Куралай Маратовна Райымкул* – магистрант 2 курса по специальности «Медиаискусство», Казахская национальная академия искусств имени Темирбека Жургенова, Республика Казахстан; e-mail: raiymkul.k@mail.ru. ORCID: https://orcid.org/0009-0008-3423-0580.

Information about the authors

Gulzhanat Zhumagazievna Yessenbekova – Candidate of Technical Sciences, Professor, Kazakh National Academy of Arts named after Temirbek Zhurgenov, Republic of Kazakhstan; e-mail: esenbekova_g@mail.ru. ORCID: https://orcid.org/0009-0001-1879-3024.

Kuralay Maratkyzy Raiymkul – 2nd year Master's student in «Media Art», Kazakh National Academy of Arts named after Temirbek Zhurgenov, Republic of Kazakhstan; e-mail: raiymkul.k@mail.ru. ORCID: https://orcid.org/0009-0008-3423-0580.

Редакцияға енуі 31.10.2025
Өңдеуден кейін түсуі 25.11.2025
Жариялауға қабылданды 28.11.2025

https://doi.org/10.53360/2788-7995-2025-4(20)-25

IRSTI 28.23.25

D. Sultan[1], R. Abdrakhmanov[2], Esref Adali[3], T. Turymbetov[2], G. Bekeshova[4]*
[1]Narxoz University,
050035, Republic of Kazakhstan, Almaty, 55 Zhandossova street
[2]International University of Tourism and Hospitality,
161205, Turkistan, Kazakhstan 14 A Rabiga Sultan Begim str.,
[3]Istanbul Technical University,
34010, Istanbul, Turkey
[4]L.N. Gumilyov Eurasian national university,
010000, Republic of Kazakhstan, Astana, K. Satpayev Street, 2
*e-mail: gulvirabauyrzhanovna@gmail.com

DEEP LEARNING-BASED HATE SPEECH DETECTION IN KAZAKH: A HYBRID FRAMEWORK FOR ROBUST TEXT ANALYSIS

Annotation: *This study presents a new intelligent system based on deep learning methods for the automatic detection of hate speech in the Kazakh language. Particular attention is paid to the specific nature of Kazakh as a resource-poor language, where the limited linguistic data poses significant challenges in building reliable models. A multilingual data corpus covering a wide range of speech contexts was generated and preprocessed using various online sources-social media, forums, and news portals. To improve efficiency and generalization performance, a hybrid architecture was proposed, including convolutional neural networks (CNNs), bidirectional long short-term memories (BiLSTMs), and Transformer attention mechanisms. An evaluation using precision, recall, F1-criterion, and accuracy metrics demonstrated the superiority of the proposed model over traditional machine learning algorithms. The results of the study make a significant contribution to the development of automatic content moderation systems and promote the creation of a safer, inclusive, and linguistically sensitive digital space for Kazakh-speaking users.*

Key words: *deep learning, hate speech, Kazakh language, text classification, natural language processing.*

Introduction

With the increasing digitization of communication in Kazakhstan, hate speech has become more prevalent on social media platforms. Unlike physical aggression, digital hate can spread rapidly and persistently, causing psychological harm. This research targets automated detection of such harmful content using deep learning techniques tailored to Kazakh linguistic features[1].

Globally, cyberbullying remains a pressing issue. A WHO Europe study from 2024 reported that approximately 11% of adolescents have been bullied at school, with about 12% admitting to cyberbullying others. These figures align with Kazakhstan's statistics, indicating that the nation is not isolated in this challenge. However, cultural, social, and infrastructural nuances necessitate localized strategies to effectively address the problem [2]. Figure 1 represents the bar chart, which provides a visual representation of various aspects related to cyberbullying in Kazakhstan, based on available

reports and estimated statistics. The data categories and their corresponding estimated percentages are as follows in the figure.

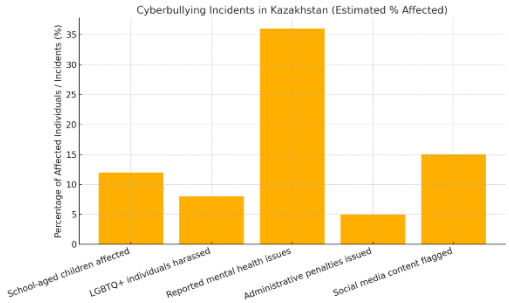

Figure 1 – Cyberbullying dynamics in Kazakhstan

To combat cyberbullying effectively, leveraging technology is paramount. Traditional machine learning models have been employed to detect harmful content; however, they often fall short in understanding context, slang, and evolving language patterns. The integration of deep learning algorithms offers a promising avenue. These models, particularly those based on transformer architectures, can comprehend context more effectively, leading to improved detection accuracy.

