statistiniais skaičiavimais ir mašininiu mokymusi. Dabartiniai rezultatai, gauti naudojantis „TextAttributor“ beta versija, išsamiai išnagrinėti.

ESMINIAI ŽODŽIAI: kompiuterinė lingvistika, ukrainiečių kalba, jausmingumo analizė, autorystės priskyrimas, stilometrija, teksto klasifikavimas.

1. INTRODUCTION

Textological research gained new scientific importance with the advancement of computational linguistics methods, which contributed to the formation of a new direction that can be considered digital textology. Within this emerging discipline, we conceptualize digital textology as harnessing automated corpus linguistics techniques, coupled with mathematical models for quantitative text analysis. Guided by the current tasks of modern quantitative linguistics and natural language processing, our team has developed a system for automatic linguistic-statistical analysis of media texts, which has been implemented as a web application named TextAttributor (TextAttributor 1.0 2024). The system operates in four tasks: 1) statistical text parameterization; 2) stylometry: determining the linguistic-statistical features of idiolect; 3) attribution: determining the degree of similarity between texts; and 4) sentiment analysis: identifying negative sentiment lexicon in texts. Additionally, within the system, two linguistic expert conclusions are automatically generated based on the results of the second and fourth tasks. The multifunctionality of the TextAttributor system requires users to familiarize themselves with its operating principles. The purpose of this article is to acquaint the academic and educational philological community with the system's creation methodology, architecture, and operational results to provide a clear understanding of its functions and analytical capabilities, which are particularly relevant. This approach is especially relevant for analyzing large volumes of internet communications and enhancing information defence strategies during the ongoing Russian-Ukrainian war. The concept of developing an automatic system of Ukrainian-language text attribution arose from the analysis of research in the field of Ukrainian corpus linguistics (Darčuk 2013) and statistical studies of the author's style (Darčuk *et al.* 2021; Darchuk, Sorokin 2022; Zuban' 2019) conducted using the tools of the Ukrainian Language Corpus (KUM) by the project authors.

During the development of the TextAttributor system, the following methods were employed: component analysis, distributive analysis, and content analysis. Content analysis, a quantitative and qualitative method for extracting information from text, utilized natural language processing and statistical techniques. Quantization and sentiment analysis were also applied. Sentiment analysis automatically identifies text tonality based on both the emotional colouring and the author's assessment of events or objects, which was achieved utilizing dictionary- and rule-based statistical calculations.

Additionally, machine learning methods, including deep learning, were incorporated. The statistical structure of a text understood as its quantitative model, enables the identification of its functional style, authorship, and period of creation. In this work, the statistical structure components were extracted by analyzing lexical and grammatical statistical features using indexing techniques and Euclidean distance metrics. The Euclidean distance, which measures the distance between two points in an n-dimensional Euclidean space, is one of the most widely used metrics in linguostatistics for cluster analysis in stylometry and authorship attribution.

The novelty of the results is the first implementation in computational linguistics of an automated morpho-syntactic-semantic stylometric model for text analysis, based on 15 statistical indices that primarily parameterize the morphological, as well as syntactic and semantic structure of the text. A stylometric model is a set of statistical parameters of a text (lexical, morphological, syntactic, etc.), based on which a comparative analysis of this text is carried out with the stylometric model of the functional style. In our opinion, it is possible to obtain a strict, deterministic, scientifically grounded system of common and distinctive features in styles, genres, etc. by using static methods in the construction of a stylometric model. For the first time, the stylometric model introduces the toxicity index parameter, which characteristically defines media texts during the Russian-Ukrainian war. This model is implemented not only in the function of statistical parameterization of texts, which is typical for systems of this type, but also in the function of comparing the analyzed text with the media style of the Ukrainian language, and for stylometric and attribution tasks involving two or more texts. The theoretical significance extends to the validation of two methods for determining text toxicity and authorship in Ukrainian: 1) through dictionary and rule-based approaches, and 2) via machine learning, including deep learning.

2. RELATED WORKS

In modern Ukrainian linguistics, substantial linguistic-statistical studies have been conducted on the idiolects of Ukrainian writers and poets, including Taras Shevchenko, Lesya Ukrainka, Vasyl Stus (Zuban’ 2019), Ivan Franko (Buk 2021), Roman Ivanychuk (Lototska 2022), Valerii Shevchuk (Volos, Levchenko 2023), Mariia Matios, Yurii Andrukhovych, Oksana Zabuzhko (Karasov, Levchenko 2024), Ivan Drach, Mykola Vingranovsky (Darchuk *et al.* 2024), Lina Kostenko (Zuban’ 2019; Darchuk *et al.* 2024), among others. These studies were conducted on representative text samples (long texts) using predominantly one or a few (no more than five) statistical parameters, often without applying stylometric comparison. Furthermore, due to the laborious nature of conducting linguistic-statistical experiments, Ukrainian linguistics has largely overlooked comprehensive statistical parameterization, comparison of analyzed texts with standard statistical parameters of functional styles, determination of the degree of text similarity, as well as stylometric and attribution analysis of short texts. The use of the TextAttributor system in linguistic-statistical research, aimed at performing these tasks automatically, will not only provide frequency characteristics of linguistic phenomena but also enable efficient stylometric analysis resulting in an expert opinion.

The TextAttributor system has advantages compared to its counterparts in global science. Most existing systems and models also focus on the use of individual linguistic-statistical modules to identify one or a few, mostly formal (n-grams, most frequent words, letters), text parameters (Eder 2015; Canhasi *et al.* 2022; LIWC-22; Khomytska *et al.* 2023; ALIAS). In contrast, TextAttributor 1.0 offers a comprehensive approach, including interactive statistical analysis of the lexical, morphological, syntactic, and semantic structure of the text in real-time across 15 parameters.

Attribution of texts is actively developed in foreign textual studies, in particular, Burrows’s method (Burrows 2002; Argamon 2007; Eder, Rybicki 2013; Burrows *et al.* 2014) effectiveness is tested in many studies on large volumes of textual data of various styles like: English prose of the early 20th century (Hoover 2004); modern English poetry (Hoover 2005); poetic works in Latin (Rybicki, Eder 2011); prose works of major genres in English, French, Italian, German, Polish, Hungarian, as well as Latin and Arabic (Rybicki, Eder 2011; Evert *et al.* 2015; Jannidis *et al.* 2015); political texts in English, including the attribution of speeches of American presidents (Savoy 2015). One more automated linguistic analysis method called "Linguistic Inquiry and Word Count" (LIWC) is implemented as a commercial application and adapted to several languages (Meier *et al.* 2019; LIWC-22).

In recent years, the tasks of automatic text attribution have predominantly used either rule-based stylometric methods or deep learning methods (Eder 2015; Canhasi *et al.* 2022). Both methods were implemented in the TextAttributor system, effectively functioning but yielding different outcomes: 1) The rule-based method allows automating the generation of linguistic conclusions about idiolect and text toxicity; 2) The machine learning method determines only the degree of toxicity and the similarity of texts and can be used in classification tasks for monitoring textual content.

The recent machine learning methods for sentiment analysis and text authorship identification are based on statistical approaches (TF-IDF, Latent Dirichlet Allocation) and deep learning techniques with various architectures (CNN, LSTM, BERT) in combination with stylometric methods (Gupta *et al.* 2019; Canhasi *et al.* 2022; Bonetti *et al.* 2023). The reported results are extremely dependent on the number of classes (types of toxicity, authors), genre, length of the text and, most importantly, the volume of the properly annotated text corpus used for the training procedure. Thus, high accuracy rates in the tasks of detecting toxicity and hate speech exceed 90% using a corpus containing tens of thousands of documents (Alkomah, Ma 2022). For example, more than 90% accuracy in authorship attribution is achieved for English-language literary works across 1000 authors, using a corpus of 6000 documents for 15 authors. In turn, for corpora containing less than 1500 documents the reported accuracy is about 70% (Khan *et al.* 2023). For the Ukrainian language, text authorship identification and sentiment/toxicity analysis are understudied problems (Lupei *et al.* 2020), moreover, such systems do not produce linguistic expertise and cannot be used as tools for linguistic research.

