

UDC 81'322:004.8

Text Mining for Big Linguistic Data in Modern Philological Conditions: Dynamic Conditions of Uncertainty, Turbulence and Risks

Svitlana Goncharenko

Kyiv National University of Technologies and Design, Kyiv

<https://orcid.org/0000-0002-7740-4658>

Abstract. *The article considers the possibilities of Text Mining technology in the analysis of large linguistic data in conditions of uncertainty, turbulence and systemic risks. The main methods of processing text corpora, directions of application of the technology in modern philology and prospects of its integration with methods of artificial intelligence and natural language processing are characterized.*

Keywords: *Digital Philology, Text Mining, Big Linguistic Data, Natural Language Processing (NLP), Language Corpora; Information Uncertainty.*

Introduction. The modern development of digital humanities is accompanied by an unprecedented growth in the volume of text information generated in the scientific and educational [1], social, political, and media spheres (billions of text documents are created every day, presented in different languages, genres, and formats [2]). Such dynamics lead to the formation of Big Linguistic Data arrays, the traditional methods of processing, analysis, and analytics of which are not sufficiently effective [3]. At the same time, the modern global environment is characterized by a high degree of uncertainty, turbulence, and systemic risks, which is described by modern concepts of VUCA, BANI, SHIVA, and WARMI - in these conditions, the requirements for speed, objectivity, and scalability of text information analysis methods are significantly increasing [4]. Philology is gradually transforming from a predominantly descriptive discipline into a field that actively uses methods of intelligent data analysis, artificial intelligence, and computational linguistics [5], [6]. A special place among these technologies is occupied by Text Mining - a set of methods for automatically extracting knowledge from large collections of unstructured texts [7]; unlike classical content analysis, Text Mining technology allows you to identify hidden patterns [8], semantic relationships, thematic structures, emotional characteristics of texts and the dynamics of changes in language processes based on the analysis of millions of documents [9].

The relevance of the study is determined by the need to develop modern approaches to processing large linguistic corpora in the conditions of constant growth in information flows, the complexity of the communication environment and the acceleration of the digital transformation of the humanities.

The aim of the article is to study the capabilities of Text Mining technology during the analysis of large linguistic data, determine the main areas of its application in modern philology, as well as assess the prospects for using methods of intellectual text analysis in conditions of high uncertainty and systemic instability.

The Main Part. The development of Big Data technologies led to a qualitative change in approaches to the study of language processes. If previously the analysis of literary works, journalistic materials or specialized texts was carried out mainly manually, today the researcher has the opportunity to simultaneously analyze millions of text documents of various origins. Such a transformation significantly expands the research possibilities of modern philology, transferring many tasks from the category of local analysis to the field of comprehensive research of large language arrays.

Text Mining technology is understood as a complex of computational methods for the automated detection of knowledge, patterns and hidden dependencies in unstructured text information = this technology combines the methods of Natural Language Processing, machine learning, statistical analysis, artificial intelligence, information retrieval and intelligent data analysis. Their joint use makes it possible to move from the analysis of individual documents to the study of global linguistic processes occurring in the digital information space.

In modern philology, the objects of analysis are national language corpora, electronic libraries, archival documents, scientific publications, media materials, social networks, forums, blogs, official state resources, multilingual databases, and digital collections of cultural heritage. The scale of such information resources requires the use of high-performance algorithms for automatic text processing, since manual analysis becomes practically impossible both from the point of view of time costs and from the point of view of ensuring the objectivity of the results.

Functional capabilities of Text Mining technology significantly exceed traditional methods of computer text processing: because modern algorithms provide automatic determination of the topic of documents, extraction of key terms, recognition of named entities, analysis of tonality, detection of semantic connections between concepts, construction of thematic models, clustering of documents, classification of texts, detection of language anomalies, identification of author's style. Thanks to this, the researcher gets the opportunity to analyze not only the content of the texts, but also the hidden patterns of the development of the language system.