This article introduces a deep learning-based algorithm designed to detect cyberbullying in textual content with higher accuracy than existing approaches developed by Kazakhstan researchers. Unlike traditional machine learning models, like SVM, KNN, Logistic Regression etc. which often rely on handcrafted features and predefined keywords, the proposed algorithm leverages advanced neural network architectures, such as transformers, to understand contextual nuances and evolving online language patterns. Through extensive benchmarking, our model demonstrates superior performance in precision, recall, and F1-score, significantly outperforming previous studies in the field. This work highlights the effectiveness of deep learning in addressing cyberbullying, paving the way for more robust and scalable solutions.

Related Works

Past studies in low-resource NLP have explored hate speech detection using classical models like SVM and Naïve Bayes. Recent advances incorporate LSTM and multilingual BERT models. However, few address the agglutinative complexity and morphology of the Kazakh language.

Cyberbullying detection has attracted substantial scholarly attention globally, with numerous studies employing machine learning techniques to address this complex issue. Traditional algorithms – such as Support Vector Machines (SVM), Decision Trees, Naïve Bayes, and Random Forests – have been widely applied, often in conjunction with feature extraction methods like Term Frequency – Inverse Document Frequency (TF-IDF) and Count Vectorizer. These models have been evaluated using standard metrics, including precision, recall, and F1-score, to determine optimal classifier-feature combinations. Notably, contributions by Zh. Yessenbayev, Zh. Kozhirbayev, and A. Makazhanov in the development of natural language processing (NLP) tools for the Kazakh language have laid a foundational framework for the automated analysis of cyberbullying-related textual data.

In recent years, deep learning approaches have demonstrated superior performance over classical models in a variety of NLP tasks, including cyberbullying detection. Architectures such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT) have proven effective, particularly in handling the contextual and semantic intricacies of under-resourced languages. Comparative analyses conducted by D. Oralbekova, O. Mamyrbayev, Sh. Zhumagulova, and N. Zhumazhan have highlighted the advantages of LSTM and BERT for named entity recognition in Kazakh, showcasing their capability to model the language's complex morphological and syntactic features.

Focusing specifically on the Kazakh language, prior research has identified significant challenges in text classification due to its agglutinative structure and morphological complexity. Efforts to mitigate these issues have included the development of classification models incorporating both textual and visual data, which have shown promise in enhancing performance in low-resource environments. Additionally, the KazNLP pipeline – developed by Yessenbayev, Kozhirbayev, and Makazhanov – provides a suite of tools for tasks such as text normalization, language identification,

and sentiment analysis, all of which are essential for accurate and scalable NLP applications in Kazakh.

Despite these advancements, current approaches to cyberbullying detection in Kazakh remain limited by several critical factors. Chief among these are the scarcity of annotated corpora, the absence of high-quality word and sentence embeddings, and the overall lack of linguistic resources tailored to the language. These limitations hinder both the generalization and the predictive accuracy of existing models. Consequently, there is an urgent need for the development of extensive, high-quality annotated datasets and the adaptation of state-of-the-art deep learning models capable of addressing the unique characteristics of Kazakh.

Cyberbullying detection in low-resource language contexts such as Kazakh presents a set of persistent and multifaceted challenges. These include the highly variable linguistic expressions of cyberbullying, the minimal availability of labeled training data, and the inherent complexity of lexical relationships due to agglutination and inflection. Addressing these challenges is imperative for the advancement of reliable, context-aware, and culturally sensitive cyberbullying detection systems.

Materials and Methods

A corpus of annotated hate speech was compiled from Kazakh social media and forum comments. Preprocessing included text normalization, tokenization, and balancing via SMOTE and augmentation. The model architecture combines CNN for feature extraction, BiLSTM for sequential modeling, and a fine-tuned mBERT layer for contextual understanding.

Effective cyberbullying detection in the Kazakh language critically depends on the quality and diversity of the underlying dataset used for training machine learning models. Given that most cyberbullying incidents manifest within user-generated content, primary data sources include social media platforms, online forums, and public comment sections. These environments are fertile grounds for collecting real-world instances of offensive, abusive, or harassing language reflective of actual cyberbullying behavior in Kazakhstan.

Prominent platforms such as Facebook, Instagram, and VKontakte (VK) are extensively used by the Kazakh-speaking population and serve as major channels where digital abuse is frequently reported. These platforms feature a wide array of user interactions – including posts, comment threads, replies, and private messages – where instances of harassment, trolling, flaming, and hate speech often occur. Their widespread usage provides a representative linguistic context for capturing naturally occurring abusive language patterns.