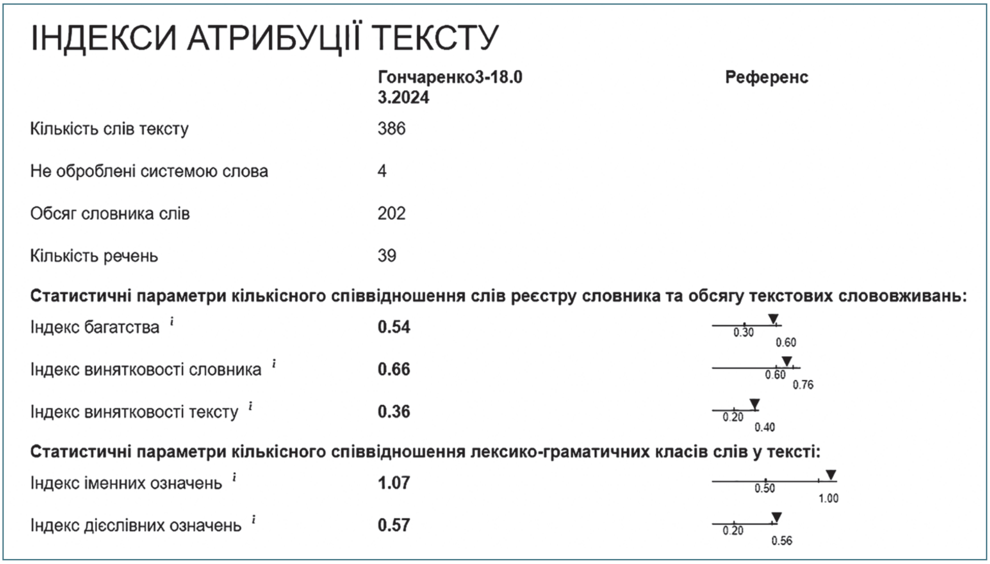
3. GENERAL CHARACTERISTICS OF THE TEXTATTRIBUTOR SYSTEM OPERATING

During the development of TextAttributor, the following tasks were accomplished: 1) Theoretical Linguistics: A methodology was developed for formalized morphological, statistical, stylometric, attributional, and sentiment analyses; 2) Software Engineering: The structure of linguistic databases and the software for the aforementioned automatic analyses were developed; 3) Experimental Linguistics: The system was tested on Ukrainian media texts, and an analytical module for automatic stylometric examination was implemented. The automated linguistic system, implemented as a web application, processes user-entered media-style texts, generating numerical values for statistical parameters that characterize the text's attributes. The system follows a structured operation sequence:

1. **Text Tokenization:** Breaking down the text into sentences and words for analysis.
2. **Morphological Labeling:** Assigning labels to words based on their grammatical forms.
3. **Contextual Analysis:** Updating the morphological labels concerning context.
4. **Syntactic Analysis:** Establishing binary relationships between words in sentences to understand their grammatical structure.
5. **Syntactic Relation Establishment:** Determining syntactic connections based on predefined rules.
6. **Emotionally Negative Vocabulary Matching:** Identifying words from a predefined set of negative vocabulary.
7. **Statistical Parameter Evaluation:** Assessing statistical parameters for the input text using automatically compiled frequency vocabularies.
8. **Visualization of Results**: Presenting statistical parameterization outcomes graphically, with empirical values and confidence intervals.
9. **Comparison of Results**: Visualizing statistical parameterization outcomes for two media-style texts, facilitating comparison.
10. **Euclidean Distance Evaluation:** Quantifying the dissimilarity between two media-style texts.
11. **Expert Opinion Generation**: Generating two expert opinions: one on text attribution and another on text toxicity through linguistic examination.
12. **Toxicity Detection and Authorship Identification:** Leveraging machine learning techniques, particularly neural network models, for detecting toxicity and identifying authors.

The results of the system are organized in the following partitions: Text Attribution Index Group, Text Attribution Expert Opinion, Comparison of Text Attribution, Linguistic Expertise of Text Toxicity, and Neural Network Opinions.

In **Text Attribution Index Group** (Fig. 1), statistical parameters are organized based on the input text: column 1 displays the Index title, column 2 presents the empirical numeric value, and column 3 provides a visual reference by comparing the empirical value of the index with confidence interval thresholds (lower and upper) derived from Ukrainian media style texts. The empirical numeric value is represented on the scale by a filled inverted triangle.



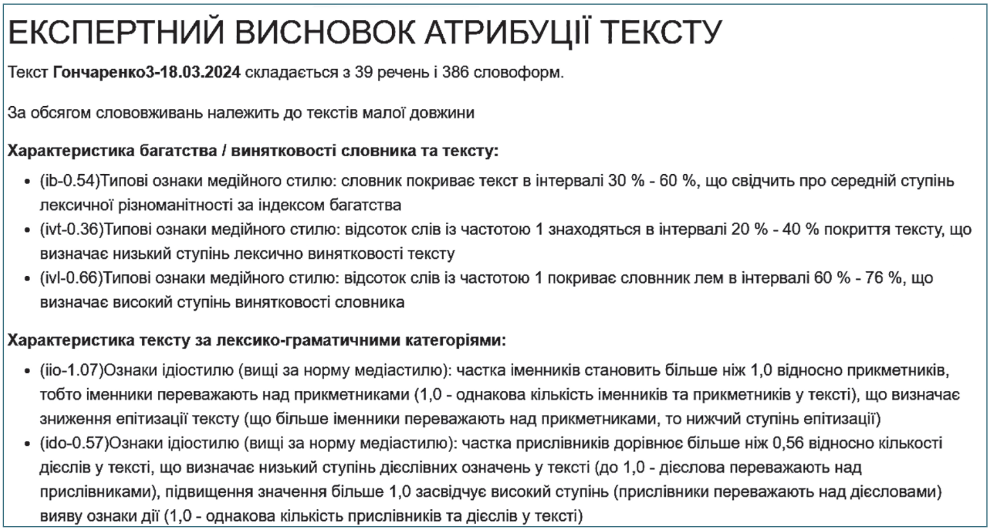
ІНДЕКСИ АТРИБУЦІЇ ТЕКСТУ

	Гончаренко3-18.0 3.2024	Референс
Кількість слів тексту	386	
Не оброблені системою слова	4	
Обсяг словника слів	202	
Кількість речень	39	
Статистичні параметри кількісного співвідношення слів реєстру словника та обсягу текстових слововживань:		
Індекс багатства i	**0.54**	▼ 0.30 0.60
Індекс винятковості словника i	**0.66**	▼ 0.60 0.76
Індекс винятковості тексту i	**0.36**	▼ 0.20 0.40
Статистичні параметри кількісного співвідношення лексико-граматичних класів слів у тексті:		
Індекс іменних означень i	**1.07**	▼ 0.50 1.00
Індекс дієслівних означень i	**0.57**	▼ 0.20 0.56

FIGURE 1: A fragment of the text attribution result

In **Text Attribution Expert Opinion** (Fig. 2), we present conclusions regarding the typical or individual linguistic statistical features extracted from the analyzed text for each estimated index. These conclusions are derived from answers to the question: Does the numerical value of the index fall within the confidence interval of the media style of the Ukrainian language? Based on a comparison of numerical values with threshold values of the confidence interval, the system generates three possible answers (typical signs of media style, signs of idiostyle lower than normal media style, and signs of idiostyle higher than normal media style).

