**European Journal of Research Development and Sustainability (EJRDS)**
Available Online at: https://www.scholarzest.com
Vol. 6 No 01 January 2025
ISSN: 2660-5570

# DEVELOPING NLP TOOL FOR LINGUISTIC ANALYSIS OF UZBEK LANGUAGES

**Ass. professor Marhamat Haydarova Yunusovna**
**Assistant Guzal Shikhnazarova Alisherovna**
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

**Article history:**
**Received:** 24th September 2024
**Accepted:** 24th October 2024

**Abstract:**
Automatic processing of unstructured texts in natural languages is one of the urgent problems of computer analysis and synthesis of texts. It is possible to separately highlight the task of text normalization, usually implying the implementation of such processes as tokenization , stemming and lemmatization . Existing stemming algorithms are mostly focused on synthetic languages, in which form formation using morphemes prevails. The Uzbek language is an example of an agglutinative language, characterized by polysemantic affixal and service morphemes. Although the Uzbek language has many differences, for example, from the English language, nevertheless, it can be successfully processed by stemming algorithms.

**Keywords:** NLP, computer linguistics, tokenization, simple words, compound words, paired words, repeated words.

## INTRODUCTION

Grammatical structure is a critical component of human communication, necessary to ensure the clarity and comprehensibility of written and spoken language [1]. Errors in grammatical construction can lead to misunderstandings and difficulties in comprehension by the audience. Automated correction of grammatical inaccuracies has become a key topic of many studies in the field of natural language processing (NLP). Recently, languages such as English and Chinese have attracted significant attention in research due to the emergence of extensive pre-trained models and significant datasets aimed at grammatical error correction (GEC), achieving performance levels comparable to human competence [1-2]. In contrast, languages such as Arabic, Russian, and Uzbek have not been as widely researched, which is mainly due to their classification as resource-poor languages with limited training data. Uzbek, which is the 54th most spoken language in the world and has over 44 million speakers [2-3], faces limitations in research, especially in the field of GEC, due to this data gap. Recent research in the field of automatic grammar correction (GEC) based on Transformer models has led to significant advances and insights. Another study combined the Transformer model with a Generative Adversarial Network for automatic English GEC. This method was validated on datasets and contributed to further improvements in GEC, highlighting the potential of combining Transformers with other deep learning methods. Recent work has highlighted the importance of language modeling in GEC, indicating that comparing the probabilities of proposed edits can lead to good performance. This approach leverages the language understanding capabilities of Transformer models to evaluate and select the most likely corrections [4].

The use of Transformers in GEC is not limited to English. For example, recent approaches in Korean and Indonesian GEC have adapted Transformers, demonstrating competitive and promising performance compared to traditional RNN-based encoder-decoder models. These achievements highlight the flexibility and efficiency of the model across languages [5,6]. Despite these achievements, challenges remain. One study found that a standard Transformer-based GEC model failed to realize grammatical generality even in simple settings with limited vocabulary and syntax. Text classification is a critical task in the field of natural language processing (NLP), where the goal is to classify a text document into predefined classes. This task is essential in many real-world applications, such as sentiment analysis, spam detection, and topic modeling. Given the vast amount of unstructured data generated daily, text classification provides a means to make sense of this data and derive meaningful insights. In recent years, deep learning models have been widely used in text classification, yielding excellent results. However, most research works on text classification have focused on high-resource languages such as English. There is a significant gap in text classification research for low-resource languages, and Uzbek is no exception. The main objective of this work is to contribute to the NLP research community by solving the text classification problem for Uzbek using a multi-label news categorization task as an example. We present a new Uzbek text classification dataset and evaluate the performance of different models on this dataset. The models range from traditional rule-based approaches such as word- and character-based support vector machine (SVM ) to more advanced deep learning models such as recurrent neural networks (RNN ) and convolutional neural networks (CNN). Uzbek is spoken by over 30 million people and is mainly used in Uzbekistan and neighboring Central Asian countries. It is a Turkic language that has been heavily influenced by both Russian and Arabic and Persian languages due to geographical and historical reasons. Since Uzbek is a low-resource language, limited research and resources are available for natural language processing (NLP) tasks in Uzbek, making the creation and use of NLP resources an important step towards advancing the digitalization of the Uzbek language. Despite this, the Uzbek language has a rich literary history and continues to be an important part of the cultural heritage of the Uzbek people. Its official alphabet is Latin, and its grammar is close to other languages of the Turkic family, which differs significantly from more commonly studied languages in NLP, such as English and Chinese. This presents a challenge for NLP tasks in Uzbek, as models trained in these languages may be ineffective in processing the nuances of Uzbek text. Developing NLP resources and models specifically for the Uzbek language can help advance research in this area and promote the use of technologies in Uzbek-speaking communities. 1. The rest of the paper is organized as follows: we provide an overview of text classification and highlight some recent NLP work in Uzbek in Section 2. This is followed by the methodology in Section 3, where we describe the data collection process and dataset creation. In the Experiments section, we describe the models used for evaluation. Moving on, Section 5 covers the experimental results, followed by Section 6, where we discuss the effects and their implications.