The technological cycle of Text Mining includes several interdependent stages. At the first stage, collection of text data from various digital sources is carried out, followed by their integration into a single information space; then pre-processing of the texts is performed, which involves cleaning data from noise information, normalization of symbols, tokenization, removal of stop words, lemmatization or stemming, morphological analysis and formation of structured representations of the text. The quality of the performance of this stage largely determines the accuracy of the following analytical procedures, especially when working with languages characterized by a developed inflectional system.

The next step is feature extraction, which allows you to convert text information into mathematical models suitable for automated processing. The most common approaches are the Bag-of-Words (BoW) model, TF-IDF, vector representations of words (Word2Vec, FastText, GloVe), contextual language models of the Transformer family, as well as methods for constructing sentence and document embeddings. The use of similar models allows to take into account both the frequency characteristics of words, but also their semantic relations, the context of use, and the peculiarities of the language structure.

Of particular importance in modern philology are the methods of thematic modeling, which automatically highlight latent themes of large text collections. Their application allows us to study the evolution of scientific terminology, the dynamics of socio-political discourse, changes in the conceptual structures of literary works, the peculiarities of the functioning of professional vocabulary, the development of media discourse and the transformation of the linguistic picture of the world under the influence of social processes. Additional opportunities are provided by sentiment analysis, which is used to study the emotional coloring of texts, public sentiment and the peculiarities of communicative behavior of various social groups.

In the context of the SHIVA and WARMI concepts, the ability of Text Mining technologies to promptly respond to a rapidly changing information environment is of particular importance, because the increasing speed of information dissemination, high variability of social processes, information load, active use of generative artificial intelligence and an increase in the amount of unreliable content require the development of intelligent systems for continuous monitoring of text flows. Automated analysis allows for timely detection of new lexical units, changes in terminological systems, the spread of new discursive practices, signs of information manipulations and the transformation of communicative models.

For modern philology, Text Mining technology opens up new areas of research: it provides automated construction of author profiles, identification of intertextual connections, determination of genre affiliation of texts, study of language changes in the diachronic aspect, analysis of translation correspondences, study of the features of professional communication, monitoring of the development of scientific language and research of the processes of formation of digital language culture. The possibilities of comparative analysis of different languages, national corpora and multilingual information resources are significantly expanded.

However, the widespread implementation of Text Mining technologies is accompanied by a number of methodological limitations; These include the heterogeneity of text data, the presence of linguistic ambiguity, the polysemy of lexical units, the influence of cultural context, the insufficient representativeness of individual corpora, the problems of interpreting the results of machine analysis and the need for

expert verification of automatically obtained conclusions. In addition, modern neural network models require significant computing resources, high-quality training samples and constant adaptation to changes in the language environment.

Thus, the development of Text Mining is associated with the integration of artificial intelligence methods, large language models, semantic knowledge graphs, multimodal data analysis and intelligent systems for supporting scientific research. Such integration forms the basis of a new generation of digital tools for philological analysis, focused on processing very large arrays of text information in a mode close to real time.

Conclusions. Text Mining technology is becoming one of the key tools of modern digital philology, providing effective processing and intelligent analysis of large linguistic data, i.e. its application allows to significantly increase the objectivity of research, automate the analysis of large-scale text corpora, reveal hidden patterns of language development and predict the transformation of communicative processes.

In conditions of growing uncertainty, turbulence and systemic risks, reflected by modern concepts of SHIVA and WARMI, the importance of intelligent text analysis technologies continues to grow. It is the optimal synergistic use of machine learning and artificial intelligence methods [10] that provides the possibility of operational research of dynamic changes in information flows, identification of new language trends and support for making scientifically sound decisions (including in the field of philological research).

