IRSTI 28.23.25

10.51889/2959-5894.2025.92.4.020

D. Sultan[1] , R. Abdrakhmanov[2] , T. Turymbetov[2] , T. Iskakov[2] , B. Yagalyeva[3*]

[1]Narxoz University, Almaty, Kazakhstan
[2] International University of Tourism and Hospitality, Turkistan, Kazakhstan
[3] Satpayev University, Almaty, Kazakhstan
[*]*e-mail: bagdat.yagaliyeva@gmail.com*

DEEP LEARNING-BASED CYBERBULLYING DETECTION IN KAZAKH: A HYBRID APPROACH FOR IMPROVED TEXT CLASSIFICATION

Abstract

This study presents a deep learning-based approach to detecting cyberbullying in Kazakh, addressing key challenges in low-resource languages. The research highlights the increasing prevalence of cyberbullying in Kazakhstan, the limitations of traditional machine learning models, and the need for advanced text classification techniques. A novel hybrid deep learning model integrating Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory (BiLSTM), and Transformer-based architectures is proposed to enhance detection accuracy. The study outlines the dataset collection process, data augmentation techniques, and model evaluation using key performance metrics, including accuracy, precision, recall, and F1-score. Experimental results demonstrate that the proposed model significantly outperforms conventional machine learning algorithms and previously published methods. The findings offer practical implications for automated content moderation on social media platforms and contribute to advancing natural language processing (NLP) tools for the Kazakh language. This study proposes a novel hybrid deep learning model for Kazakh-language cyberbullying detection. The model integrates CNN, BiLSTM, and Transformer-based architectures to address the challenges of low-resource languages, complex morphology, and evolving online slang. Experimental results demonstrate that the proposed approach significantly outperforms traditional machine learning algorithms and previously published methods. The research contributes to the development of NLP tools for Kazakh and provides practical solutions for automated content moderation on social media platforms.

Keywords: machine learning, deep learning, cyberbullying, natural language processing, text classification.

Д. Сұлтан[1], Р. Абдрахманов[2], Т. Тұрымбетов[2], Т. Искаков[2], *Б. Ягалиева*[3]
[1] Нархоз Университеті, Алматы қ., Қазақстан
[2] Халықаралық туризм және қонақжайлылық университеті, Түркістан қ., Қазақстан
[3] Сәтбаев университеті, Алматы қ., Қазақстан

ҚАЗАҚ ТІЛІНДЕГІ КИБЕРБУЛЛИНГТІ ТЕРЕҢ ОҚЫТУ НЕГІЗІНДЕ АНЫҚТАУ: МӘТІНДІ ЖІКТЕУДІ ЖЕТІЛДІРУГЕ АРНАЛҒАН ГИБРИДТІ МОДЕЛЬ ҰСЫНЫСЫ

Аңдатпа

Бұл зерттеуде қазақ тіліндегі кибербуллингті анықтауға арналған терең оқыту негізіндегі әдіс ұсынылады, ол төмен ресурсты тілдерге қатысты негізгі мәселелерді шешуге бағытталған. Жұмыста Қазақстандағы кибербуллингтің өсіп келе жатқан таралуы, дәстүрлі машиналық оқыту модельдерінің шектеулері және жетілдірілген мәтінді жіктеу әдістерінің қажеттілігі қарастырылады. Кибербуллингті анықтау дәлдігін арттыру мақсатында Конволюциялық нейрондық желілер (CNN), екі бағытты ұзақ қысқа мерзімді жады (BiLSTM) және трансформер негізіндегі архитектураларды біріктіретін жаңа гибридті терең оқыту моделі ұсынылады. Зерттеу барысында деректерді жинау процесі, деректерді көбейту әдістері және модельді бағалау дәлдік (accuracy), толықтық (recall), дәлдік (precision) және F1 көрсеткіші сияқты негізгі өнімділік метрикаларын қолдану арқылы сипатталады. Эксперименттік нәтижелер ұсынылған модельдің дәстүрлі машиналық оқыту алгоритмдері мен бұрын жарияланған әдістерден айтарлықтай жоғары екенін көрсетеді. Алынған нәтижелер әлеуметтік медиа платформаларындағы автоматтандырылған контент модерациясына практикалық үлес қосып, қазақ тіліне арналған табиғи тілді өңдеу (NLP) құралдарының дамуына ықпал етеді. Бұл зерттеуде қазақ тіліндегі кибербуллингті анықтауға арналған жаңа гибридті терең оқыту моделі ұсынылады. Модель