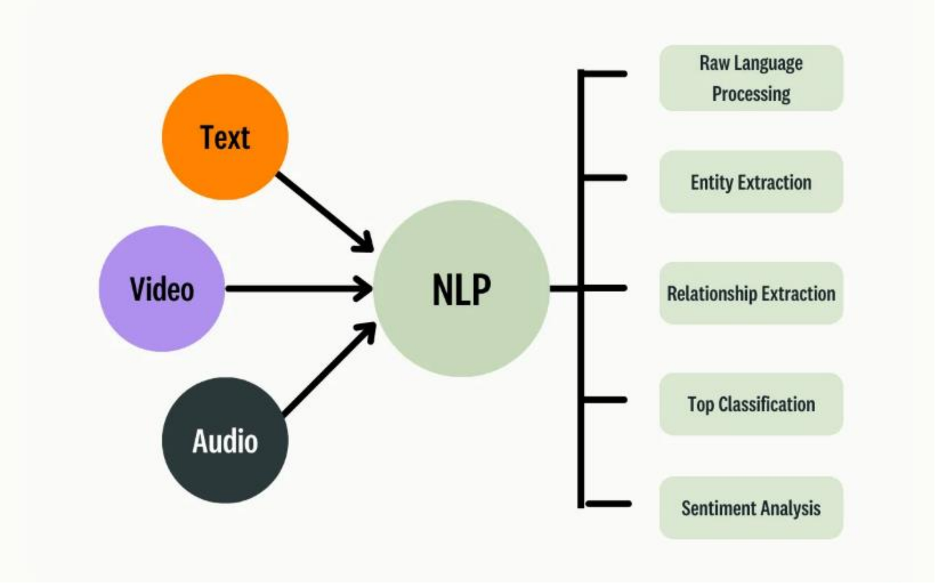


**Fig-1.** Natural language processing (NLP)

Natural language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages. Its goal is to make machines able to understand, interpret, and generate human language to communicate with humans. NLP allows machines to process large amounts of natural language data and extract relevant information from it to be used for various tasks such as sentiment analysis, text summarization, machine translation, question answering, and more.

## MATERIAL AND METHODS

*Definition and Importance.* In general terms, NLP is a sub-discipline of AI[7-8] which focuses on the ability of computers to analyze, understand, and generate natural language. It helps machines interact with humans in a way that they both can understand – by interpreting the words spoken by us or written in texts etc. In simpler terms, NLP enables machines to read our queries and respond accordingly in a meaningful way. The importance of NLP lies in its potential applications like providing virtual assistants for customer service inquiries or providing accurate translations from one language to another in real-time.

*Real-world applications.* NLP has seen several successful applications across many industries - from customer service automation to healthcare. For example, virtual assistants are powered by NLP techniques[9-10] like intent recognition, where the system identifies the user's query intent based on natural language input. Other examples include automatic summarization tools, which can summarize an article or email into a few concise sentences; automated grammar checks, which correct spelling mistakes; machine translation systems which translate text from one language into another; and voice search systems which allow users to search using voice commands instead of typing their query. NLP technology can also be used for automated sentiment analysis, where the system can detect emotions conveyed through words or phrases – this has potential applications in marketing research and customer decision-making processes.

*Components of NLP.* NLP consists[10] of several components, including speech recognition (recognizing spoken words):

*Syntax:* This component of natural language processing (NLP) deals with the arrangement of words in a sentence, ensuring grammatical correctness and structural integrity. Parsing techniques and rule-based algorithms are commonly employed to analyze and generate well-formed sentences.

*Semantics:* The field of semantics in NLP focuses on the underlying meaning of words and sentences. It goes beyond the surface-level understanding and delves into concepts, objects, and the relationships between them. NLP algorithms can infer the context and extract valuable insights by deciphering the intended message.

*Pragmatics:* Considered an essential component of language understanding, pragmatics considers the larger context in which communication occurs. It involves understanding the speaker's intention, the listener's interpretation, and the social norms that shape the interaction. Pragmatics is crucial in capturing nuances, sarcasm, and implying meaning beyond the literal.

*Morphology:* Morphology examines the structure and formation of words in a language. It analyzes root words, prefixes, and suffixes to understand how they contribute to the overall meaning of a word. By dissecting and understanding these linguistic units, NLP systems can effectively handle word variations and capture their intended essence.

