MACHINE LEARNING-BASED HATE SPEECH DETECTION IN THE KAZAKH LANGUAGE

MILANA BOLATBEK
Al-Farabi Kazakh National University
bolatbek.milana@gmail.com

SHYNAR MUSSIRALIYEVA
Al-Farabi Kazakh National University
mussiraliyevash@gmail.com

MOLDIR SAGYNAY
Al-Farabi Kazakh National University
sagynaymoldir11@gmail.com

Abstract

Modern text data processing and classification methods require extensive use of machine learning and neural networks. Categorizing text into different classes has become a crucial task in many fields. This paper presents a multi-class text classification model utilizing a Modified TF-IDF (MTF-IDF) approach in combination with Long Short-Term Memory (LSTM) neural networks, XGBoost, and MLPClassifier algorithms. Additionally, the study explores the integration of TF-IDF and CountVectorizer (MTF-IDF) methods for text vectorization, aiming to enhance classification efficiency.

The research findings indicate that the LSTM model achieved the highest accuracy rate of 89%, demonstrating superior performance. The MLPClassifier model achieved 85% accuracy, while XGBoost obtained 81% accuracy. Moreover, the integration of TF-IDF and MTF-IDF methods significantly improved the detection of rare but essential words, enhancing the overall performance of the models.

This study is dedicated to addressing the problem of automated detection of harmful content in the Kazakh language. Hate speech in the digital space refers to any online material that harms individuals or communities through aggression, manipulation, discrimination, or the intentional spread of socially damaging narratives. The results provide a solid foundation for future research aimed at the early identification and mitigation of hate speech in the digital space, contributing to a safer online environment.

Keywords: hate speech, bullying, violent extremism, TF-IDF, MTF-IDF, LSTM, MLPClassifier.

1. Introduction

Nowadays, a large amount of information is exchanged on the Internet, including on social networks, the spread of dangerous content related to violence, racism, bullying and extremism is a pressing issue. Automatic detection and filtering of such content is an important task for companies working with big data, law enforcement agencies and social media platforms.

Monitoring content on the Internet should not only be ethical, but also meet legal requirements. In many countries of the world, social media platforms are required to operate within legislative restrictions, which increases their demand for automatic moderation. In addition, the spread of hate speech on the Internet is especially dangerous for young people, since their critical thinking skills are not yet fully developed. Such a situation can lead to the spread of various radical ideas and falling under the influence of dangerous groups. This issue is also relevant in Kazakhstan, since the level of Internet use among young people is very high. In this regard, automatic detection and mitigation and early detection of harmful digital contenton the Internet is one of the important tasks (Gorwa et al. 2020).

One of the most effective ways to combat destructive information is to use artificial intelligence and machine learning technologies. These technologies allow you to automatically analyze messages on the Internet and classify their content. In particular, it is important to develop algorithms and neural networks for detecting hate speech in the Kazakh language, as existing tools predominantly target high-resource languages and do not adequately address Kazakh-language online communication. This approach allows you to analyze texts semantically and divide them into categories based on their meaning. In addition, it is necessary to increase the information literacy of society by identifying false information on social networks, not sharing it, and raising awareness about cybersecurity (Barakhin et al. 2019).

In Kazakhstan, the fight against harmful content on the Internet is being carried out in several directions. The government has created special monitoring departments to monitor dangerous content on social networks. Technology companies and public organizations are also paying attention to this issue and taking various measures to prevent the spread of destructive information. Active efforts are being made to work together with the administrations of social networks, remove dangerous content, and prevent its spread (Kumisbekov et al. 2022).

In this article, authors consider the creation of a multi-category text classification model using methods such as TF-IDF, MTF-IDF, LSTM, XGBoost, MLPClassifier to solve the problem of automatic detection of dangerous content on the Internet. The TF-IDF and MTF-IDF methods are used for text vectorization, and the XGBoost and MLPClassifier algorithms are used for its classification. In addition, the LSTM neural network improves the classification quality by analyzing long-term dependencies. By combining these methods, we propose an effective hybrid model that automatically detects and early detects hate speech. This model can be widely used in detecting dangerous content on social networks, filtering spam, classifying news, and processing text data in general.