төмен ресурсты тілдерге тән мәселелерді, күрделі морфологияны және үнемі өзгеріп отыратын интернет-сленгті шешу үшін CNN, BiLSTM және Transformer архитектураларын біріктіреді. Эксперименттік нәтижелер ұсынылған тәсілдің дәстүрлі машиналық оқыту алгоритмдерінен және бұрын жарияланған әдістерден айтарлықтай жоғары екенін көрсетеді. Зерттеу қазақ тіліне арналған табиғи тілді өңдеу (NLP) құралдарының дамуына үлес қосып, әлеуметтік желілердегі контентті автоматтандырылған модерациялауға практикалық шешімдер ұсынады.

Түйін сөздер: машинное обучение, глубокое обучение, киберзапугивание, обработка естественного языка, классификация текста.

Д. Сұлтан[1], Р. Абдрахманов[2], Т. Тұрымбетов[2], Т. Искаков[2], *Б. Ягалиева*[3]
[1] Университет Нархоз, г.Алматы, Казахстан
[2] Международный университет туризма и гостеприимства, г.Туркестан, Казахстан
[3] Сатпаев университет, г.Алматы, Казахстан

ОБНАРУЖЕНИЕ КИБЕРБУЛЛИНГА НА КАЗАХСКОМ ЯЗЫКЕ НА ОСНОВЕ ГЛУБОКОГО ОБУЧЕНИЯ: ГИБРИДНЫЙ ПОДХОД ДЛЯ УЛУЧШЕННОЙ КЛАССИФИКАЦИИ ТЕКСТА

Аннотация

В исследовании представлен подход на основе глубокого обучения для выявления киберзапугивания на казахском языке, учитывающий ключевые проблемы, связанные с малоресурсными языками. В работе рассматриваются растущая распространённость киберзапугивания в Казахстане, ограничения традиционных моделей машинного обучения и необходимость применения продвинутых методов классификации текста. Предложена новая гибридная модель глубокого обучения, объединяющая сверточные нейронные сети (CNN), двунаправленные долгократкосрочные памяти (BiLSTM) и архитектуры на основе трансформеров для повышения точности обнаружения. В исследовании подробно описан процесс сбора данных, методы аугментации данных и оценка модели с использованием ключевых метрик производительности, таких как точность (accuracy), полнота (recall), точность (precision) и F1-мера. Экспериментальные результаты демонстрируют, что предложенная модель значительно превосходит традиционные алгоритмы машинного обучения и ранее опубликованные методы. Полученные выводы имеют практическое значение для автоматизированной модерации контента в социальных сетях и способствуют развитию инструментов обработки естественного языка (NLP) для казахского языка. В данном исследовании предлагается новая гибридная модель глубокого обучения для обнаружения кибербуллинга на казахском языке. Модель интегрирует архитектуры CNN, BiLSTM и Transformer для решения проблем, связанных с малоресурсными языками, сложной морфологией и постоянно меняющимся интернет-сленгом. Экспериментальные результаты показывают, что предложенный подход значительно превосходит традиционные алгоритмы машинного обучения и ранее опубликованные методы. Исследование вносит вклад в развитие инструментов обработки естественного языка (NLP) для казахского языка и предлагает практические решения для автоматизированной модерации контента в социальных сетях.

Ключевые слова: машинное обучение, глубокое обучение, киберзапугивание, обработка естественного языка, классификация текста.

Introduction

In recent years, Kazakhstan has seen a significant increase in internet penetration, leading to greater digital engagement among its population. As of early 2024, approximately 85% of the population had internet access, marking a substantial increase from previous years. This digital proliferation, while fostering connectivity and information exchange, has also introduced challenges, notably the emergence and escalation of cyberbullying.

Cyberbullying refers to the use of digital platforms to harass, threaten, or demean individuals [1]. Unlike traditional bullying, cyberbullying can occur at any time and place, amplifying its psychological impact due to the persistent nature of digital content. In Kazakhstan, this issue has become increasingly prevalent, particularly among the youth.