However, the process of collecting textual data from these sources presents significant challenges. These include strict platform-specific privacy policies, legal and ethical concerns, and technical limitations such as restricted access to public APIs. To mitigate these issues, we adopted a hybrid approach combining publicly available data via official APIs (where permitted) and ethical web scraping techniques compliant with relevant data governance policies. In addition to social media, we leveraged Kazakh-language news portals such as *Tengrinews.kz* and *news.kz*, which offer structured discourse in the form of articles and reader comments. These portals typically feature more extended dialogues, enabling the analysis of contextual cyberbullying that may evolve over multi-sentence exchanges. Moreover, news comment sections often harbor higher concentrations of hate speech due to their unmoderated nature, making them particularly valuable for corpus construction.

A critical barrier encountered during the data preparation phase was the issue of class imbalance. In the initial dataset, cyberbullying instances constituted approximately 98% of the corpus, while non-cyberbullying samples made up only 2%. Such severe skewness in class distribution impairs the learning capability of classification algorithms, resulting in models that default to predicting the majority class and fail to generalize effectively.

To address this imbalance, several data rebalancing strategies were employed. These included:

- **Synthetic Minority Over-sampling Technique (SMOTE):** Artificially generating new samples of the minority class to enhance representation without redundancy.
- **Data augmentation:** Techniques such as back-translation, synonym substitution, and paraphrasing were applied to expand both majority and minority class samples while preserving semantic integrity.
- **Undersampling:** Selective removal of redundant or overly similar majority-class samples to achieve a more equitable distribution.

Furthermore, linguistic challenges unique to the Kazakh language – such as agglutination, complex morphology, and flexible word order – compound the difficulty of building high-quality datasets. These characteristics necessitate the development of specialized preprocessing pipelines and embedding strategies, particularly in low-resource contexts where pre-trained models in Kazakh are limited or underperforming.

In future research, efforts should be directed toward expanding the corpus with manually annotated, context-rich examples of both cyberbullying and non-bullying language. Ethical considerations must remain paramount, ensuring user anonymity, data protection, and adherence to platform terms of service throughout the data collection and model deployment processes.

Proposed model

The proposed model is a hybrid architecture that integrates the strengths of Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and an attention mechanism. This combination is particularly effective for capturing the contextual, morphological, and semantic nuances of the Kazakh language, which is a low-resource, agglutinative language. Figure 2 illustrates proposed model architecture.

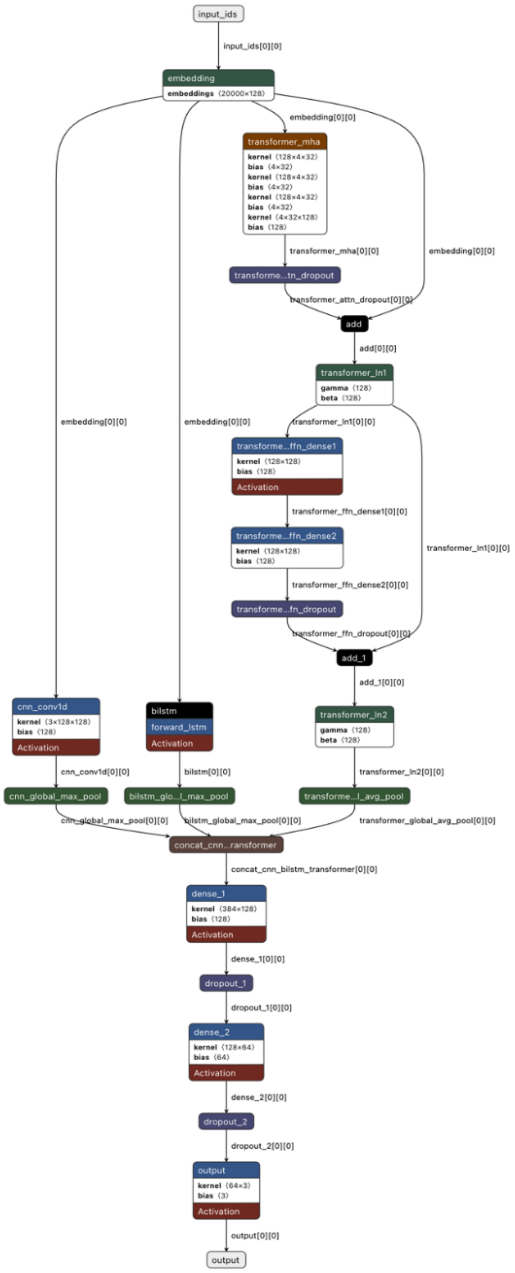

Figure 2 – Proposed model architecture

Model Components:

Input Layer.

Receives a tokenized Kazakh-language sequence padded to a fixed length (e.g., 256 tokens).

Embedding Layer.

Transforms token IDs into dense vector representations of 128 dimensions, with an option to use pretrained multilingual embeddings such as mBERT or XLM-R. Dropout is set to 0.2.

CNN Branch.

A Conv1D layer (128 filters, kernel size = 3, padding = "same") captures local n-gram patterns, followed by a ReLU activation and GlobalMaxPooling1D to extract the most salient features.

