# AUTOMATIC DETECTION TECHNOLOGIES FOR STOPWORDS IN UZBEK LANGUAGE

**Aziz Iskandarov**
iskandarovazizbek7@gmail.com
**Shermatov Boburjon**
bobursherметov912@gmail.com
**Elmurod Kuriyozov**
elmurod1202@urdu.uz
Urgench State University

**Annotatsiya.** Nomuhim so‘zlar (stopwrods) tabiiy tilni qayta ishlashda (NLP) keng tarqalgan masala bo‘lib, matnni tahlil qilish, ma’lumotlarni qidirish va boshqa NLP ilovalarida muammolarni keltirib chiqarishi mumkin. Ushbu tadqiqot ishida o‘zbek tilida nomuhim so‘zlarni avtomatik aniqlash texnologiyalarini ishlab chiqishga e’tibor qaratilgan. Biz o‘zbek matnlarida nomuhim so‘zlarni aniqlash uchun mashinali o‘rganish (machine learning) algoritmlariga, xususan, Support Vector Machines (SVM) va Random Forest (RF) tasniflagichlariga asoslangan metodologiyani taqdim etamiz. Bizning metodologiya o‘zbek tilida o‘tkazilgan oldingi tadqiqotdan olingan “Maktab korpusi” deb nomlangan o‘zbek tilidagi matnlarning qo‘lda annotatsiya qilingan ma’lumotlar to‘plamidan foydalangan holda baholanadi va baholash natijalari yuqori aniqlik ko‘rsatkichlariga erishdi. Natijalarimiz shuni ko‘rsatadiki, taklif etilayotgan yondashuv o‘zbek matnlarida nomuhim so‘zlarni aniqlashning samarali usuli bo‘lib, o‘zbek tilidagi matnlarni qayta ishlash uchun NLP ilovalari ish faoliyatini yaxshilashda ishlatilishi mumkin.

**Kalit so‘zlar:** *O‘zbek tili, nomuhim so‘zlarni aniqlash, mashinali o‘qitish.*

**Abstract.** Stopwords are a common problem in natural language processing (NLP), which can cause issues in text analysis, information retrieval, and other NLP applications. This research paper focuses on the development of automatic detection technologies for stopwords in the low resource Uzbek language. We present an approach based on machine learning algorithms, specifically Support Vector Machines (SVM) and Random Forest (RF) classifiers, to identify stopwords in Uzbek texts. The approach is evaluated using a manually annotated dataset of Uzbek language texts called “School Corpus”, from a previously conducted research, and the evaluation results achieve high accuracy rates. Our results demonstrate that the proposed approach is an effective method for identifying stopwords in Uzbek texts and can be used to improve the performance of NLP applications for Uzbek language processing.

**Keywords:** *Uzbek language, Stopwords detection, Machine learning.*

**Аннотация.** Стоп-слова — распространенная проблема при обработке естественного языка (NLP), которая может вызывать проблемы при анализе

текста, поиске информации и других приложениях NLP. Это исследовательская работа посвящена разработке технологий автоматического обнаружения стоп-слов в узбекском языке. Мы представляем подход, основанный на алгоритмах машинного обучения, в частности на машинах опорных векторов (SVM) и классификаторах случайного леса (RF), для выявления стоп-слов в узбекских текстах. Подход оценивается с использованием вручную аннотированного набора данных текстов на узбекском языке под названием «Школьный корпус» из ранее проведенного исследования, и результаты оценки достигают высоких показателей точности. Наши результаты показывают, что предложенный подход является эффективным методом определения стоп-слов в узбекских текстах и может использоваться для повышения производительности приложений NLP для обработки узбекского языка.

**Ключевые слова:** *Узбекский язык, обнаружение стоп-слов, машинное обучение.*

## 1. Introduction.

Stopwords are commonly used words that are removed from text during NLP preprocessing, as they carry little meaning and can lead to noise in the analysis. However, the set of stopwords can vary depending on the language and domain of the text. In Uzbek language, there is a lack of comprehensive stopword lists available for NLP applications, which can affect the performance of text analysis and information retrieval tasks. Advancements in the field of Machine Learning [1] and Neural Networks [2], [3] make it possible to successfully identify stopwords from given texts.

In this paper, we propose a machine learning-based approach for automatic detection of stopwords in Uzbek language texts. Our approach utilizes Support Vector Machines (SVM) and Random Forest (RF) classifiers to identify stopwords. The proposed approach is evaluated using the already available annotated dataset of Uzbek texts called “School Corpus” [4], and we demonstrate its effectiveness in identifying stopwords in Uzbek texts.

**Uzbek language.** Uzbek is a Turkic language spoken by more than 30 million people mainly in Uzbekistan, but also in other Central Asian countries and by diaspora communities. It has a rich morphology and syntax, with complex inflectional and derivational systems, which can make language processing tasks challenging. The Uzbek language has its unique features and characteristics, such as vowel harmony and postpositions instead of prepositions. As there is a growing need for NLP applications to process Uzbek language data, developing effective techniques for tasks such as automatic detection of stopwords is crucial for improving the accuracy and effectiveness of these applications [5].

## 2. Related Work.

Several studies have been conducted on automatic stopword detection in different languages, including English, French, Chinese, and other languages [6], [7]. The most commonly used methods for stopword detection are rule-based and frequency-based approaches. However, these approaches require manual effort and do not always perform well, especially for languages with complex morphology and syntax.

Machine learning-based approaches have been shown to be effective in stopword detection tasks. In recent years, various machine learning algorithms, such as SVM, RF, and Naive Bayes (NB), have been applied to stopword detection tasks in different languages [8], [9].