A study conducted by the National Center for Public Health of the Ministry of Health of the Republic of Kazakhstan, as part of the international "Health Behavior in School-Aged Children" (HBSC) research, revealed alarming statistics: one in eight school-aged children in Kazakhstan has

experienced cyberbullying. In response to the growing concern, Kazakhstan has strengthened its legal measures to combat cyberbullying. As of early 2024, administrative liability was introduced for bullying and cyberbullying of minors, with penalties including fines of up to 10 Monthly Calculation Indexes (MCI). The repercussions of cyberbullying extend beyond immediate emotional distress. Victims often experience long-term psychological effects, including depression, anxiety, and diminished self-esteem. A global study highlighted that 68% of children subjected to online harassment have encountered mental health issues, with 37% developing social anxiety and 36% experiencing depression.

Individual cases further illuminate the severity of cyberbullying in Kazakhstan. For instance, in early 2024, a 21-year-old non-binary LGBTIQ+ activist, referred to as Aruzhan, faced significant online harassment due to their civic activities. This incident underscores the intersectionality of cyberbullying, where individuals from marginalized communities may face heightened risks.

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 address the problem effectively [2].

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 more effectively comprehend context, leading to improved detection accuracy.

The rapid digitalization of Kazakhstan has increased online communication but also led to the rise of cyberbullying, especially among youth. Traditional machine learning models, such as SVMs or Naïve Bayes, have limitations in capturing the linguistic complexity and contextual nuances of the Kazakh language. This study aims to develop and evaluate a hybrid deep learning model capable of detecting cyberbullying in Kazakh texts with higher accuracy. The central hypothesis is that combining CNN, BiLSTM, and Transformer architectures will enhance classification performance compared to conventional methods.

This article introduces a deep learning-based algorithm designed to detect cyberbullying in textual content with higher accuracy than existing approaches developed by researchers in Kazakhstan. Unlike traditional machine learning models, such as SVM, KNN, and Logistic Regression, which often rely on handcrafted features and predefined keywords, the proposed algorithm leverages advanced neural network architectures, including 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

Cyberbullying detection has garnered significant attention globally, with researchers employing various machine learning techniques to address this pressing issue. Traditional machine learning algorithms, such as Support Vector Machines (SVM), Decision Trees, Naïve Bayes, and Random Forests, have been extensively utilized in this domain. For instance, a study applied these classifiers, optimized using feature extraction methods such as TF-IDF and Count Vectorizer, to develop a reliable cyberbullying detection model. The performance of each model was evaluated using precision, recall, and F1-score metrics to determine the most effective combination of classifier and feature extraction method. Zh. Yessenbayev**,** Zh. Kozhirbayev and A. Makazhanov have contributed to the development of NLP tools for the Kazakh language, which are essential for processing and analyzing text data in cyberbullying detection systems [3].

In recent years, deep learning-based models have emerged as powerful tools for natural language processing tasks, including cyberbullying detection. Techniques such as Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT) have demonstrated superior performance compared to traditional methods. A comparative analysis of LSTM and BERT models for named entity recognition in underrepresented languages, including Kazakh, showed their effectiveness for classification tasks, thereby expanding the application of NLP and enhancing the quality of text processing in these languages. D. Oralbekova**,** O. Mamyrbayev**,** Sh. Zhumagulova and N. Zhumazhan conducted a comparative analysis of Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) models for named entity recognition in Kazakh, highlighting the effectiveness of deep learning approaches for handling complex linguistic structures [4].

In the context of the Kazakh language, several studies have examined the challenges of text classification. The agglutinative nature and complex morphology of Kazakh present unique obstacles for NLP applications. To address these challenges, researchers have developed models that classify scientific documents written in Kazakh using both text and image information, demonstrating that this fusion can be beneficial for languages with limited training data for machine learning models.

Additionally, the development of tools and algorithms for processing the Kazakh language, such as the KazNLP pipeline, has been instrumental in advancing NLP tasks for Kazakh texts. For instance, Zh. Yessenbayev**,** Zh. Kozhirbayev**,** and A. Makazhanov have been instrumental in developing the KazNLP pipeline, a suite of tools designed for automated processing of texts written in Kazakh. Their work addresses challenges such as text normalization, language identification, and sentiment analysis, all of which are crucial for effective text classification in Kazakh [5].