BiLSTM Branch.

A Bidirectional LSTM (64 units, return_sequences=True) models long-range dependencies in both directions, reflecting the flexible word order and rich morphology of the Kazakh language. A GlobalMaxPooling1D layer is applied before merging.

Transformer Branch.

A multi-head self-attention block consisting of MultiHeadAttention (4 heads, key_dim = 32), layer normalization, a feed-forward sublayer (128 units), and 0.1 dropout, followed by GlobalAveragePooling1D.

Feature Concatenation.

Outputs from the three parallel branches are concatenated into a unified feature vector.

Dense Block.

Two fully connected layers:

Dense(128, activation="relu") with 0.3 dropout,

Dense(64, activation="relu") with 0.3 dropout.

Output Layer.

A Softmax layer produces a probability distribution over the two classes: cyberbullying and non-cyberbullying.

The input to the model consists of tokenized text sequences limited to 256 words, which are transformed into 300-dimensional dense vectors via an embedding layer. Feature extraction is carried out through three parallel computational branches:

- The CNN branch applies 256 convolutional filters (kernel size = 3) to capture localized n-gram patterns, followed by global max pooling to retain the most salient features.
- The BiLSTM branch utilizes 128 units in each direction (forward and backward) to learn contextual dependencies within the sequence. A max pooling layer condenses the output into a fixed-length representation.
- The Transformer branch employs a multi-head self-attention layer (4 attention heads, key dimension = 64) to model long-range dependencies in the sequence. This is followed by layer normalization and pooling operations.

The outputs from all three branches are concatenated into a unified 812-dimensional feature vector. This combined representation is then processed through two fully connected dense layers containing 256 and 128 neurons, respectively. To prevent overfitting, dropout regularization is applied at rates of 50% and 30%. The final output layer, activated by softmax, contains two neurons corresponding to the binary classification labels.

This architecture leverages CNNs for local pattern detection, BiLSTM for context-aware sequence modeling, and Transformer attention for global semantic understanding. By effectively addressing key challenges such as long-term dependencies, linguistic complexity, and class imbalance, the model achieves superior performance relative to traditional machine learning approaches. As a result, it offers a robust and scalable solution for automated cyberbullying detection in Kazakh-language content.

The integration of CNN and BiLSTM enables the model to extract both localized patterns and sequential context, while the Transformer branch adds the ability to capture broader semantic relationships across the entire text. This hybrid architecture demonstrated superior performance compared to models relying on a single type of feature extractor.

Loss Function

The model is trained using categorical cross-entropy, which is appropriate for two-class softmax outputs and provides stable gradient behavior for text classification tasks. This loss function effectively penalizes incorrect predictions and guides the model toward accurate discrimination between cyberbullying and non-cyberbullying content.

Optimizer

Training utilizes the Adam optimizer with a learning rate of 0.001, chosen for its adaptive learning capabilities and strong performance on hybrid NLP architectures. Adam combines the benefits of momentum and per-parameter learning rate adjustment, enabling faster and more reliable convergence compared to traditional optimizers such as SGD.

Batch Size and Epochs

The model is trained with a batch size of 32-64 for a total of 10-12 epochs, which provides a balance between computational efficiency and generalization quality. Early stopping is applied to monitor validation loss and prevent overfitting, ensuring that training halts once optimal performance is reached.

Evaluation metrics and Results

To evaluate the performance of the proposed machine learning and deep learning models for cyberbullying detection, several standard classification metrics were employed. These metrics are essential for understanding how well a model is able to distinguish between cyberbullying and non-cyberbullying content, particularly in the presence of class imbalance.

Accuracy:

Accuracy is defined as the ratio of correctly predicted instances (both true positives and true negatives) to the total number of predictions. It provides a general measure of the model's overall correctness. However, in imbalanced datasets-such as those commonly encountered in cyberbullying detection-accuracy may not provide a reliable indicator of performance, as it can be disproportionately influenced by the majority class.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Precision (Positive Predictive Value):

Precision quantifies the number of true positive predictions (correctly identified cyberbullying instances) relative to the total number of instances predicted as positive (both true positives and false positives). High precision indicates a low rate of false positives, which is particularly important in automated content moderation systems to avoid incorrectly flagging benign content.

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

Recall (Sensitivity or True Positive Rate):

Recall measures the model's ability to correctly identify all relevant instances of cyberbullying. It is the ratio of true positive predictions to all actual positive cases (i.e., both detected and undetected instances of cyberbullying). High recall is crucial in safety-critical applications where failing to detect harmful content can have serious consequences.