Regarding the stopwords detection for Uzbek language, there has been a great deal of work, mostly focusing on the creating of the stopwords list for the language by Madatov et. al. [10], [11], which they introduce a methodology to extract Uzbek stopwords based on TF-IDF method. Apart from that, there has been a recent upward trend in creating NLP resources and tools for the Uzbek language, such as Latin-Cyrillic transliteration tool [12], datasets for sentiment analysis [13] and text classification [14], as well as tools for stemming [15], [16] and lemmatizers [17].

**3. Dataset.**

To perform our analysis on Uzbek stopwords, we used previously established ground rules and methodology for the Uzbek language stopwords by Madatov et. al. [18], who created the list of stopwords from their own dataset called “Uzbek Corpus”, which was constructed using online study materials comprising of 731,156 words, of which 47,165 are unique words.

We performed several preprocessing steps on the dataset, including the re-creation of annotated dataset by annotating the “School Corpus” they created using the above-mentioned unigrams. Further preprocessing steps took place, such as tokenization, stemming, and removing punctuation and numbers. We also converted all the text to lowercase to reduce the dimensionality of the feature space. The dataset was randomly split into training and testing sets with a ratio of 80:20.

**4. Methodology.**

4.1. Feature Extraction.

We used two feature extraction techniques to represent the text for classification: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). The BoW representation counts the occurrence of each word in the text, while TF-IDF measures the importance of a word in a document relative to the entire corpus.

**4.2. Classification.**

We used two classifiers, SVM and RF, to classify the texts into stopword and non-stopword categories, as the machine learning methods for classification prove themselves to perform good [8]. The SVM classifier was implemented with a linear kernel and a regularization parameter of 0.1. The RF classifier was implemented with 100 decision trees and a maximum depth of 10.

## 5. Evaluation and Results.

We evaluated the performance of our approach using several metrics, including accuracy, precision, recall, and F1-score. The results are presented in the following section.

Our experiments show that the SVM classifier achieved an accuracy of 98.2% and an F1-score of 0.98, while the RF classifier achieved an accuracy of 97.8% and an F1-score of 0.98. The precision and recall scores were also high, indicating that the classifiers performed well in identifying stopwords in Uzbek texts.

## 6. Conclusion and Future Work.

In this paper, we presented a machine learning-based approach for automatic detection of stopwords in Uzbek language texts. Our approach utilizes Support Vector Machines (SVM) and Random Forest (RF) classifiers to identify stopwords in Uzbek texts, and it was evaluated using a manually annotated dataset of 500 Uzbek texts. The results demonstrate that the proposed approach achieves high accuracy rates in identifying stopwords in Uzbek texts.

Our approach can be useful in various NLP applications for Uzbek language processing, such as text analysis, information retrieval, and text classification. By identifying and removing stopwords, the performance of these applications can be improved significantly.

Future work includes expanding the dataset and exploring other feature extraction techniques and machine learning algorithms to improve the performance of the approach. Additionally, the approach can be extended to other languages with similar characteristics to Uzbek, such as other Turkic languages. Overall, the proposed approach provides a useful tool for improving the accuracy and effectiveness of NLP applications for Uzbek language processing.

## References

1. J. Mattiev and B. Kavšek, “CMAC: clustering class association rules to form a compact and meaningful associative classifier,” in *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part I 6*, 2020, pp. 372–384.

2. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” *arXiv preprint arXiv:2301.12711*, 2023.

3. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodr\’\iguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243.

4. K. Madatov, S. Bekchanov, and J. Vičič, “Accuracy of the Uzbek stop words detection: a case study on" School corpus",” *arXiv preprint arXiv:2209.07053*, 2022.

5. U. Salaev, E. Kuriyozov, and C. Gómez-Rodr\’\iguez, “SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language,” in *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings*, 2022, pp. 199–206. [Online]. Available: www.scopus.com

6. J. Savoy, “A stemming procedure and stopword list for general French corpora,” *Journal of the American Society for Information Science*, vol. 50, no. 10, 1999, doi: 10.1002/(SICI)1097-4571(1999)50:10<944::AID-ASI9>3.0.CO;2-Q.

7. L. Dolamic and J. Savoy, “Brief communication: When stopword lists make the difference,” *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, 2010, doi: 10.1002/asi.21186.

8. J. Mattiev, U. Salaev, and B. Kavsek, “Word Game Modeling Using Character-Level N-Gram and Statistics,” *Mathematics*, vol. 11, no. 6, p. 1380, 2023.

9. J. Mattiev and B. Kavšek, “ACHC: Associative Classifier Based on Hierarchical Clustering,” in *Intelligent Data Engineering and Automated Learning–IDEAL 2021: 22nd International Conference, IDEAL 2021, Manchester, UK, November 25–27, 2021, Proceedings 22*, 2021, pp. 560–571.

10. K. Madatov, S. Bekchanov, and J. Vičič, “Lists of uzbek stopwords,” Univerza na Primorskem, Inštitut Andrej Marušič, 2021.

11. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.

12. U. Salaev, E. Kuriyozov, and C. Gómez-Rodr\’\iguez, “A machine transliteration tool between Uzbek alphabets,” in *CEUR Workshop Proceedings*, 2022, pp. 42–50. [Online]. Available: www.scopus.com

13. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com

14. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.

15. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a morphological and syntactic tagged corpus for the Uzbek language,” *arXiv preprint arXiv:2210.15234*, 2022.

16. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” *arXiv preprint arXiv:2210.16011*, 2022.

17. M. Sharipov and O. Sobirov, “Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language,” *arXiv preprint arXiv:2210.16006*, 2022.

18. X. Madatov, M. Sharipov, and S. Bekchanov, “O`ZBEK TILI MATNLARIDAGI NOMUHIM SO`ZLAR,” *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2021