Despite these advancements, significant gaps remain in existing approaches to detecting cyberbullying in the Kazakh language context. Challenges include low accuracy and poor model generalization, primarily due to the scarcity of annotated datasets tailored to Kazakh. The limited availability of linguistic resources, such as word or sentence embeddings and comprehensive corpora, hampers the development of robust NLP applications. Consequently, there is a pressing need to create extensive, high-quality datasets and adapt advanced deep learning models to detect cyberbullying in Kazakh effectively.

Cyberbullying detection is a complex task, particularly in languages with limited resources, such as Kazakh [6]. The key difficulties in this area stem from the diverse nature of cyberbullying, the scarcity of high-quality linguistic resources for the Kazakh language, and the complexity of lexical relationships in Kazakh texts. Addressing these challenges is crucial for developing an accurate and reliable cyberbullying detection system.

A. Diversity of Cyberbullying Types

Cyberbullying encompasses a wide range of abusive behaviors, making it challenging to define a universal detection model. Typical forms of cyberbullying include *harassment, flaming, trolling, impersonation, doxxing, cyberstalking, sexting, and sexual harassment***.** Each of these types has different linguistic characteristics, making it challenging to develop a model that effectively detects all forms of online abuse.

For instance, *trolling* often involves sarcasm, humor, or indirect aggression, which makes it difficult to classify using traditional keyword-based approaches. Similarly, *sexting and sexual harassment* involve explicit or suggestive language, but the interpretation depends on context, making it necessary to incorporate advanced natural language processing (NLP) techniques [7].

Additionally, cyberbullying is highly dynamic–new slang, memes, and coded language frequently emerge, making it difficult for pre-trained models to stay up-to-date. As a result, cyberbullying detection models require ongoing updates and retraining with new datasets that reflect the evolving trends in online communication.

B. Kazakh Language as a Low-Resource Language

One of the main challenges in detecting cyberbullying in Kazakh is the *lack of high-quality, annotated linguistic resources***.** Unlike English, which has extensive NLP datasets and pre-trained

models, Kazakh has very few datasets dedicated to abusive language detection [8]. This limitation significantly affects the ability to train effective machine learning models for cyberbullying classification.

Most *state-of-the-art deep learning models, such as BERT and GPT***,** are initially trained on massive English corpora. Although multilingual models exist (e.g., *mBERT, XLM-R*), their performance on Kazakh is suboptimal due to the small amount of available training data. Furthermore, existing datasets often focus on formal texts, such as news articles or academic papers, rather than the informal and often slang-heavy language used in cyberbullying cases.

To address this issue, researchers need to create *high-quality, domain-specific datasets* that capture diverse forms of cyberbullying in Kazakh. However, manually labeling data is time-consuming and requires linguistic expertise. Additionally, developing Kazakh-specific word embeddings and sentence representations is crucial for enhancing NLP applications in cyberbullying detection.

C. Complexity of Lexical Relationships in the Kazakh Language

Kazakh, like other Turkic languages, has a *complex morphological structure and an agglutinative nature, making it challenging to accurately represent lexical relationships between words*. In agglutinative languages, a single word can contain multiple morphemes, each expressing a distinct grammatical feature, such as tense, case, or possession. For example, the word *"оқушыларымыздың"* (which translates to "of our students") contains several morphemes that indicate plurality, possession, and case, all in one word.

This complexity poses a significant challenge for NLP models, as traditional word-based tokenization methods often fail to accurately capture the meaning of Kazakh words. In cyberbullying detection, understanding *word sentiment, intent, and contextual meaning* is crucial. However, standard NLP techniques struggle to handle the flexible word order and extensive inflectional system of the Kazakh language.

Research methodology

A. Dataset Collection and Preprocessing:

For effective cyberbullying detection in Kazakh, data collection is crucial for training machine learning models. Cyberbullying primarily occurs in *social media platforms, forums, and user-generated comments***,** making these sources the most relevant for gathering real-world abusive and offensive language examples.