In **Comparison of Text Attribution** (Fig. 4, subsection 4.2), upon selecting "Calculate Vector Distance", tabulated data is promptly updated as follows: Column 1 lists user-entered texts along with the previously analyzed text; Column 2 displays the number of sentences in each text; Column 3 shows the word count; Column 4 presents the Euclidean distance between the analyzed text and each text in the table; Column 5, labelled "Compare", initiates an automatic comparison of statistical indices between the analyzed text and the selected text. Upon activating the corresponding "Compare" element, comprehensive information on the statistical comparison between two texts is provided in the "Text Attribution Indices" group (Fig. 3).



ЕКСПЕРТНИЙ ВИСНОВОК АТРИБУЦІЇ ТЕКСТУ

Текст **Гончаренко3-18.03.2024** складається з 39 речень і 386 словоформ.

За обсягом слововживань належить до текстів малої довжини

Характеристика багатства / винятковості словника та тексту:

- (ib-0.54)Типові ознаки медійного стилю: словник покриває текст в інтервалі 30 % - 60 %, що свідчить про середній ступінь лексичної різноманітності за індексом багатства
- (ivt-0.36)Типові ознаки медійного стилю: відсоток слів із частотою 1 знаходяться в інтервалі 20 % - 40 % покриття тексту, що визначає низький ступінь лексично винятковості тексту
- (ivl-0.66)Типові ознаки медійного стилю: відсоток слів із частотою 1 покриває словнник лем в інтервалі 60 % - 76 %, що визначає високий ступінь винятковості словника

Характеристика тексту за лексико-граматичними категоріями:

- (iio-1.07)Ознаки ідіостилю (вищі за норму медіастилю): частка іменників становить більше ніж 1,0 відносно прикметників, тобто іменники переважають над прикметниками (1,0 - однакова кількість іменників та прикметників у тексті), що визначає зниження епітизації тексту (що більше іменники переважають над прикметниками, то нижчий ступінь епітизації)
- (ido-0.57)Ознаки ідіостилю (вищі за норму медіастилю): частка прислівників дорівнює більше ніж 0,56 відносно кількості дієслів у тексті, що визначає низький ступінь дієслівних означень у тексті (до 1,0 - дієслова переважають над прислівниками), підвищення значення більше 1,0 засвідчує високий ступінь (прислівники переважають над дієсловами) вияву ознаки дії (1,0 - однакова кількість прислівників та дієслів у тексті)

FIGURE 2: An example of generated expert opinion

In **Linguistic Expertise of Text Toxicity** (subsection 4.3), a lexical-semantic interpretation of the calculated toxicity index is presented (Fig. 6).

In **Neural Network Opinions** we present the output of our deep learning model, which quantifies text toxicity on a scale from 0 (non-toxic) to 1 (high toxicity). We also assess text similarity to known authors, with values ranging from 0 (no similarity) to 1 (high similarity), based on the authors the model was trained on. Results are shown only for tests with a similarity measure exceeding 0.1. Our current system operates in demo mode, utilizing a limited corpus of author texts to gauge similarity with user-entered text.

4. ATTRIBUTION OF UKRAINIAN-LANGUAGE TEXTS: EXPERIMENTS, RESULTS AND DISCUSSIONS

4.1. Statistical Parameterization Indices

The developed stylometric model solves two tasks:

1) Identifying individual features of an author’s style (stylometry);

2) Assessing the similarity between two texts based on linguistic and statistical parameters (attribution).

19 parameters are utilized to statistically analyze the user's input text (refer to Fig. 1 for the Text Attribution Index Group). The initial four parameters encompass basic quantitative data: word count, unprocessed words, dictionary size, and sentence count. The word count excludes vocabulary not processed by the system, such as dialectisms, muscovitisms, occasionalisms, etc. The remaining 15 parameters consist of computable indices that yield scalar quantities.

For the stylometric analysis, we are going to validate Hypothesis 1. This hypothesis states that by analyzing empirical data encompassing statistical parameters within and beyond the confidence interval of the Ukrainian media style, it is possible to discern unique authorial features within a given media text. Two sets of confidence intervals were calculated for each linguistic parameter based on Ukrainian media text samples: one for texts over 1000 words (sample size: 124,576 words) and one for texts up to 1000 words (sample size: 15,635 words).

The confidence interval for each statistical parameter was established using the standard deviation, σ, of the Average Numerical Value of the Index (ANVI) within the text sample, denoted as ANVI ± σ. The standard deviation estimates how much the empirical numerical values of statistical indices may deviate from the average numerical value of the index within the media style. To ensure accuracy, the confidence interval for σ was determined through a linguistic and statistical experiment. Consequently, the system offers three potential conclusions when testing Hypothesis 1: 1) the index's numerical value falls within the confidence interval; 2) the index's numerical value is below the lower threshold of the interval; or 3) the index's numerical value exceeds the upper threshold of the interval.

Let us consider the outcomes derived from the parameterization of Text 1 authored by Oleksii Honcharenko the blogger (see Fig. 1). This analysis is grounded on statistical parameters and their subsequent interpretation encapsulated in the generated expert opinion as illustrated in Fig. 2.

The statistical metrics concerning vocabulary and text volume reveal a quantitative relationship. In Text 1, all these indicators are within the typical range for the media style:

1) **The Variety Index (IB)** represents the ratio of unique words to the overall text size, serving as a measure of lexical richness. The IB value of 0.54 falls within the 30%–60% range, indicating a moderate level of lexical diversity;

2) **The Text Exclusiveness Index (IVT)** expresses the ratio of the number of hapax legomena occurring once in the text to the text volume. IVT of 0.36 falls within the 20%–40% range, indicating low lexical exclusiveness);

3) **The Vocabulary Exclusiveness Index (IVL)** expresses the ratio of the number of hapax legomena occurring only once in the text to the total vocabulary volume. IVL of 0.66 falls within the 60%–76% range, indicating a high degree of lexical exclusiveness.

Next, let us consider the statistical parameters of the quantitative correlation of lexical and grammatical word classes in Text 1:

1) **The Nominal Attributes Index, or Epithetization, (IIO)** expresses the ratio of the number of nouns to the number of adjectives. IIO of 1.07 is higher than the media style norm of 1.0, indicating a prevalence of nouns over adjectives resulting in fewer epithets in the text;

2) **The Verb Attributes Index (IDO)** expresses the ratio of the number of adverbs to the number of verbs. IDO of 0.57 is higher than the media style norm of 0.56. This proportion of adverbs indicates a low degree of verb attributes in the text;

3) **The Nominalization Degree (STN)** expresses the ratio of the number of nouns to the number of verbs. STN of 1.76 is lower than the media style norm of 3.00. A decrease in this index indicates a predominance of verbs over nouns and a low degree of nominalisation of the text;

4) **The Pronominalization Index (IPRO)** expresses the ratio of the number of personal pronouns to the volume of the text word usage. IPRO of 0.09 is higher than the media style norm and is a sign of the idiostyle. Personal pronouns cover more than 7% of the text, indicating the degree of coherence in the test;

5) **The Modality Index (IMOD)** expresses the ratio of the number of particles to the number of words in the text. IMOD of 0.07 is higher than the media style norm of 4% and is a sign of the idiostyle. Particles cover more than 4% of the text, which determines the degree to which the author expresses various evaluations of phenomena and situations in the text;

6) **The Substantivity Index (ISUB)** expresses the ratio of the number of nouns to the volume of text word usage. ISUB of 0.26 is lower than the media style norm. Nouns cover less than 30% of the text, which determines the low degree of static text.