2. Literature Review

Recent studies have explored diverse machine learning and natural language processing approaches for text classification, cyberbullying detection, fake news identification, and keyword extraction in large-scale online environments.

For example, (Alqahtani and Ilyas 2024) addresses the problem of detecting and monitoring cyberbullying on social media platforms such as Facebook, Instagram, and X. The authors developed a system capable of automatically distinguishing aggressive and neutral posts using a stacking-based ensemble machine learning method. This approach integrated multiple feature extraction techniques and combined classifiers including Decision Trees, Random Forest, Linear SVM, Logistic Regression, and K-Nearest Neighbors. The ensemble effectively leveraged the strengths of each model, achieving an overall accuracy of 94%, which outperformed conventional machine learning baselines and demonstrated the potential of stacking classifiers to mitigate online aggression.

While text classification often relies on traditional term weighting schemes, (Li et al. 2011) critically evaluates the limitations of the classical TF-IDF model, particularly in contexts such as news keyword extraction, where frequently used but semantically significant words can be underweighted. To address this, the authors proposed MTF-IDF, a modified probabilistic approach that enhances the differentiation of important terms. Empirical evaluations showed that MTF-IDF significantly improved keyword identification and information retrieval performance compared to the conventional TF-IDF, highlighting its suitability for applications like news content analysis and search optimization.

Similarly, (Fan and Qin 2018) focused on improving text classification by introducing TF-IDCRF, an algorithm designed to correct inaccurate weighting of feature categories inherent in standard TF-IDF. By adjusting the IDF computation to consider inter-feature relationships and pairing the method with a Naïve Bayes classifier, the study demonstrated that TF-IDCRF yielded superior classification accuracy across multiple text corpora relative to other enhanced TF-IDF variants.

Beyond general text classification, other research has addressed the detection of misinformation. (Shakil and Alam 2022) presented an approach combining an improved TF-IDF algorithm with a Naïve Bayes classifier for fake news detection on the Internet. By incorporating linguistic structure and contextual features, the method achieved classification accuracy improvements from 85% to 93%, outperforming models that relied solely on TF-IDF or lacked probabilistic modeling. This demonstrates the value of integrating NLP-driven feature enhancement with probabilistic classification techniques for identifying disinformation.

Cyberbullying detection in settings with limited labeled data has also been investigated. (Schnitzler et al. 2016) proposed a session-based system leveraging an ensemble of one-class classifiers to process large volumes of unlabeled text data.

This approach requires only a small number of positive examples for initial training and demonstrated strong performance in detecting aggressive posts under conditions of data scarcity. Experimental comparisons indicated that the ensemble method outperformed both single-window and fixed-window strategies, establishing it as an effective framework for real-time cyberbullying detection when annotated data are limited or unavailable.

Collectively, these studies demonstrate that combining traditional and enhanced text representation methods (e.g., MTF-IDF, TF-IDCRF), ensemble learning, and probabilistic classifiers can yield substantial improvements in text classification accuracy across applications ranging from cyberbullying detection to fake news identification. They also underscore the importance of designing systems capable of operating effectively in dynamic, low-resource, or partially labeled environments typical of modern online platforms.

3. Methodology and methods

As part of the study, the texts were divided into five categories and a special dataset was created. These categories were divided into racism, bullying, violence, nationalism and neutral texts. Each category was assigned a corresponding numerical value: racism – 0, bullying – 1, violence – 2, nationalism – 3, neutral texts – 4. Thus, the texts were classified into specific categories and the basis for automatic detection of hate content was formed.

Machine learning and deep learning methods are widely used to detect destructive texts. These methods are implemented using text preprocessing, semantic analysis, and efficient classification algorithms. The research used TF-IDF, MTF-IDF, XGBoost, MLPClassifier, and LSTM neural networks. The features of each method and their role in detecting hate speech were considered.

TF-IDF is a classic method for determining the importance of words in a text. It calculates the frequency (TF) of each word and determines its importance (IDF) in the entire set of documents. This method allows you to convert the text into a digital format and helps to select important words when detecting hate speech.

MTF-IDF is an improved version of the TF-IDF method. In this method, the importance of each word in the text is modified depending on the class distribution. That is, words that are rare but frequently used in hate speech are given more importance. This approach allows you to more effectively detect hate speech.