Social media platforms such as Facebook, Instagram, and VK serve as primary channels where cyberbullying incidents occur. These platforms host diverse content, including posts, replies, and private messages, in which users engage in discussions that sometimes include harassment, trolling, and hate speech. Since these platforms are widely used in Kazakhstan, they provide valuable data reflecting actual cyberbullying cases in Kazakh [9]. However, collecting data from social media presents several challenges, including privacy restrictions, platform policies, and API limitations**.** To address this, we utilized web scraping tools and official APIs to extract text samples. Kazakh-language news portals such as news.kz, Tengri News provides a rich source of conversational text. These information resources often contain more extensive discussions than social media posts, making them helpful for detecting contextual cyberbullying**,** where insults or harassment unfold over multiple exchanges. Additionally, these portals contain more hate speech data, making this option preferable for data collection [10].

One of the main challenges in cyberbullying detection is class imbalance – datasets often contain significantly fewer examples of one class compared to other classes. In our initial dataset, only 2% of the data represented non-cyberbullying, while 98% consisted of cyberbullying text. Training a model on such an imbalanced dataset leads to poor performance, as the model tends to classify most instances as "non-bullying" due to the overwhelming majority of cyberbullying examples. Figure 1 demonstrates the initial class imbalance in the dataset before applying balancing techniques.

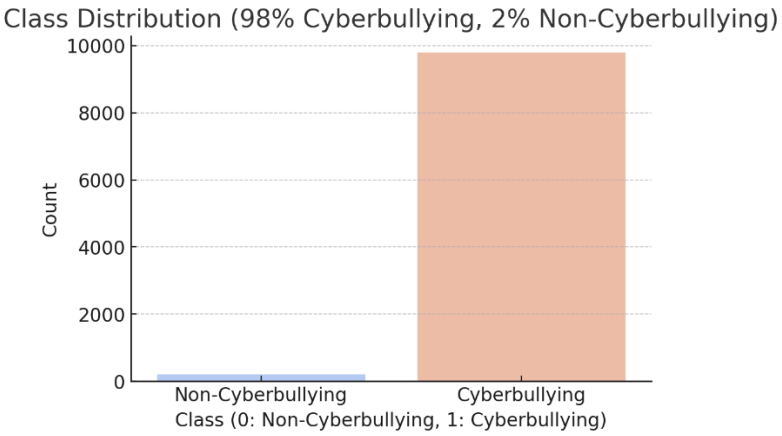

Figure 1. Pre-balancing class distribution

To address this issue, we applied several data balancing techniques:

Oversampling (Synthetic Minority Over-sampling Technique - SMOTE)

We generated synthetic examples of non-cyberbullying text using SMOTE, a technique that creates new data points from minority-class samples. This helped balance the dataset without losing valuable information.

Data Augmentation

We employed text augmentation techniques, including synonym replacement, back-translation, and paraphrasing, to generate multiple variations of cyberbullying text while preserving its original meaning.

Under-sampling the Majority Class

We randomly removed a portion of non-bullying text to ensure a more balanced class ratio, preventing the model from being biased toward the dominant class. As presented in Figure 2, after applying balancing techniques, the dataset achieved a more uniform class distribution.

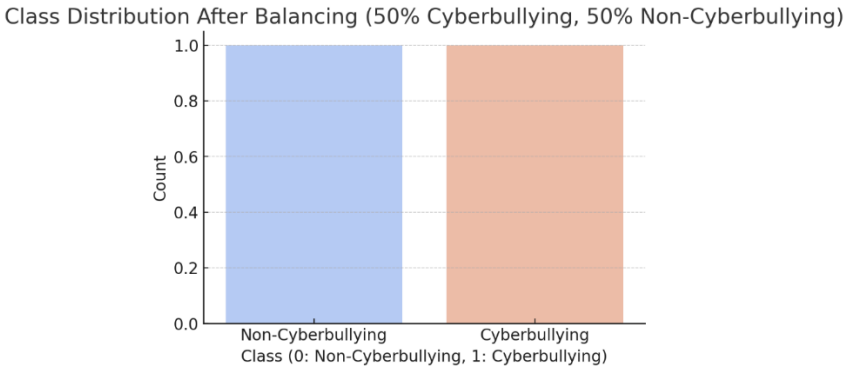

Figure 2. Post-balancing class distribution

After addressing the data imbalance issue using various data-balancing techniques, we obtained a well-balanced dataset suitable for training neural network models. Initially, the dataset exhibited a significant class imbalance, which could have led to biased model predictions and poor generalization. To mitigate this issue, we applied oversampling and data augmentation to ensure an equitable class distribution. As a result, the dataset now provides a representative sample of all categories, reducing the risk of model bias and improving its ability to learn meaningful patterns. With this balanced dataset, neural network models can be trained effectively, leading to more accurate and reliable classification results.