Let us explain the statistical parameters of the quantitative ratio of phrases and sentences in Text 1:

1) **The Dynamism Index (IDYN)** expresses the ratio of the number of verb phrases to the number of noun phrases. IDYN of 0.74 is higher than the media style norm of 0.6 and is a sign of the idiostyle. The prevalence of verb phrases over noun phrases (IDYN greater than 1.0) determines the dynamic and rapid unfolding of events in the text's meaning;

2) **The Coherence Index (IZV)** expresses the ratio of the number of prepositions and conjunctions to the number of sentences in the text. IZV of 0.66 falls within the typical range of 0.45 to 0.90 for media style, indicating a predominance of the number of sentences over the number of prepositions and conjunctions. This determines the average degree of coherence between the realities, phenomena, and situations described in Text 1.

Particular attention will be given to psycholinguistic statistical parameters (Chuhunov 2009) in Text 1:

1) **The Trager coefficient (KT)** expresses the ratio of the number of verbs to the number of adjectives in the text, indicating the emotional stability/ instability of the speaker. KT of 0.61 is higher than the media style norm of 0.5 and is a sign of the idiostyle. According to psychological interpretation, this indicates the low emotionality of the author/speaker, indecisiveness to take active practical actions and anxiety. The value in the range of 1.35 ± 0.05 determines normal emotional stability and regulation for the author;

2) **The Action Certainty Coefficient (KOD)** expresses the ratio of the number of verbs to the number of nouns in the text. This ratio indicates the degree of authorial socialization and the syntactic completeness of the statement. KOD of 0.57 falls within the typical range of 0.30 to 0.60 for media style, indicating a low degree of objectified action coefficient. According to psychological interpretation, this suggests a low level of emotionality in the author/speaker, indecision regarding active practical actions, and anxiety. A value within the range of 1.35 ± 0.05 determines normal emotional stability and regulation of emotions by the author, which, based on linguistic features, indicates a normal degree of syntactic completeness in sentences and correct syntactic structure;

3) **The Aggressiveness Coefficient (KA)** expresses the ratio of the number of verbs and verb usage to the total number of words in the text, which marks the aggressiveness of the language in the psycholinguistic aspect. KA of 0.15 is higher than the media style norm of 13% and is a sign of idiostyle. According to psychological interpretation, values below 60% indicate a low degree of aggressiveness, the author's suppressed emotions, and his indecision. Approaching 60% determines the normal degree of aggressiveness and emotional stability of the author, while values above 60% indicate high aggressiveness of the author and emotional excitement.

Finally, let us consider the semantic statistical parameter of Text 1: **The Text Toxicity Index (ITOX)** is calculated from the frequency of negative vocabulary in the text (for more details, see section 4.3). ITOX value of 0.36 falls within the typical range of 0.1 to 0.7 for media style. ITOX shapes cognitive and pragmatic attitudes towards negative emotions.

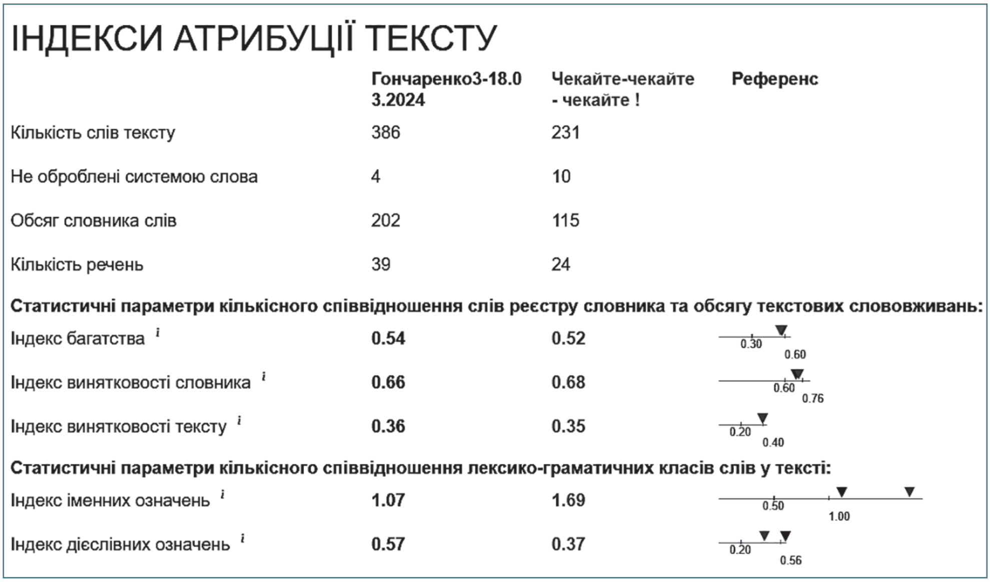
4.2. Text comparison by statistical parameters

The task of assessing the degree of similarity between two texts involved testing Hypothesis 2. This hypothesis states that if the empirical data resulting from the statistical parameterization of two texts show only a small difference in the numerical values of the indices, then these texts belong to the same author. This hypothesis is tested within the system in two ways:

1) by comparing the numerical values of each index in the two texts;

2) by determining the vector distance between the two texts.

These problems are implemented in the Comparison of Text Attribution section. The first task involves comparing empirical data of texts using indices on the scale of media style confidence interval (Fig. 3). As depicted in Fig. 3, the quantitative characteristics of the parameters of both texts are displayed in different colours for the convenience of the user. The visualization compares the empirical values of the indices of both texts with confidence interval thresholds (lower and upper) derived from Ukrainian media-style texts.

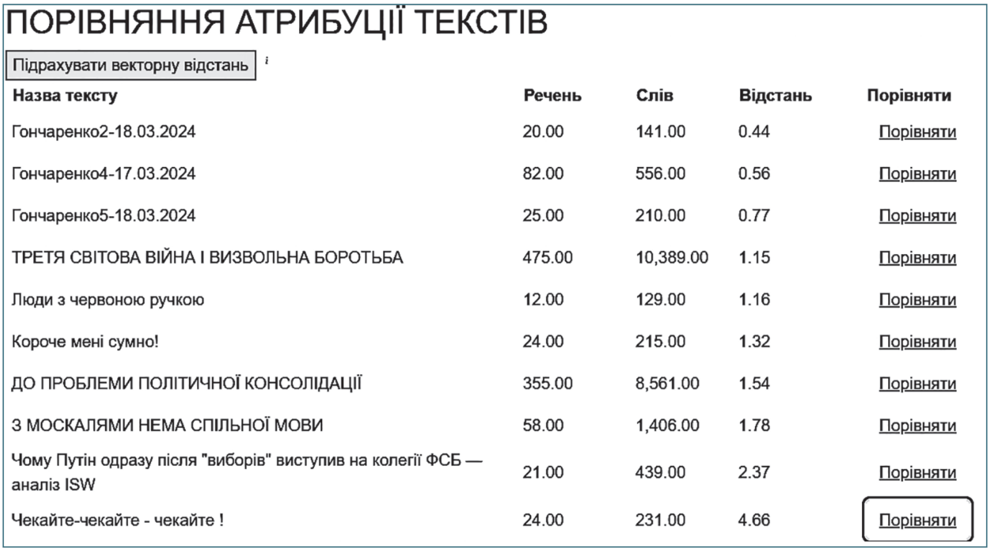


ІНДЕКСИ АТРИБУЦІЇ ТЕКСТУ

	Гончаренко3-18.03.2024	Чекайте-чекайте - чекайте !	Референс
Кількість слів тексту	386	231	
Не оброблені системою слова	4	10	
Обсяг словника слів	202	115	
Кількість речень	39	24	
Статистичні параметри кількісного співвідношення слів реєстру словника та обсягу текстових слововживань:			
Індекс багатства *i*	**0.54**	**0.52**	0.30 0.60
Індекс винятковості словника *i*	**0.66**	**0.68**	0.60 0.76
Індекс винятковості тексту *i*	**0.36**	**0.35**	0.20 0.40
Статистичні параметри кількісного співвідношення лексико-граматичних класів слів у тексті:			
Індекс іменних означень *i*	**1.07**	**1.69**	0.50 1.00
Індекс дієслівних означень *i*	**0.57**	**0.37**	0.20 0.56