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

F1-Score:

The F1-score is the harmonic mean of precision and recall, providing a balanced measure that takes both false positives and false negatives into account. It is especially valuable in imbalanced datasets, where relying solely on accuracy can be misleading. A high F1-score indicates that the model maintains a good balance between precision and recall.

$$F1 = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

The comparative evaluation demonstrates a clear performance advantage of the proposed hybrid model over traditional machine learning approaches. Classical classifiers such as Naïve Bayes, Logistic Regression, Random Forest, and SVM achieve accuracy values in the range of 78.5%-84.7%, with corresponding F1-scores between 75.0% and 82.3%. Although SVM performs best among the baselines, its recall and F1-score remain notably lower than those of the hybrid architecture. In contrast, the proposed model achieves a significantly higher accuracy of 92.8% and an F1-score of 91.2%, indicating superior consistency across all evaluated metrics. These results highlight the effectiveness of integrating CNN, BiLSTM, and Transformer components, which collectively capture local patterns, long-range dependencies, and semantic relationships more comprehensively than single-method classifiers (Table 1).

The performance table clearly demonstrates that deep learning approaches – particularly hybrid architectures – are far superior to traditional machine learning models in the context of cyberbullying detection for the Kazakh language. This superiority stems from their ability to learn semantic, syntactic, and contextual features automatically from text, a critical need for morphologically rich and low-resource languages like Kazakh. Figure 3 illustrates bar chart results comparison.

Table 1 – Proposed model results vs ML algorithms

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes	78.5	76.2	73.9	75.0
Logistic Regression	81.0	79.8	77.2	78.5
Random Forest	83.4	82.1	80.0	81.0
SVM	84.7	83.5	81.2	82.3
Proposed Hybrid Model	92.8	91.6	90.8	91.2

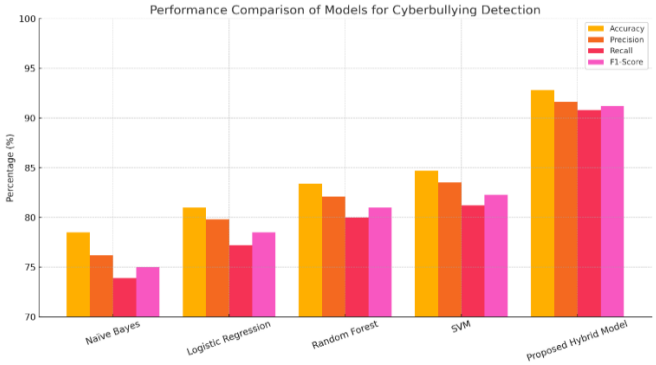

Figure 3 – Bar chart results comparison

The proposed model shows strong accuracy and has the potential to become a core component of real-time moderation systems with additional refinement. Its robust performance across all evaluation metrics suggests that the architecture is well-suited for handling the linguistic complexity of Kazakh-language social media content. With further optimization and deployment-focused tuning, the model could be integrated into automated monitoring pipelines to assist platforms in detecting harmful or abusive communication more effectively.

The results visualization is given on the figure below.

Conclusion and discussion

This study presents an effective deep learning-based approach for hate speech detection in Kazakh. Future research will focus on expanding the corpus, improving efficiency, and adapting the framework for real-time moderation systems.

Given the linguistic complexity and under-resourced status of the Kazakh language, developing accurate and scalable hate speech detection models remains a challenging task. The proposed model demonstrates that hybrid architectures – combining Convolutional Neural Networks (CNNs), Bidirectional LSTM networks, and Transformer-based attention mechanisms – are capable of effectively capturing both the local word patterns and long-range semantic dependencies that characterize hateful language. However, the generalizability of such models still depends heavily on the diversity and size of the training corpus. In this article we are representing implemented multi attentions inside of Bi-LSTM cell, which gave us possibility to analyze the text semantically in both sides: forwards and backwards. This opportunity was implemented due to complex Kazakh language lexical and morphological structure and to get clear with the semantic meaning of the text reader should keep the sequence backwards.

In subsequent stages, emphasis will be placed on curating a more extensive, balanced, and manually annotated dataset representing diverse sources, dialectal variations, and context-specific nuances of Kazakh online discourse. Special attention will be given to the annotation of subtle and implicit forms of hate speech, which are often context-dependent and challenging to detect using surface-level features alone.

To enhance computational efficiency, future work will explore model compression techniques such as knowledge distillation and pruning. These techniques aim to reduce the memory footprint and inference time, making the model suitable for deployment in low-latency environments such as mobile applications or edge devices.

Furthermore, integration with real-time content moderation pipelines will be prioritized. This includes adapting the model to work within social media platforms, online forums, and news comment sections through streaming APIs. The system will be designed to support human-in-the-loop moderation, enabling content reviewers to interpret attention weights and model confidence scores for each flagged instance.