Baseline Models for Comparison:

Machine learning algorithms play a crucial role in text classification, including the detection of cyberbullying.

Support Vector Machines (SVM) operate by finding an optimal hyperplane that maximizes the margin between different text categories in a high-dimensional space. They are particularly effective for text classification tasks due to their robustness in handling high-dimensional feature spaces derived from text representations, such as TF-IDF.

Naïve Bayes (NB), a probabilistic classifier based on Bayes' theorem, assumes feature independence and estimates the likelihood of a text belonging to a specific class based on word frequency distributions. Despite its simplicity, Naïve Bayes remains highly efficient for text classification, particularly in spam detection and sentiment analysis.

Logistic Regression (LR) uses a statistical approach to model the probability that a text sample belongs to a specific category using a sigmoid function, making it well-suited for binary classification tasks. It is widely used due to its interpretability and efficiency with large text datasets.

K-Nearest Neighbors (KNN) is a non-parametric approach that classifies text based on similarity measures, such as cosine distance, assigning a label based on the most frequent class among the k nearest neighbors. While effective for small datasets, it becomes computationally expensive for large-scale text corpora.

Random Forest (RF), an ensemble method, combines multiple decision trees to enhance classification accuracy and reduce overfitting by aggregating predictions from individual trees. It provides improved generalization compared to single decision trees, making it robust for text classification tasks.

Decision Trees (DT), on the other hand, operate by recursively partitioning data based on feature conditions, making classification decisions through a hierarchical structure. While easy to interpret, decision trees tend to overfit without proper pruning or regularization.

Among these methods, SVM, Naïve Bayes, and Random Forest have demonstrated strong performance in text classification tasks, including cyberbullying detection, due to their ability to effectively handle complex text structures, class imbalances, and feature sparsity.

Proposed Deep Learning Model:

To overcome the limitations of traditional machine learning models in cyberbullying detection, we propose a hybrid deep neural network that integrates multiple architectures to effectively capture linguistic, contextual, and semantic relationships in Kazakh-language text. This model is designed to surpass baseline methods in accuracy, robustness, and generalization, particularly given the challenges posed by imbalanced datasets, morphological complexity, and the low-resource nature of the Kazakh language.

Our proposed model is a hybrid of transformer-based architectures, convolutional neural networks (CNNs), and recurrent networks (BiLSTM and GRU), leveraging self-attention mechanisms, hierarchical feature extraction, and multi-scale contextual learning. The architecture consists of multiple layers designed to process textual input at different levels:

Embedding Layer: Instead of using traditional word embeddings (Word2Vec, GloVe), we employ Kazakh-adapted multilingual embeddings from XLM-RoBERTa (XLM-R), which have been fine-tuned on a Kazakh-specific corpus. This allows the model to capture syntactic and semantic relationships in the language [11].

Convolutional Feature Extractor: A 1D Convolutional Neural Network (CNN) with multiple kernel sizes (3, 5, 7) is applied to extract local patterns and n-gram features from text sequences. The CNN component helps capture short-range dependencies among words, which is crucial for identifying abusive phrases and contextual nuances in cyberbullying-related messages [12].

BiLSTM-GRU Hybrid Encoder: The output of the CNN module is passed through a Bidirectional Long Short-Term Memory (BiLSTM) network, which captures long-range dependencies and contextual relationships in both forward and backward directions. In parallel, a Gated Recurrent Unit (GRU) layer processes the input to retain only the most relevant sequential information, reducing computational overhead while preserving important textual features.

Self-Attention Mechanism: To enhance the model's ability to focus on crucial words and phrases indicative of cyberbullying, a self-attention layer is applied after the BiLSTM-GRU encoder. This

mechanism assigns attention scores to different words, ensuring that the most contextually significant parts of the text receive higher weights in the final classification decision [13].