FIGURE 3: Comparing the vector distance of texts

The second task is to calculate the vector distance between two texts using the Euclidean distance formula. This involves summing the squares of the differences between the numerical values of 15 statistical parameters of the two texts. The square root of this sum is then calculated. The numerical values of the vector distance indicate the extent of statistical dissimilarity between the texts and provide a measure of the distance between them. While the fluctuations in the vector distance value for media texts are not well-known, the distance can be 0 when comparing identical texts. This implies that the texts are different if the vector distance is greater than 0. Moreover, the greater the numerical value of the vector distance, the greater the linguistic and statistical differences between the texts. Euclidean distance comparison is presented in a tabular format. In Column 4 of Fig. 4, numerical values represent the Euclidean distances. Fig. 4 displays the comparison of the analyzed text by Oleksii Honcharenko (Text 1) with three other texts by the same author: line 1 for Text 2 (distance 0.44); line 2 for Text 3 (distance 0.56); line 3 for Text 4 (distance 0.77). The empirical data demonstrate that the similarity among texts by a single author remains within 1, while texts by other authors (lines 4–10) exhibit distances greater than 1.



ПОРІВНЯННЯ АТРИБУЦІЇ ТЕКСТІВ

Назва тексту	Речень	Слів	Відстань	Порівняти
Гончаренко2-18.03.2024	20.00	141.00	0.44	Порівняти
Гончаренко4-17.03.2024	82.00	556.00	0.56	Порівняти
Гончаренко5-18.03.2024	25.00	210.00	0.77	Порівняти
ТРЕТЯ СВІТОВА ВІЙНА І ВИЗВОЛЬНА БОРОТЬБА	475.00	10,389.00	1.15	Порівняти
Люди з червоною ручкою	12.00	129.00	1.16	Порівняти
Короче мені сумно!	24.00	215.00	1.32	Порівняти
ДО ПРОБЛЕМИ ПОЛІТИЧНОЇ КОНСОЛІДАЦІЇ	355.00	8,561.00	1.54	Порівняти
З МОСКАЛЯМИ НЕМА СПІЛЬНОЇ МОВИ	58.00	1,406.00	1.78	Порівняти
Чому Путін одразу після "виборів" виступив на колегії ФСБ — аналіз ISW	21.00	439.00	2.37	Порівняти
Чекайте-чекайте - чекайте !	24.00	231.00	4.66	Порівняти

FIGURE 4. Systematization of the Euclidean distance between texts

A comparison of the statistical parameters of the texts convinces us that the proposed attribution methodology reveals a stylometric model of Ukrainian-language texts. This model can also be used to determine the authorship of the text. We will confirm this with a visual presentation of the attribution of 7 texts by three different authors according to 15 statistical parameters using the Uniform Manifold Approximation and Projection (UMAP) method (McInnes *et al.* 2020), the algorithm of which is also implemented in the TextAttributor system. This technique allows to reduce the dimensionality of the feature space by projecting a 15-dimensional space onto the plane. In Fig. 5, points of different shapes (P1–P7 for Author 1, F1–F7 for Author 2, M1–M7 for Author 3) correspond to 7 randomly selected texts from different authors.

As we can see, after dimensionality reduction, the point mappings of the texts from the three authors can be visibly segregated into distinct areas. This example confirms that the parameters defined by the TextAttributor are important for describing the stylistic features of texts and for determining the authorship of the text (even in a reduced form).

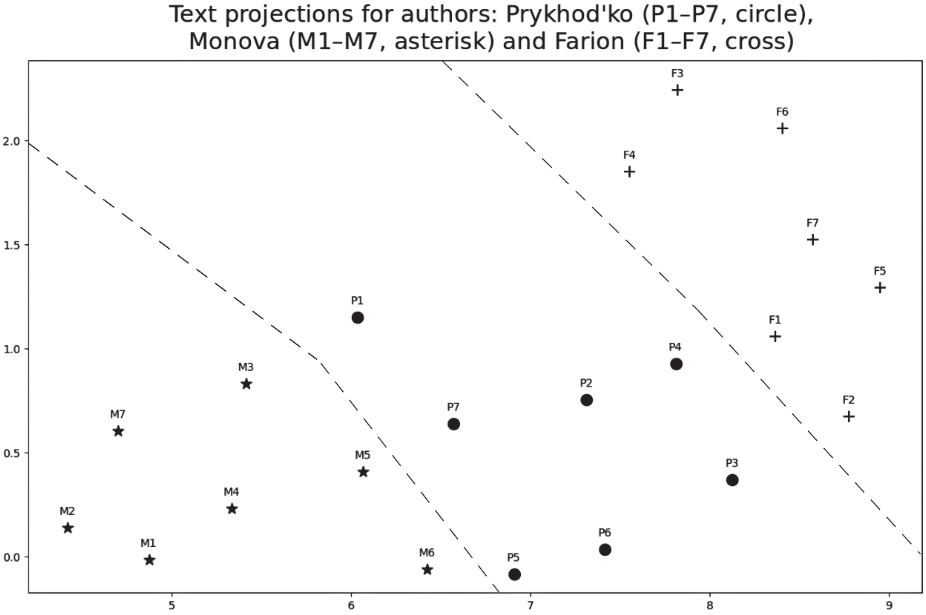


FIGURE 5: Graphical representation of text attribution using the UMAP method

4.3. Indexing the text's negative sentiment

While developing a system of statistical parameters to determine the style and authorship of the text, we recognized that text toxicity could serve as an attribution index. We therefore created a separate module to automatically determine the toxicity index in the TextAttributor.

Today, the information component has become particularly important in hybrid warfare. Therefore, the material of our study was a research corpus of online media texts of political discourse with a volume of 10 million words. We use the term "toxic text" in a broad sense. These texts are characterized by harassment, threats, obscenity, cyber-bullying, trolling, and identity-based hate texts, and contain emotiogens, which are phenomena and objects that cause negative emotions in a person (e.g. *war*, *air raid*, *corruption*).

This project aimed to determine the potential for the realization of negative sentiment in Ukrainian texts and to develop an automated system for identifying toxic content. This task comprised two consecutive subtasks: first, developing a system of toxic text linguistic examination with determination of the toxicity indices through dictionary analysis and rule-based methods; second, constructing a training dataset for machine learning purposes and the development of a neural network for predicting text toxicity indices.

Let us look at the completion of the first task. This involved the construction of a lexicographic database housing three distinct dictionaries:

1) Emotiogen dictionary: A compilation of 5000 words (according to the meanings of lexical-semantic variants) with negative sentiment tones, rated on a scale of –2. Examples include *immoral*, *impunity*, *bribery* and *steal*;

2) Hate Speech Dictionary: Comprising 3000 words, including names of individuals (1620), obscene terms (613), and abusive language (787). Examples include *westerner* (*западенеуь*: a derogatory name for people from the western regions of Ukraine) and *huckster*;

3) Toxic Phrases: A collection of 1500 idiomatic expressions conveying negative emotions. Examples include *grimaces like a monkey*, *kisses his arse* and *opens his mouth*.