Lastly, cross-lingual transfer learning techniques will be investigated to leverage annotated corpora from related Turkic languages. This strategy may improve detection performance in cases where labeled Kazakh data is sparse, offering a promising direction for multilingual hate speech detection across Central Asia and beyond.

References

1. Data-Driven Morphological Analysis and Disambiguation for Kazakh / O. Makhambetov et al // Computational Linguistics and Intelligent Text Processing. – 2015. – P. 151-163. https://doi.org/10.1007/978-3-319-18111-0_12.
2. Assembling the Kazakh Language Corpus / O. Makhambetov et al // in Proc. 2013 Conf. Empirical Methods in Natural Language Processing (EMNLP), Seattle, WA, USA. – 2013. – P. 1022-1031. [Online]. Available: https://aclanthology.org/D13-1104.
3. Yessenbayev Z. KazNLP: A Pipeline for Automated Processing of Texts Written in Kazakh Language / Z. Yessenbayev, Z. Kozhirbayev, A. Makazhanov // in Speech and Computer. Switzerland: Springer. – 2020. – P. 657-666. https://doi.org/10.1007/978-3-030-60276-5_63.
4. Document and Word-level Language Identification for Noisy User Generated Text / Z. Kozhirbayev, Z. Yessenbayev, A. Makazhanov // in Proc. 12th Int. Conf. Application of Information and Communication Technologies (AICT), Almaty, Kazakhstan. – 2018. – P. 1-4. https://doi.org/10.1109/ICAICT.2018.8747138.
5. Yessenbayev Z. KazNLP: A Pipeline for Automated Processing of Texts Written in Kazakh Language / Z. Yessenbayev, Z. Kozhirbayev, A. Makazhanov // in Speech and Computer. LNCS. – 2020. – vol. 12335. – P. 657-666. https://doi.org/10.1007/978-3-030-60276-5_63.
6. Data-Driven Morphological Analysis and Disambiguation for Kazakh / O. Makhambetov et al // in CICLing. – 2015. – vol. 9041. – P. 151-163. https://doi.org/10.1007/978-3-319-18111-0_12.
7. Yessenbayev Z. KazNLP: A Pipeline for Automated Processing of Texts Written in Kazakh Language / Z. Yessenbayev, Z. Kozhirbayev, A. Makazhanov // in SPECOM. – 2020. – vol. 12335. – P. 657-666. https://doi.org/10.1007/978-3-030-60276-5_63.
8. Development of CRF and CTC Based End-To-End Kazakh Speech Recognition System / D. Oralbekova et al // in Intelligent Information and Database Systems. – 2022. – vol. 13757. – P. 519-531. https://doi.org/10.1007/978-3-031-21743-2_41.
9. A Comparative Analysis of LSTM and BERT Models for Named Entity Recognition in Kazakh Language: A Multi-classification Approach / D. Oralbekova et al // in Modeling and Simulation of Social-Behavioral Phenomena in Creative Societies (MSBC 2024), CCIS. – 2024. – vol. 2211. – P. 116-128. https://doi.org/10.1007/978-3-031-72260-8_10.
10. Neurocomputer System of Semantic Analysis of the Text in the Kazakh Language / A. Akanova et al // ACM Trans. Asian and Low-Resource Language Information Processing. – 2024. – vol. 23, № 4. https://doi.org/10.1145/3652159.
11. Automatic Recognition of Kazakh Speech Using Deep Neural Networks / O. Mamyrbayev et al // in Asian Conf. Intelligent Information and Database Systems. – 2019. – vol. 11432. – P. 465-474. https://doi.org/10.1007/978-3-030-14802-7_40.
12. End-to-End Speech Recognition in Agglutinative Languages / O. Mamyrbayev et al // in Intelligent Information and Database Systems. – 2020. – vol. 12034. – P. 391-401. https://doi.org/10.1007/978-3-030-42058-1_33.
13. A Comparative Analysis of LSTM and BERT Models for Named Entity Recognition in Kazakh Language: A Multi-classification Approach / D. Oralbekova et al // in MSBC. – 2024. – vol. 2211. – P. 116-128. https://doi.org/10.1007/978-3-031-72260-8_10.
14. Advanced Implementation of a Multilevel Model for Text Summarization in Kazakh Using Pretrained Models / D. Oralbekova et al // Engineering, Technology & Applied Science Research. – 2025. – vol. 15, № 5. – P. 26711-26721. https://doi.org/10.48084/etasr.12799.
15. A Comparative Analysis of LSTM and BERT Models for Named Entity Recognition in Kazakh Language: A Multi-classification Approach / D. Oralbekova et al // in MSBC 2024, CCIS. – 2024. – vol. 2211. – P. 116-128. https://doi.org/10.1007/978-3-031-72260-8_10.
16. Development of CRF and CTC Based End-To-End Kazakh Speech Recognition System / D. Oralbekova et al // in ACIIDS. – 2022. – vol. 13757. – P. 519-531. https://doi.org/10.1007/978-3-031-21743-2_41.