XGBoost is a powerful machine learning algorithm based on gradient boosting. It is adapted to work with large amounts of text data and provides high accuracy of results. XGBoost helps to classify destructive texts more accurately by combining several tree models. MLPClassifier is a fully connected multilayer neural network. This model analyzes the meaning of each word or phrase, processing it as a vector. It is especially effective for processing large amounts of data and learning their complex patterns.

LSTM is a recurrent neural network that allows you to process texts while preserving temporal and semantic connections. This method helps to take into account the long-term dependencies of hate speech in the text structure. LSTM shows high efficiency in classifying disinformation and dangerous content, better understanding the context. In addition, an effective hybrid model was created by combining the results of TF-IDF and MTF-IDF with the LSTM neural network. This approach allowed for a deeper analysis of the meaning of texts and improved the accuracy of identifying hate speech.

This study investigated methods for automatic detection of hate speech using TF-IDF, MTF-IDF, XGBoost, MLPClassifier, and LSTM neural networks. This research utilized a dataset constructed in earlier studies (Bolatbek M. et al. 2024). Since the LSTM method can analyze the meaning of the text more deeply, it can be widely used in the future to better detect hate speech in the Kazakh language. Combining these machine learning methods will allow for early detection and mitigation of the spread of dangerous content on the Internet.

4. Results

Throughout the study, machine learning and deep learning methods were used to detect hate content and their effectiveness was evaluated. The results of each model were analyzed comparatively, and their advantages and disadvantages were identified. Texts were vectorized using the TF-IDF and MTF-IDF methods. These methods allowed us to select important words and calculate their numerical value. During text processing, the steps of removing stop words, converting the text to lowercase, and removing special characters were performed.

Among the machine learning methods, the XGBoost and MLPClassifier models were used. These models were combined with the TF-IDF and MTF-IDF methods, and their effectiveness was tested. As a result, the XGBoost model showed average accuracy, while the MLPClassifier showed good results (Figure 1, Figure 2).

Figure 1: MLPClassifier on confusion matrix

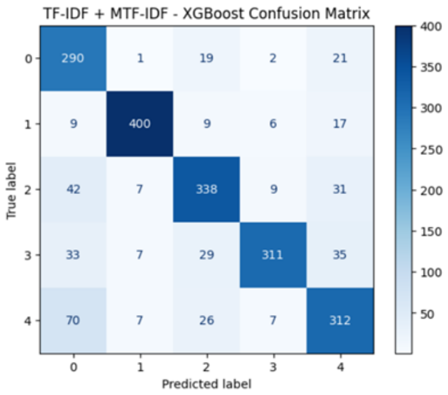
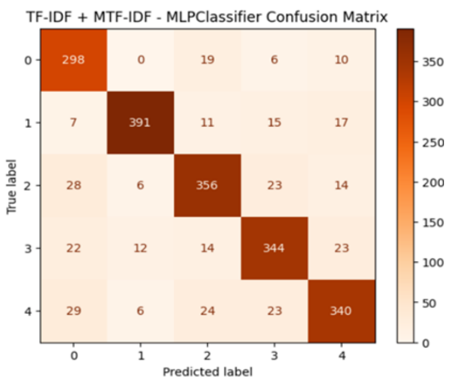


Figure 2: XGBoost on confusion matrix



Among the deep learning methods, the LSTM (Long Short-Term Memory) neural network was used. LSTM showed high efficiency in analyzing the semantic structure of texts and identifying long-term dependencies. This method showed excellent results in identifying hidden connections between texts and classifying hate speech. The LSTM neural network was combined with the TF-IDF and MTF-IDF methods, and the obtained indicators were compared with other models (Figure 3, Figure 4).