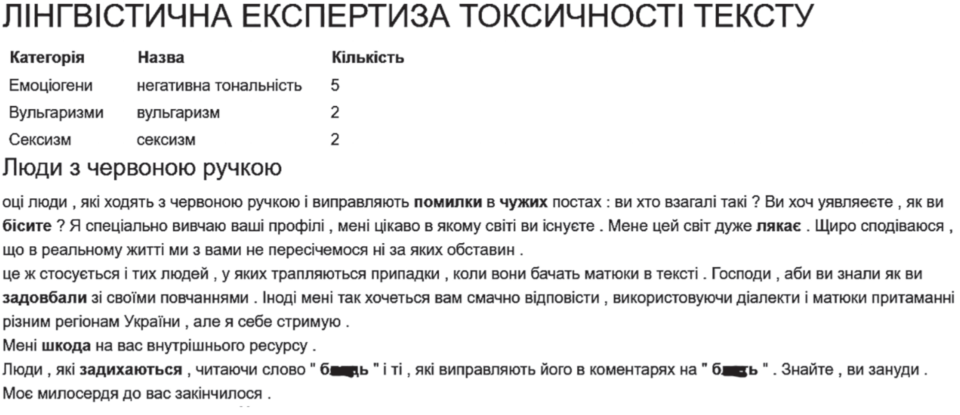
Each entry in these lists was annotated with semantic features, with the Hate Speech Dictionary having 18 features and the Toxic Phrases list having 26. For instance, words in the Hate Speech Dictionary were tagged with characteristics like 's' for sexism, 'r' for racism, and 'e' for ageism. This lexicographic database, a multiparametric system, assigns semantic features to units (words or phrases) which, combined with frequency data from analyzed text, allows for toxicity analysis.

The toxicity index of the text is computed using the formula:

$$Itox = (e + |K|\ (m + t)) \ /\ n * 10,$$

Here, *n* is the text volume; *e* is the count of emotional words; *m* is the count of hate speech words; *t* is the count of toxic phrases; *K* is a coefficient equal to –2, it emphasises hate speech words and toxic phrases, which on a scale of five digits (+2, +1, 0, −1, −2) corresponds to −2. Index values vary from 0 to +1.

For instance, the toxicity index of the text "People with a red pen (Люди з червоною ручкою)" (12 sentences, 129 words) is 1.01, indicating its high toxicity, as the threshold for such short texts ranges from 0.1 to 0.7. Simultaneously, the system generates the results of an automatic linguistic examination of the text toxicity (refer to Fig. 6), comprising: 1) a statistical map of the semantic classes of the negative vocabulary according to the classification markers of lexicographic lists (emotiogens – 5, vulgarisms – 2, sexism – 2); 2) a text with specific words of negative sentiment, verbalizing the categories of the statistical map. Let us take a look at an excerpt of the analyzed real-life text (Fig. 6): *those people who walk around with a red pen and correct **mistakes** in **other people's** posts: who even are you? Do you have any idea how **annoying** you are? I study your profiles on purpose, I'm curious about the world you exist in. This world **scares** me a lot. I sincerely hope that in real life we will not cross paths under any circumstances. (оці люди, які ходять з червоною ручкою і виправляють **помилки** в **чужих** постах: ви хто взагалі такі? Ви хоч уявляеєте, як ви **бісите**? Я спеціально вивчаю ваші профілі, мені цікаво в якому світі ви існуєте. Мене цей світ дуже **лякає**. Щиро сподіваюся, що в реальному житті ми з вами не пересічемося ні за яких обставин).* Here, the system automatically highlights in bold the following words with a negative component: *mistakes (помилки), other people's (чужих), annoying (бісите), scares (лякає).*



ЛІНГВІСТИЧНА ЕКСПЕРТИЗА ТОКСИЧНОСТІ ТЕКСТУ

Категорія	Назва	Кількість
Емоціогени	негативна тональність	5
Вульгаризми	вульгаризм	2
Сексизм	сексизм	2

Люди з червоною ручкою

оці люди , які ходять з червоною ручкою і виправляють **помилки** в **чужих** постах : ви хто взагалі такі ? Ви хоч уявляеєте , як ви **бісите** ? Я спеціально вивчаю ваші профілі , мені цікаво в якому світі ви існуєте . Мене цей світ дуже **лякає** . Щиро сподіваюся , що в реальному житті ми з вами не пересічемося ні за яких обставин .
це ж стосується і тих людей , у яких трапляються припадки , коли вони бачать матюки в тексті . Господи , аби ви знали як ви **задовбали** зі своїми повчаннями . Іноді мені так хочеться вам смачно відповісти , використовуючи діалекти і матюки притаманні різним регіонам України , але я себе стримую .
Мені **шкода** на вас внутрішнього ресурсу .
Люди , які **задихаються** , читаючи слово " **б[redacted]ь** " і ті , які виправляють його в коментарях на " **б[redacted]ь** " . Знайте , ви зануди .
Моє милосердя до вас закінчилося .

FIGURE 6: A fragment of text toxicity automated linguistic examination

5. MACHINE LEARNING APPROACHES: EXPERIMENTS, RESULTS AND DISCUSSIONS

Let us focus on the second operation mode of the TextAttributor system dedicated to machine learning techniques. At the input for model training, we have text corpora labelled according to the goal of each subtask: (a) toxicity detection text information and (b) authorship identification. Each corpus is divided into training and control sets, as well as, for sufficiently large corpus, a validation set. The parameters of the chosen problem-solving model are estimated on the training set. Hyperparameters of the model are adjusted by the validation set. The final performance indicators of the model are measured on a control set, taking into account the availability of computing resources, which are necessary for the model's operability. A corpus for each subtask has been developed in parallel with the TextAttributor machine learning component, so the current models were trained on relatively small data including short Internet texts of Ukrainian-language blogs, comments, articles, etc.

5.1. Toxicity detection

As a result of a series of experimental studies, a computationally efficient architecture was chosen based on the fastText method and its tools (Joulin *et al.* 2016). This method provides word embeddings and estimates the probability distribution of documents according to predefined classes. Words are presented in vector form based on automatic word splitting into parts (subwords), which simulates the openness of the dictionary. The subword presentation is important since the Ukrainian language is highly inflective, and plenty of unseen words as well as words with spelling errors must be covered. The effectiveness of the used approach is also determined by the technical conditions of system operation and available textual resources for model training.

A prepared corpus of approximately 12,000 text documents was used for binary classification to detect toxic content. The best result, based on the generalized metric (F1 = 79.4%), was achieved with the following hyperparameter settings: a word vector dimensionality of 56, an initial learning rate of 0.15, 500 training epochs, lexical context represented by bigrams, subword lengths ranging from 2 to 5 characters, and a decision threshold of 0.4. Additionally, the model demonstrated a sensitivity of 85% while achieving an accuracy of approximately 70%.

5.2. Authorship identification

Experimental research of models for text authorship identification was carried out using a part of the corpus of publicly available socio-political texts, in which it is possible to identify the person who is the author of the document. A total of 702 texts from four authors who published the largest number of documents between 60 and 1000 characters were selected. The largest number of publications by the author is 304 (115 thousand characters), and the smallest is 109 (49 thousand characters). For testing, 25 texts for each selected author were randomly sampled. The rest of the documents made up the training set.

The best result, according to the generalized metric (F1 = 81%), was achieved with the following hyperparameter values: a word embedding space dimension of 50, an initial learning rate of 0.1, 50 training epochs, lexical context represented by bigrams, subword lengths ranging from 2 to 5 characters, and a decision-making threshold of 0.5. Notably, for certain model configurations, the sensitivity to the author with the largest number of documents in the dataset reached 100%, with an accuracy exceeding 90%. These results demonstrate the feasibility of developing an effective system capable of accurately covering texts by authors who are sufficiently represented in the training set.