*Phonetics:* In the realm of speech processing, phonetics is concerned with the study of sounds produced and perceived in language. It explores the physical properties of speech sounds, their acoustic characteristics, and the articulatory processes involved in their production. With a deeper understanding of phonetics, NLP algorithms can accurately transcribe speech and convert it into written text.

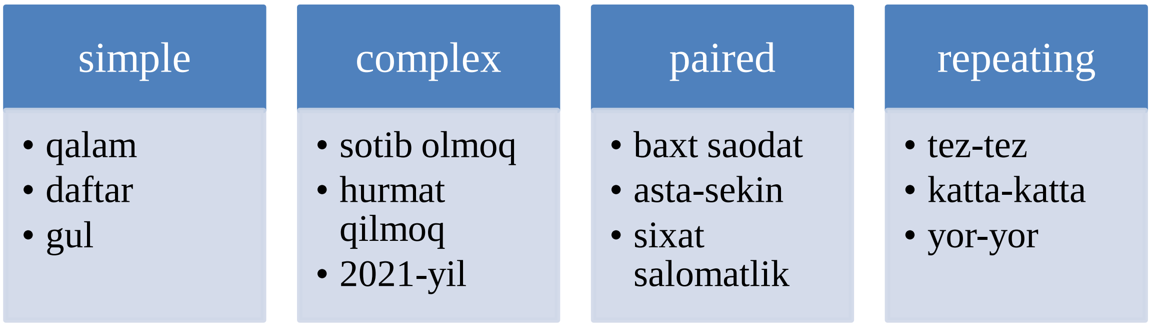
*Discourse:* This fundamental aspect of NLP involves unraveling how sentences connect to deliver coherence and coherence in a text. It considers language elements' overall structure, flow, and organization within a larger textual context. Discourse analysis enables NLP systems to build meaningful representations of written or spoken text by analyzing relationships between sentences.

*Statistical and Machine Learning Models:* Modern NLP heavily relies on statistical methods and machine learning models, encompassing advanced techniques such as deep learning. These models process large volumes of textual data, learning patterns, and relationships to understand and generate human-like language. By leveraging these powerful algorithms, NLP systems can achieve remarkable language understanding and generation levels[11].

These components work together to enable different types of tasks related to natural language processing, such as speech synthesis generation (generating audible speech) and sentiment analysis (understanding user feedback).

## RESULTS AND DISCUSSION

The morphological composition of words in the Uzbek language includes two main parts: root and affix morphemes. The root (o'zak ) is a morpheme that has a lexical meaning and does not have affixes in its composition. An affix is a morpheme that can have several functions. When attached to a root, an affix forms a new word or grammatical meaning. The normal form (base) of words in the Uzbek language ( negiz ) can consist only of a root morpheme or of a root and affix morphemes simultaneously[13]. The base of words can be derivative and non-derivative. A derivative base consists of a root and a word-forming affix. For example, " gul " (flower) is a non-derivative base, " guldon " (vase) is a derivative base. Word formation can occur in various ways, including by adding an affix to the root " gul+chi " (florist), by concatenating two simple words through a hyphen " baxt-saodat " (happiness), " tez-tez " (often). Thus, in the modern Uzbek language, words by their structure are: simple, complex, paired, repeating (Fig. 2).



| simple | complex | paired | repeating |
|---|---|---|---|
| • qalam | • sotib olmoq | • baxt saodat | • tez-tez |
| • daftar | • hurmat qilmoq | • asta-sekin | • katta-katta |
| • gul | • 2021-yil | • sixat salomatlik | • yor-yor |

Fig.-2. Types of words by structure

The developed algorithm varies depending on the types of words, which correspond to the following designations[12,14]:

*ws* – simple words (with a derivative or non-derivative base);
*wc* – compound words;
*wp* – paired words;
*wr* – repeating words.

Of greatest interest is the task of normalizing words like wc and wp , for example, respectively, "2021-yillar" and "sixat-salomatlik". The essence of the algorithm is that affixes occurring after the hyphen are removed from the word form. If the original word has no affixes, then it is considered a stemma . That is, "2021-yillar" is truncated to "2021-yil" (2021), "sixat-salomatlik" is truncated to "sixat-salomat" (healthy)

## CONCLUSION

Discrete text representation methods, each word in the corpus is considered unique and converted into a numeric form based on the various methods discussed above. The paper presents several advantages and disadvantages of the various methods. We summarize them in general. Methods that generate discrete numeric values of the text are easy to understand, implement, and interpret. The algorithms can be used to filter out simple and meaningless words. And this, in turn, helps in training and generalizing the model faster. The direct proportionality of the vocabulary to the corpus size can be pointed out as a disadvantage of the methods. Complex problems in NLP can be solved with the help of distributed text representation algorithms. Distributed text representations can be used to understand and learn a language corpus. An example of this is learning the words in a corpus and how they relate to each other. Today, distributed text representations are widely used in developing supervised learning models to solve complex NLP tasks in object recognition**.**