Figure 3: LSTMClassifier to classification report

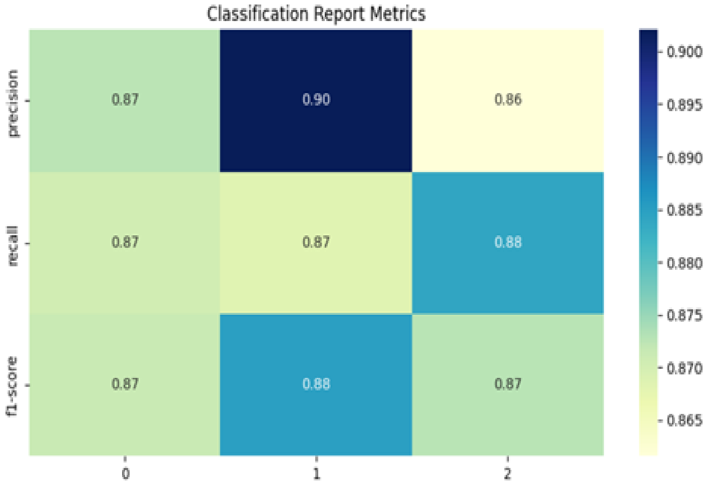
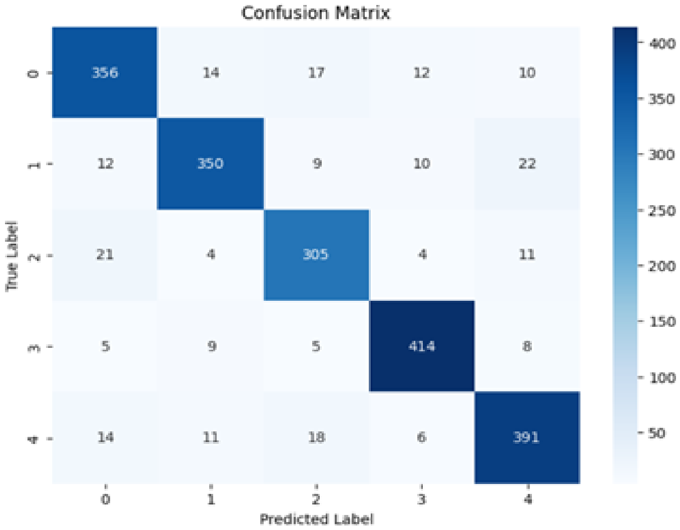


Figure 4: LSTMClassifier on confusion matrix



The results of the study showed that the LSTM neural network showed the highest performance, especially since it allows for deep analysis of the text structure, which provides more accurate results. Although the XGBoost and MLPClassifier models showed relatively good results, they were limited in analyzing contextual information in depth. The result obtained by combining the TF-IDF and MTF-IDF methods improved the performance of the XGBoost and MLPClassifier models. This approach helped to better identify rare but meaningful words, increasing the accuracy of the overall model (Table 1).

Table 1: Results by machine learning and deep learning methods

Model	Precision	Recall	F1-score
XGBoost (TF-IDF + MTF-IDF)	0.82	0.81	0.81
MLPClassifier (TF-IDF + MTF-IDF)	0.85	0.85	0.85
LSTM (TF-IDF + MTF-IDF)	0.89	0.88	0.89

In general, the results of the study showed that it is effective to combine different methods for automatic detection of destructive web content. Deep learning methods, especially LSTM, achieved high accuracy in detecting hate speech. However, in cases where resource saving and fast decision-making are required, XGBoost and MLPClassifier models can be used in combination with TF-IDF and MTF-IDF methods. This approach allows to detect hate speech at an early stage and prevent its spread.

5. Conclusion

This study used machine learning and deep learning methods to identify hate speech on the Internet. In order to classify Kazakh texts, text data was vectorized using TF-IDF and MTF-IDF methods, which were processed by XGBoost, MLPClassifier, and LSTM neural networks.

The results showed that the LSTM model achieved the highest accuracy, achieving 89%. This model showed high performance compared to other methods due to its ability to deeply analyze the semantic structure of the text. The MLPClassifier model showed a good result with an accuracy of 85%, while XGBoost showed a result of 81%, which indicated its lower efficiency compared to other models.

In the course of the study, the combination of TF-IDF and MTF-IDF methods improved the classification results and contributed to the increase in the performance of the models. This approach made it possible to more effectively identify rare but meaningful words. In addition, the use of neural networks helps to deeply understand the context of destructive texts and identify hidden connections between them.

In general, the results obtained show that the LSTM neural network is the most effective method for classifying hate speech in the Kazakh language. However, in order to save resources and increase processing speed, it is also reasonable to use the MLPClassifier and XGBoost models. The results of this study can serve as an important basis for future research aimed at automatically detecting hate speech on the Internet.