The trained neural network models were integrated into the TextAttributor system in the “client-server” architecture, where the results of two tasks are presented: (a) toxicity determination and (b) authorship detection. The result of the system’s work on the text of the “Law of Ukraine on Higher Education”: Toxicity index (estimate in the form of probability: how toxic the text is) – 0.04, similarity with authors (estimations in the form of the probability of authorship for authors known to the system): 0.49 for I. Farion, 0.30 for Oleksii Honcharenko, and 0.22 for Mariana Bezuhla.

6. CONCLUSIONS

The TextAttributor system is currently undergoing testing phases to address attribution and toxicity determination issues within Ukrainian-language texts. This system is both convenient and effective for various technological studies. Over 5 months, the TextAttributor web application has automatically analyzed 784 Ukrainian texts by 226 users.

Our findings demonstrate the effectiveness of our method, which involves quantizing verbal elements based on formal grammatical and semantic parameters, particularly for negative emotionality. This is achieved through a combination of dictionary and rule-based approaches, complemented by machine learning techniques. Experimental machine learning research for toxicity detection and authorship identification by text allowed us to obtain results comparable to results of experiments with similar model characteristics (number of documents and authors) reported for other languages. Subword models showed enhanced robustness against lexical openness and textual errors.

The conducted research paves the way for solving such problems for Ukrainian as detecting and monitoring toxic content, hate speech and misinformation, authorship verification and deobfuscation, author profiling and diarizing, psycholinguistic profiling, style modelling, and tracking the source of fake content, etc. The developed text analysis and attribution scheme can also be applied to other languages.

Looking ahead, we plan to integrate our automatic text attribution and toxicity detection tool with semantic, syntactic, psycholinguistic, and sociolinguistic analysis. This integration will give us a deeper understanding of the impact on the reader. By using semantic analysis of lexical taxonomy, we aim to identify narrative elements inherent in the text, thereby increasing the reliability of text attribution in authorship identification procedures. In addition, we plan to expand our text corpora and utilize multilingual large language models to further extend the capabilities of our system as well as to implement tools for statistical text comparison for all styles of the Ukrainian language.

Acknowledgements

The system has been created within a project supported by the British Embassy in Kyiv and carried out by a team of educators and researchers of the Ukrainian Language and Applied Linguistics Chair, Taras Shevchenko National University of Kyiv.

We would like to express our sincere appreciation to the British Embassy in Kyiv for their financial support of this research project. We are grateful to Svitlana Yavorska and Iryna Bezkorovayna for their guidance and assistance throughout the project. Special thanks to Kostiantyn Goncharenko for the organization and documentary support of the project. Furthermore, we are grateful to the students of the educational program “Applied (Computer) Linguistics and English Language” of Taras Shevchenko National University of Kyiv. They have contributed their time, expertise, and efforts to forming the corpus of Ukrainian toxic texts. We are grateful to Oksana Tolochko, a graduate of our bachelor’s degree program, for creating a tonal dictionary of the Ukrainian language, which we used to compile a lexicographic list of negative emotiogens. Additionally, we are deeply appreciative of the efforts made by Mykola Kostikov in the data collection process. We would like to express our sincere gratitude to all members of the research community who contributed to this study with their advice and consultations.

REFERENCES

ALIAS – *Automated Linguistic Identification & Assessment System*. Available at: https://aliastechnology.com/alias-overview/.

Alkomah Fatimah, Ma Xiaogang 2022: A Literature Review of Textual Hate Speech Detection Methods and Datasets. – *Information* 13(6), 273. DOI: 10.3390/info13060273.

Argamon Shlomo 2007: Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. – *Literary and Linguistic Computing* 23, 131–147. DOI: 10.1093/llc/fqn003.

Bonetti Andrea, Martínez-Sober Marcelino, Torres Julio, Vega Jose, Pellerin Sebastien, Vila-Francés Joan 2023: Comparison between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks. – *Applied Sciences* 13(10), 6038. DOI: 10.3390/app13106038.

Buk Solomija 2021: *Long prose fiction by I. Franko: electronic corpus, frequency dictionaries and other interdisciplinary contexts*, LNU imeni Ivana Franka [Ivan Franko National University of Lviv].

Burrows John 2002: "Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship. – *Literary and Linguistic Computing* 17(3), 267–287. DOI: 10.1093/llc/17.3.267.

Burrows Steven, Uitdenbogerd Alexandra, Turpin Andrew 2014: Comparing techniques for authorship attribution of source code. – *Software: Practice and Experience* 44(1), 1–32. DOI: 10.1002/spe.2146.

Canhasi Ercan, Kadriu Arbana, Misini Arta 2022: A Survey on Authorship Analysis Tasks and Techniques. – *SEEU Review* 17(2), 153–167. DOI: 10.2478/seeur-2022-0100.

Chuhunov Vadym 2009: *Diahnostyka v psykhoterapii ta psykhoterapevtychnyi diahnoz*, Kharkiv: Nauka.

Darchuk Natalia, Sorokin Viktor 2022: Parameterization of the Ukrainian Text Corpus based on parsing. – *Computational Linguistics and Intelligent Systems*, Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS-2022) 1: *Main Conference*, eds. V. Lytvyn, N. Sharonova, I. Jonek-Kowalska, A. Kowalska-Styczen, V. Vysotska, Ye. Kupriianov, O. Kanishcheva, O. Cherednichenko, Th. Hamon, N. Grabar, Poland, 12–13 May, 2022, 256–265.

Darchuk Natalia, Zuban' Oksana, Sorokin Viktor 2024: On stylistic differentiation in Ukrainian poetic speech based on the syntactic structure of sentences. – *Bulletin of Taras Shevchenko National University of Kyiv. Literary Studies. Linguistics. Folklore Studies* 1(35), 5–10. DOI: 10.17721/1728-2659.2024.35.01.

Darčuk Natalija 2013: *Kompjuterne anotuvannja ukrajins'koho tekstu: rezul'taty i perspektyvy*, Kyiv: Osvita Ukrajiny.

Darčuk Natalija, Vasyl'jeva Iryna, Vasyl'jev Oleksij 2021: Vektorna model' analizu stylistyky tekstiv. – *Ukrajins'kyj fizyčnyj žurnal* 66(5), 373–378.

Eder Maciej 2015: Does size matter? Authorship attribution, small samples, big problem. – *Digital Scholarship in the Humanities* 30(2), 167–182. DOI: 10.1093/llc/fqt066.

Eder Maciej, Rybicki Jan 2013: Do birds of a feather really flock together, or how to choose training samples for authorship attribution. – *Literary and Linguistic Computing* 28(2), 229–236. DOI: 10.1093/llc/fqs036.

Evert Stefan, Proisl Thomas, Schöch Christof, Jannidis Fotis, Pielström Steffen, Vitt Thorsten 2015: Explaining Delta, or: How do distance measures for authorship attribution work? – *Corpus Linguistics 2015*: Abstract Book, United Kingdom, 21 July 2015, eds. F. Formato, A. Hardie, Lancaster: UCREL. Available at: https://zenodo.org/records/18308.

Gupta Shriya T. P., Sahoo Jajati K., Roul Rajendra K. 2019: Authorship identification using recurrent neural networks. – *ICISDM' 19*: Proceedings of the 2019 3rd International Conference on Information System and Data Mining, United States, 06 April 2019, New York: Association for Computing Machinery, 133–137. DOI: 10.1145/3325917.3325935.

Hoover David 2004: Testing Burrows's delta. – *Literary and Linguistic Computing* 19(4), 453–475. DOI: 10.1093/llc/19.4.453.