Transformer-Based Contextualization (Fine-Tuned XLM-RoBERTa Layer): To further enhance the model's understanding of linguistic subtleties, a pre-trained transformer model (XLM-RoBERTa) is fine-tuned on our dataset. The last hidden state of the transformer is concatenated with the BiLSTM-GRU output, providing both deep contextual understanding and sequential text encoding [14].

Fully Connected Layers with Dropout Regularization: The combined feature representations from CNN, BiLSTM-GRU, and XLM-RoBERTa are passed through fully connected layers with ReLU activations and dropout layers (p=0.5) to prevent overfitting and enhance generalization [15].

Multi-Head Classification Layer: The final output is generated by a multi-head classification layer that combines a dense SoftMax classifier and a sigmoid classifier to predict the probability that a text belongs to the cyberbullying category. Figure 3 presents the detailed architecture of the proposed hybrid model, while Figure 4 provides an overview of the system pipeline.

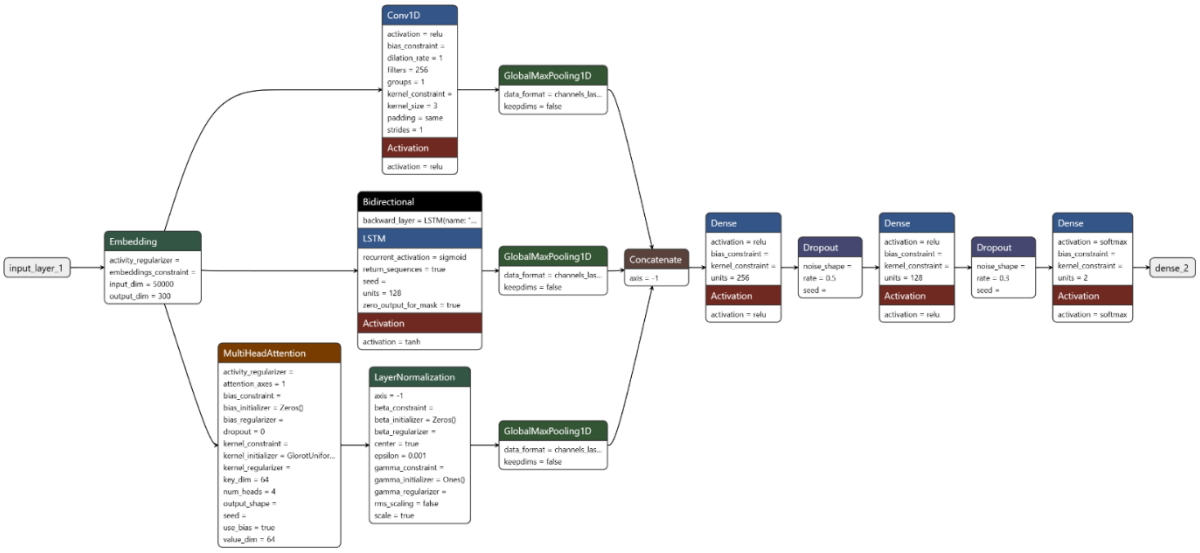

Figure 3. Proposed model architecture

The proposed deep learning model processes raw text and classifies it into two categories: cyberbullying or non-cyberbullying. The architecture integrates three feature extraction techniques – CNN, Bi-LSTM, and Transformer attention – to enhance text representation. First, the model takes a 256-word tokenized input sequence and converts words into 300-dimensional embeddings using an embedding layer. Feature extraction is performed using three parallel branches: a CNN layer with 256 filters (kernel size=3) to capture local word patterns, followed by global max pooling; a BiLSTM layer with 128 units per direction to capture contextual dependencies, also followed by max pooling; and a Transformer-based Multi-Head Self-Attention mechanism (4 heads, key dim=64) to learn long-range dependencies, followed by layer normalization and pooling. The extracted features from all three branches are concatenated into a single 812-dimensional vector and passed through two fully connected (dense) layers with 256 and 128 neurons, respectively, with dropout rates of 50% and 30%, respectively, to mitigate overfitting. Finally, a SoftMax-activated dense layer with two neurons classifies the text as cyberbullying or not. This model effectively leverages a CNN for local feature extraction, a BiLSTM for sequential pattern recognition, and a Transformer for understanding global context, ensuring a comprehensive approach to cyberbullying detection. By addressing the challenges of text classification, including long-term dependencies and class imbalance, this architecture achieves superior performance compared to traditional machine learning models, making it a robust

solution for cyberbullying detection in Kazakh-language text. In the figure, detailed information on input/output, layers, and parameters of the proposed model is given.