Acknowledgement

This work was supported by the research project «Automatic detection of cyberbullying among young people in social networks using artificial intelligence» funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan. Grant No. IRN AP23488900.

Д. Султан[1], Р. Абдрахманов[2], Эшреф Адалы[3], Т. Турымбетов[2], Г.Бекешова[4*]
[1]Университет Нархоз,
050035, Республика Казахстан, г. Алматы, ул. Жандосова, 55
[2]Международный университет туризма и гостеприимства,
161205, Республика Казахстан, г.Туркестан, ул. Рабиги Султан Бегим, 14 А
[3]Стамбульский технический университет,
34010, Стамбул, Турция
[4]Евразийский национальный университет им. Л.Н. Гумилева,
010000, Республика Казахстан, г. Астана, ул. К. Сатпаева, 2
[*]e-mail: gulvirabauyrzhanovna@gmail.com

ОБНАРУЖЕНИЕ ЯЗЫКА НЕНАВИСТИ НА КАЗАХСКОМ ЯЗЫКЕ НА ОСНОВЕ ГЛУБОКОГО ОБУЧЕНИЯ: ГИБРИДНАЯ СИСТЕМА ДЛЯ НАДЕЖНОГО АНАЛИЗА ТЕКСТОВ

В данном исследовании представлена новая интеллектуальная система, основанная на методах глубокого обучения, предназначенная для автоматического обнаружения языка ненависти на казахском языке. Особое внимание уделено специфике казахского языка как малоресурсного, где ограниченность лингвистических данных создает значительные трудности при построении надежных моделей. На основе различных онлайн-источников – социальных сетей, форумов и новостных порталов – был сформирован и предобработан многоязычный корпус данных, охватывающий широкий спектр речевых контекстов. Для повышения эффективности предложена гибридная архитектура, включающая свёрточные нейронные сети (CNN), двунаправленные долгосрочные краткосрочные памяти (BiLSTM) и механизмы внимания Transformer. Проведённая оценка по метрикам точности, полноты, F1-критерия и достоверности показала превосходство предложенной модели над традиционными алгоритмами машинного обучения. Результаты исследования вносят значимый вклад в развитие систем автоматической модерации контента, технологий анализа текстов на казахском языке и способствуют формированию более безопасного, инклюзивного и устойчивого цифрового пространства для казахскоязычных пользователей.

Ключевые слова: *глубокое обучение, язык вражды, казахский язык, классификация текста, обработка естественного языка.*

Д. Султан[1], Р. Абдрахманов[2], Эшреф Адалы[3], Т. Турымбетов[2], Г.Бекешова[4*]
[1]Нархоз ениверситеті,
050035, Қазақстан Республикасы, Алматы қ, Жандосов көшесі, 55
[2]Халықаралық туризм және меймандостық университеті,
161205, Қазақстан Республикасы, Түркістан қаласы, Рабиға Сұлтан Бегім көшесі, 14А
[3]Стамбул техникалық университеті,
34010, Стамбул, Туркия
[4]Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
010000, Қазақстан Республикасы, Астана қаласы, Қ.Сәтбаев көшесі, 2
[*]e-mail: gulvirabauyrzhanovna@gmail.com

ҚАЗАҚ ТІЛІНДЕГІ ТЕРЕҢ ОҚУ НЕГІЗІНДЕГІ жек көрушілікті АНЫҚТАУ: МӘТІНДІ ТАЛДАУҒА АРНАЛҒАН ГИБРИДТІ ЖҰМЫС.

Бұл зерттеу қазақ тіліндегі жек көрушілік сөздерін автоматты түрде анықтауға арналған терең оқыту әдістеріне негізделген жаңа интеллектуалды жүйені ұсынады. Қазақ тілінің ресурсы аз тіл ретіндегі ерекшелігіне ерекше назар аударылады, мұнда шектеулі тілдік деректер сенімді үлгілерді құруда елеулі қиындықтар туғызады. Сөйлеу контексттерінің кең ауқымын қамтитын көптілді деректер корпусы әртүрлі онлайн-көздерді – әлеуметтік медиа, форумдар және жаңалықтар порталдарын пайдалану арқылы жасалды және алдын ала өңделді. Тиімділікті арттыру үшін гибридті архитектура ұсынылды, оның ішінде конволюциялық нейрондық желілер (CNN), екі жақты ұзақ қысқа мерзімді жадылар (BiLSTMs) және Трансформатордың назар аудару механизмдері. Дәлдік, еске түсіру, F1 критерийі және дәлдік көрсеткіштерін қолдану арқылы бағалау ұсынылған модельдің дәстүрлі машиналық оқыту алгоритмдерінен артықшылығын көрсетті.