Hoover David 2005: Delta, Delta Prime, and Modern American Poetry: Authorship Attribution Theory and Method. – *ACH/ALLC 2005*: Proceedings of the 17th Joint International Conference of the Association for Computers and the Humanities (ACH) and the Association for Literary and Linguistic Computing, Canada, 15–18 June 2005, University of Victoria: Humanities Computing and Media Centre, 79–80.

Jannidis Fotis, Pielström Steffen, Schöch Christof, Vitt Thorsten 2015: Improving Burrows' Delta – An empirical evaluation of text distance measures. – *DH 2015*: Digital Humanities, Abstracts of the Global Digital Humanities Conference 11, Australia, 29 June – 3 July 2015, Sydney. DOI: https://zenodo.org/records/1321296.

Joulin Armand, Grave Edouard, Bojanowski Piotr, Douze Matthijs, Jégou Hérve, Mikolov Tomas 2016: FastText.zip: Compressing text classification models. – *ArXiv e-prints 1612.03651*. DOI: 10.48550/arXiv.1612.03651.

Karasov Vitalii, Levchenko Olena 2024: Statistical profile of O. Zabuzhko's idiostyle. – *CEUR Workshop Proceedings* 3722: *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent* Systems, Lviv, April 12–13, 2024, 251–264.

Khan Talha F., Anwar Waheed, Arshad Humera, Abbas Syed N. 2023: An Empirical Study on Authorship Verification for Low Resource Language Using Hyper-Tuned CNN Approach. – *IEEE Access* 11, 80403–80415. DOI: 10.1109/ACCESS.2023.3299565.

Khomytska Iryna, Teslyuk Vasyl, Bazylevych Iryna, Karamysheva Iryna 2023: Automated Identification of Authorial Styles. – *CEUR Workshop Proceedings: Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems*, 3396, Kharkiv, April 20–21, 2023. Available at: https://ceur-ws.org/Vol-3396/paper26.pdf.

KUM – *Korpus ukrajins'koji movy*: Linhvistyčnyj portal Mova.info. Available at: http://www.mova.info/corpus.aspx.

LIWC-22 – *Linguistic Inquiry and Word Count*. Available at: https://www.liwc.app.

Lototska Nataliia 2022: Statistical Characteristics of Roman Ivanychuk's Idiolect (Based on Writer's Text Corpus). – *CEUR Workshop Proceedings* 3171, 487–500.

Lupei Maksym, Mitsa Alexander, Repariuk Volodymyr, Sharkan Vasyl 2020: Identification of authorship of Ukrainian-language texts of journalistic style using neural networks. – *Eastern-European Journal of Enterprise Technologies* 1/2(103), 30–36. DOI: 10.15587/1729-4061.2020.195041.

McInnes Leland, Healy John, Melville James 2020: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. – *ArXiv e-prints 1802.03426*. DOI: 10.48550/arXiv.1802.03426.

Meier Tabea, Boyd Ryan, Pennebaker James, Meh Matthiasl, Martin Mike, Wolf Markus, Horn Andrea 2019: "LIWC auf Deutsch": The development, psychometrics, and introduction of DE-LIWC2015. – *PsyArXiv Preprints*. DOI: 10.31234/osf.io/uq8zt.

Rybicki Jan, Eder Maciej 2011: Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words? – *Literary and Linguistic Computing* 26(3), 315–321. DOI: 10.1093/llc/fqr031.

Savoy Jacques 2015: Comparative evaluation of term selection functions for authorship attribution. – *Digital Scholarship in the Humanities* 30(2), 246–261. DOI: 10.1093/llc/fqt047.

TextAttributor 1.0 2024: *TextAttributor 1.0: The system for automatic stylometric analysis of Ukrainian media texts*. Available at: ta.mova.info.

Volos Yana, Levchenko Olena 2023: Statistical profile of Valerie Shevchuk's idiostyle (morphological level). – *Slobozhan Scientific Bulletin*, Ser. *Philology*, 3, 32–36. DOI:10.32782/philspu/2023.3.6.

Zuban' Oksana 2019: The Morphemic System Stylometric Analysis of the Ukrainian Poets' Idiostyles: Corpus Based Approach. – *Linhvistyčni studiji* 38, 96–104. DOI: 10.31558/1815-3070.2019.38.15.

Ukrainos žiniasklaidos tekstų automatinės stilometrinės analizės sistema „TextAttributor 1.0“ (metodai, priemonės, funkcionalumas)

SANTRAUKA

Sistema „TextAttributor“ statistiškai parametrizuoja ukrainiečių kalbos tekstus autorystės atpažinimo ir nuotaikų analizės aspektais. Naudodama daugiaparametrinį 15 statistinių indeksų rinkinį, „TextAttributor“ apibūdina autorines stilistines ypatybes. Ji integruoja tokius metodus kaip komponentinė analizė, paskirstymo analizė, kvantavimas ir nuotaikų analizė panaudojant žodynus ir mašininio mokymosi metodus, skirtus nemandagaus teksto ir autorystės išaiškinimui. Sistema apdoroja tekstus taikydama tokenizavimą, morfologinį žymėjimą, kontekstinę ir sintaksinę analizę ir emociškai neigiamo žodyno atitikimą, leidžiantį išsamiai vizualizuoti ir ekspertiškai įvertinti teksto priskyrimą ir neigiamą turinį. Eksperimentai patvirtina sistemos gebėjimą nustatyti unikalius autoriaus bruožus ir įvertinti teksto panašumą, ypač politinio diskurso ir nemandagios komunikacijos žiniasklaidos kontekste Rusijos ir Ukrainos karo metu. Straipsnyje pabrėžiama praktinė ir teorinė „TextAttributor“ reikšmė, demonstruojant jos efektyvumą didelėms teksto apimtims ir tikslioms analitinėms galimybėms. Sėkmingas žodynais, taisyklėmis pagrįstų ir mašininio mokymosi metodų taikymas pabrėžia sistemos tvirtumą ir universalumą. Beta versija ir tebevykdomų tyrimų rezultatai yra nuodugniai išnagrinėti, o būsimos jų kryptys nustatomos siekiant išplėsti kalbinį korpusą ir tobulinti mašininio mokymosi modelius, siekiant pagerinti tikslumą ir pritaikomumą.

The System for Automatic Stylometric Analysis of Ukrainian Media Texts TextAttributor 1.0 (Techniques, Means, Functionality)

Įteikta 2024 m. lapkričio 25 d.

NATALIIA DARCHUK
Taras Shevchenko National University of Kyiv
60 Volodymyrska Street
Kyiv, 01033, Ukraine
n.darchuk@knu.ua

OKSANA ZUBAN
Taras Shevchenko National University of Kyiv
60 Volodymyrska Street
Kyiv, 01033, Ukraine
oxana.zuban@knu.ua

VALENTYNA ROBEIKO
Taras Shevchenko National University of Kyiv
60 Volodymyrska Street
Kyiv, 01033, Ukraine
valentyna.robeiko@knu.ua

YULIIA TSYHVINTSEVA
Taras Shevchenko National University of Kyiv
60 Volodymyrska Street
Kyiv, 01033, Ukraine
Institute of the Ukrainian Language of
the National Academy of Sciences of Ukraine
4 Mykhailo Hrushevskyi Street
Kyiv, 01001, Ukraine
julivoznuk@gmail.com

VICTOR SOROKIN
Taras Shevchenko National University of Kyiv
60 Volodymyrska Street
Kyiv, 01033, Ukraine
victor.sorokin@gmail.com

MYKOLA SAZHOK
Institute for information technologies and systems
of the National Academy of Sciences of Ukraine
40 Akademika Hlushkova Avenue
Kyiv, 03187, Ukraine
sazhok@gmail.com