6. Acknowledgment

This research was carried out within the framework of the project funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant AP19576868).

References

Gorwa R., Binns R., Katzenbach C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance //Big Data & Society. – 2020. – Т. 7. – №. 1. – С. 2053951719897945.

Barakhin V. B. et al. Methods for detecting destructive information. // Physics Journal: Conference Series. – IOP Publishing, 2019. – Vol. 1405. – No. 1. – P. 012004.

Kumisbekov S. K., Sabitov S. M., Akimzhanova M. T. Issues of preventing cyberbullying at the present stage. // Bulletin of the Karaganda University "Law Series". – 2022. – Vol. 105. – No. 1. – Pp. 85–95.

Alqahtani A. F., Ilyas M. A Machine Learning Ensemble Model for the Detection of Cyberbullying //arXiv preprint arXiv:2402.12538. – 2024.

Li J. R., Mao Y. F., Yang K. Improvement and application of TF* IDF algorithm //Information Computing and Applications: Second International Conference, ICICA 2011, Qinhuangdao, China, October 28-31, 2011. Proceedings 2. – Springer Berlin Heidelberg, 2011. – С. 121-127.

Fan H., Qin Y. Research on text classification based on improved tf-idf algorithm //2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018). – Atlantis Press, 2018. – С. 501-506.

Shakil M. H., Alam M. G. R. Toxic Voice Classification Implementing CNN-LSTM & Employing Supervised Machine Learning Algorithms Through Explainable AI-SHAP //2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET). – IEEE, 2022. – С. 1-6.

Schnitzler K. et al. Using Twitter™ to influence research: discussing strategies, opportunities, and challenges. // International Journal of Nursing Studies. – 2016. – Vol. 59. – Pp. 15–26.

Bolatbek M. et al. Kazakh Language Dataset for Hate Speech Detection on Social Media Text //2024 IEEE 9th International Conference on Computational Intelligence and Applications (ICCIA). – IEEE, 2024. – С. 94-98.

Milana Bolatbek is a researcher specializing in Artificial Intelligence and Natural Language Processing. She holds PhD degree in Information Security Systems and focuses on the development of intelligent systems for text analysis, hate speech detection, and digital content monitoring. Her academic interests include deep learning, computational linguistics, and social media analytics. Milana Bolatbek has contributed to several interdisciplinary projects integrating linguistics, psychology, and AI for cybersecurity applications. She has co-authored papers published in peer-reviewed and Scopus-indexed journals and actively participates in international conferences on artificial intelligence and data science.

Shynar Mussiraliyeva is a researcher in the field of Cyber Security and Data Analytics. She is a professor of the department of Cybersecuirty and Cryptology at al-Farabi Kazakh National University and has extensive experience in machine learning, natural language processing, and intelligent information systems. Her research focuses on applying AI technologies to solve problems in cybersecurity, social media analysis, and digital communication. Shynar Mussiraliyeva has published numerous papers in international peer-reviewed and Scopus-indexed journals. She is actively involved in academic collaborations and has supervised several research projects related to AI applications in language and behavior analysis.

Moldir Sagynay is a researcher in the field of Artificial Intelligence and Computational Linguistics. She holds a Master's degree in Information Technology and focuses on developing intelligent algorithms for text processing, emotion detection, and online communication analysis. Her academic interests include neural network models, low-resource language processing, and digital safety systems.

(1) whether or not AI tools were used in writing the paper
Artificial intelligence tools (such as Grammarly and ChatGPT) were used only for language editing, text polishing, and formatting.

(2) existence of any conflict of interest (in no, information: no conflict of interest)
no conflict of interest

(3) financing of the research - if any (if not, information: no financing of the research involved)
This research was carried out within the framework of the project funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant AP19576868).

(4) each author's contribution to the research - in percentage (e.g., author 1. - 50 % - putting the thesis and writing Parts a, author 2. - ....)

Milana Bolatbek contributed approximately 30% to the research, including formulating the main concept, developing the methodology, and writing the core sections of the paper.

Shynar Mussiraliyeva contributed around 30%, focusing on theoretical analysis, proofreading of the final version of the paper.

Moldir Sagynay contributed about 40%, supporting data collection, data processing, model testing, literature review, and editing the manuscript.