The prospects for further development of Text Mining technology are associated with the improvement of semantic analysis methods, the expansion of the use of large language models, the development of multilingual corpora, the integration of knowledge graphs, and the creation of new generation intelligent analytical platforms = this will significantly expand the methodological apparatus of modern philology and ensure the transition to a comprehensive study of language processes in the context of the digital transformation of society.

References

1. I. Hrashchenko, S. Krasniuk, T. Tsalko, S. Nevmerzhytska, N. Liubymova (2024). *Effective computer modeling in educational management*. ScientificWorldJournal. – September 2024. – Issue 27, Part 3. – P. 3-17. – URL: <https://www.sworldjournal.com/index.php/swj/issue/view/swj27-03>
2. Naumenko, M. (2024). Analiz ta analityka velikykh danykh v marketynh i torhivli konkurentospromozhnoho pidpriemstva [Analysis and analytics of big data in marketing and trade of a competitive enterprise]. Grail of Science, # 40, pp. 117–128. DOI: <https://doi.org/10.36074/grail-of-science.07.06.2024.013> (in Ukrainian).

3. Krasnyuk M.T. (2006). Problemy zastosuvannya system upravlinnia korporatyvnymy znanniamy ta yikh taksonomiia [Problems of applying corporate knowledge management systems and their taxonomy]. Modeliuvannya ta informatsiini systemy v ekonomitsi - Modeling and information systems in the economy, vol. 73, p. 256 [in Ukrainian].

4. M. Krasnyuk, T. Motsyuk, S. Krasniuk (2021). Relevanz des wahlfaches "Wissensdetektion in unstrukturierten daten" bei der ausbildung von mastern der technischen und humanitären fachrichtungen. Education and science of today: intersectoral issues and development of sciences : collection of scientific papers "ΛΟΓΟΣ" with proceedings of the I International scientific and practical conference, Cambridge, United Kingdom, March 19, 2021. Vol. 3. Cambridge-Vinnytsia : P.C. Publishing House & European Scientific Platform, 2021. pp. 66–67.

5. Goncharenko S. (2025). Innovative IT technologies of computational linguistics. *Trends, Issues, and Challenges in Modern Science* : Proceedings of the 2nd International Scientific Conference (Cambridge, United Kingdom, 5 September 2025). - Lulu Press, Inc., 2025. - P. 109-113. URL: <https://test.knutd.edu.ua/handle/123456789/31072>

6. Goncharenko S. (2025). Intelligent information technologies for innovative management of advanced philology projects. *Innovations and New Directions in Scientific Research* : Proceedings of the 2nd International Scientific Conference (Manchester, United Kingdom, 20 September 2025). - Lulu Press, Inc., 2025. - P. 176-179. URL: <https://er.knutd.edu.ua/handle/123456789/31648>

7. Krasniuk S. (2024). Advanced text mining in philology. *Collection of scientific papers "SCIENTIA"* with Proceedings of the VI International Scientific and Theoretical Conference, Stockholm, Kingdom of Sweden, September 27, 2024. – Stockholm, Kingdom of Sweden : International Center of Scientific Research, 2024. – P. 68-72. – URL: <https://previous.scientia.report/index.php/archive/issue/view/27.09.2024>

8. Krasniuk S. (2024). Structured, semi-structured and unstructured batch and stream data in philology. *Science and Global Challenges in the Modern World*: proceedings of the International Scientific Conference, Lester, United Kingdom, November, 2024. – Lester, UK : IEDC ; Bookmundo, 2024.

9. Krasniuk, S., & Goncharenko, S. (2024). Big data in philology. *Collection of Scientific Papers "ΛΟΓΟΣ"*, (September 20, 2024; Paris, France), 159–165. <https://doi.org/10.36074/logos-20.09.2024.031>

10. Derbentsev, V. D., Serdiuk, O. A., Soloviov, V. M., & Sharapov, O. D. (2010). Synergistic and econophysical methods of studying dynamic and structural characteristics of economic systems. Cherkasy: Brama-Ukraine. - 2010 [in Ukrainian].