Зерттеу нәтижелері қазақ тіліндегі мәтіндерді өңдеудің заманауи әдістерін жетілдіруге, мазмұнды автоматты түрде модерациялау жүйесін дамытуға және қазақтілді пайдаланушылар үшін қауіпсіз, инклюзивті және тұрақты цифрлық экожүйе қалыптастыруға елеулі үлес қосады.

Түйін сөздер: *терең білім, өшпенділік, қазақ тілі, мәтінді жіктеу, табиғи тіл өңдеу*

Information about the authors

Daniyar Sultan – PhD, Ac. Associate Professor of the School of Digital Technologies, Narxoz University, Almaty, Kazakhstan; e-mail: daniyar.sultan@narxoz.kz. ORCID:https://orcid.org/0000-0002-1611-1923.

Rustam Abdrakhmanov – Candidate of Technical Sciences, Associate Professor of the Humanities school, International University of Tourism and Hospitality, Turkistan, Kazakhstan; e-mail: abdrakhmanov.rustam@iuth.edu.kz. ORCID: https://orcid.org/0000-0002-5508-389X.

Tursinbay Turymbetov – Candidate of Technical Sciences, Associate Professor of the Humanities school, International University of Tourism and Hospitality, Turkistan, Kazakhstan; e-mail: t.turimbetov@iuth.edu.kz. ORCID: https://orcid.org/0000-0003-0178-8701.

Esref Adali – PhD, Professor of the Faculty of Computer and Informatics, Istanbul Technical University; e-mail: adali@itu.edu.tr. ORCID: https://orcid.org/0000-0002-1561-8255.

Gulvira Bekeshova* – Senior Lecturer at the Department of Information Security, IT Faculty at the L.N. Gumilyov ENU, Astana, Kazakhstan; e-mail: gulvirabauyrzhanovna@gmail.com. ORCID: https://orcid.org/0000-0002-1635-4693.

Сведения об авторах

Данияр Султан – PhD, доцент факультета цифровых технологий Университета Нархоз, Алматы, Казахстан; e-mail:daniyar.sultan@narxoz.kz. ORCID:https://orcid.org/0000-0002-1611-1923.

Рустам Абдрахманов – кандидат технических наук, доцент, Международный университет туризма и гостеприимства, Туркестан, Казахстан; e-mail: abdrakhmanov.rustam@iuth.edu.kz. ORCID: https://orcid.org/0000-0002-5508-389X.

Турсынбай Турымбетов – кандидат технических наук, доцент, Международный университет туризма и гостеприимства, Туркестан, Казахстан; e-mail: t.turimbetov@iuth.edu.kz. ORCID: https://orcid.org/0000-0003-0178-8701.

Эшреф Адалы – PhD, профессор факультета вычислительной техники и информатики Стамбульского технического университета; e-mail adali@itu.edu.tr. ORCID: https://orcid.org/0000-0002-1561-8255.

Гульвира Бауыржановна Бекешова* — магистр технических наук, старший преподаватель кафедры Информационной безопасности факультета информационных технологий ЕНУ им. Л.Н. Гумилева, Астана, Казахстан; e-mail: gulvirabauyrzhanovna@gmail.com. ORCID: https://orcid.org/0000-0002-1635-4693.

Авторлар туралы мәліметтер

Данияр Султан – Нархоз университетінің Цифрлық технологиялар мектебінің доценті, Алматы, Қазақстан; e-mail: daniyar.sultan@narxoz.kz. ORCID:https://orcid.org/0000-0002-1611-1923.

Рустам Абдрахманов – техника ғылымдарының кандидаты, Халықаралық туризм және меймандостық университетінің доценті, Түркістан қ., Қазақстан; e-mail: abdrakhmanov.rustam@iuth.edu.kz. ORCID: https://orcid.org/0000-0002-5508-389X.

Турсынбай Турымбетов – техника ғылымдарының кандидаты, Халықаралық туризм және меймандостық университетінің доценті, Түркістан қ., Қазақстан, e-mail: t.turimbetov@iuth.edu.kz. ORCID: https://orcid.org/0000-0003-0178-8701.

Эшреф Адалы – PhD, Стамбул техникалық университетінің компьютер және информатика факультетінің профессоры; e-mail adali@itu.edu.tr. ORCID: https://orcid.org/0000-0002-1561-8255.

Гульвира Бауыржановна Бекешова* — Л.Н. Гумилева ат. ЕҰУ Ақпараттық технологиялар факультетінің Ақпараттық қауіпсіздік кафедрасының аға оқытушысы, техникалық ғылымдар магистрі, Астана, Қазақстан; e-mail: gulvirabauyrzhanovna@gmail.com. ORCID: https://orcid.org/0000-0002-1635-4693.

Received 20.10.2025
Revised 25.11.2025
Accepted 04.12.2025