## REFERENCES

1. Kazakova MA, Sultanova AP Analysis of natural language processing technology: modern problems and approaches. *Advanced Engineering Research (Rostov-on-Don)*. 2022;22 (2):169-176. https://doi.org/10.23947/2687-1653-2022-22-2-169-176
2. Bakaev, I. I. (2021). Development of a stemming algorithm for words of the Uzbek language. *Cybernetics and programming* , (1), 1-12.
3. Aripov, M., Fayzullayeva, Z., & Karimov, N. (2024, November). Annotation of texts based on semantic analysis. In *AIP Conference Proceedings* (Vol. 3244, No. 1). AIP Publishing.
4. Abdurakhmonova, N., Alisher, I., & Sayfulleyeva, R. (2022, September). MorphUz: Morphological Analyzer for the Uzbek Language. In 2022 7th International Conference on Computer Science and Engineering (UBMK) (pp. 61-66). IEEE.
5. Elov, B. B., Khamroeva, S. M., Alayev, R. H., Khusainova, Z. Y., & Yodgorov, U. S. (2023). Methods of processing the uzbek language corpus texts. International Journal of Open Information Technologies, 11(12), 143-151.
6. Abdullayeva, Z. S., & Shihnazarova, G. A. (2023, April). INTELLEKTUAL TAHLIL ASOSIDA CHET TILI BILIMINI TEKSHIRISHNI MODELLASHTIRISH. In INTERNATIONAL SCIENTIFIC CONFERENCES WITH HIGHER EDUCATIONAL INSTITUTIONS (Vol. 1, No. 14.04, pp. 150-152).
7. Fayzullayeva, Z., Olimova, M., & Karimov, N. (2024). ABSTRAKTIV ANNOTATSIYALASH YORDAMIDA MATNLARNI ANNOTATSIYALOVCHI MASTAT TIZIMINI YARATISH. *DIGITAL TRANSFORMATION AND ARTIFICIAL INTELLIGENCE*, *2*(3), 17-23.
8. Yampolskiy R.V. AI-Complete, AI-Hard, or AI-Easy: Classification of Problems in Artificial Intelligence // Midwest Artificial Intelligence and Cognitive Science Conference. – Cincinnati, 2012. – P. 1-8. – URL: https://bit.ly/3vDqhI1.
   Fayzullayeva, Z., Karimov, N., & Abdusattarov, A. (2024). Matndan sun’iy intellekt yordamida Ekstrativ xulosa olish: Matndan sun’iy intellekt yordamida Ekstrativ xulosa olish. *MODERN PROBLEMS AND PROSPECTS OF APPLIED MATHEMATICS*, *1*(01).
9. Abdullayeva, Z. S., & Shihnazarova, G. A. (2023, April). INTELLEKTUAL TAHLIL ASOSIDA CHET TILI BILIMINI TEKSHIRISHNI MODELLASHTIRISH. In *INTERNATIONAL SCIENTIFIC CONFERENCES WITH HIGHER EDUCATIONAL INSTITUTIONS* (Vol. 1, No. 14.04, pp. 150-152).
10. Inatillayevna, F. Z., & Nosirjon o‘g‘li, K. N. (2024). NLPDA MATNLARNI UMUMIYLASHTIRISH VA LEKSIK TAHLIL. *«CONTEMPORARY TECHNOLOGIES OF COMPUTATIONAL LINGUISTICS»*, *2*(22.04), 338-343.
11. Каримова, В. А., & Шихназарова, Г. А. (2015). Инглиз тилини ўрганишда билимларни текшириш жараёнини моделлаштириш. *Современное образование (Узбекистан)*, (5), 19-23.
12. Mengliyev, I., Meylikulov, S., Fayzullayeva, Z., & Kobulova, M. (2024, November). Education artificial intelligence systems and their use in teaching. In *AIP Conference Proceedings* (Vol. 3244, No. 1). AIP Publishing.
13. Alisherovna, S. G. (2024). MATHEMATICAL MODELING OF MACHINE LEARNING ALGORITHMS. «CONTEMPORARY TECHNOLOGIES OF COMPUTATIONAL LINGUISTICS», 2(22.04), 418-420.
14. Abdurakhmonova, N. Z., Ismailov, A. S., & Mengliev, D. (2022, November). Developing NLP tool for linguistic analysis of Turkic languages. In 2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) (pp. 1790-1793). IEEE.