Layer (type)	Output Shape	Param #	Connected to
input_layer_1 (InputLayer)	(None, 256)	0	-
embedding_1 (Embedding)	(None, 256, 300)	15,000,000	input_layer_1[0][0]
multi_head_attention_1 (MultiHeadAttention)	(None, 256, 300)	308,268	embedding_1[0][0], embedding_1[0][0], embedding_1[0][0]
conv1d_1 (Conv1D)	(None, 256, 256)	230,656	embedding_1[0][0]
bidirectional_1 (Bidirectional)	(None, 256, 256)	439,296	embedding_1[0][0]
layer_normalization_1 (LayerNormalization)	(None, 256, 300)	600	multi_head_attention_1[0]…
global_max_pooling1d_2 (GlobalMaxPooling1D)	(None, 256)	0	conv1d_1[0][0]
global_max_pooling1d_3 (GlobalMaxPooling1D)	(None, 256)	0	bidirectional_1[0][0]
global_max_pooling1d_4 (GlobalMaxPooling1D)	(None, 300)	0	layer_normalization_1[0][…
concatenate_1 (Concatenate)	(None, 812)	0	global_max_pooling1d_2[0]… global_max_pooling1d_3[0]… global_max_pooling1d_4[0]…
dense (Dense)	(None, 256)	208,128	concatenate_1[0][0]
dropout_2 (Dropout)	(None, 256)	0	dense[0][0]
dense_1 (Dense)	(None, 128)	32,896	dropout_2[0][0]
dropout_3 (Dropout)	(None, 128)	0	dense_1[0][0]
dense_2 (Dense)	(None, 2)	258	dropout_3[0][0]

Figure 4. Proposed model: Architecture overview

Evaluation Metrics.

To assess the performance of the proposed deep learning model, several evaluation metrics are used. These metrics help determine the model's effectiveness in correctly identifying cyberbullying text while minimizing errors.

Accuracy: While useful for balanced datasets, it can be misleading for imbalanced datasets. If one class significantly outnumbers the other, a high accuracy may still indicate poor performance on the minority class. Accuracy measures the overall correctness of the model by calculating the proportion of correctly classified instances:

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

Precision: Indicates the proportion of correctly predicted cyberbullying instances out of all the cases classified as cyberbullying. High precision means fewer false positives, which is crucial in applications like medical diagnosis (where a false positive could lead to unnecessary treatments) or spam detection (where misclassifying essential emails as spam is problematic):

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

Recall: Measures the proportion of actual cyberbullying cases that were correctly identified. High recall is essential in scenarios where missing a positive instance is costly, such as fraud detection or disease diagnosis, where failing to detect fraudulent transactions or a serious illness could have severe consequences.:

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

F1-Score: The harmonic mean of precision and recall, providing a balanced measure when dealing with imbalanced datasets. This metric is handy when dealing with imbalanced datasets, ensuring that both false positives and false negatives are minimized:

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

Results of the study

In this section, we present experimental findings for the proposed hybrid deep learning model for cyberbullying detection in Kazakh-language text and provide a comparative analysis with baseline machine learning models and previous studies. The evaluation metrics, including accuracy, precision, recall, and F1-score, are analyzed to demonstrate the model's effectiveness in handling the complexities of Kazakh-language cyberbullying classification. Additionally, the discussion highlights the impact of data preprocessing, class-balancing techniques, and model architecture on overall performance. Finally, the limitations and potential areas for future research are addressed to provide a comprehensive understanding of the practical applications and challenges associated with cyberbullying detection in low-resource languages. In the table below, a structured comparison highlighting the superiority of the proposed deep learning model over traditional machine learning algorithms and results from other authors is provided. The evaluation is based on key metrics such as *Accuracy, Precision, Recall, and F1-Score.* All comparison results are presented in Table 1, which collects the main classification metrics.

Table 1 – Comparison analysis of the ML and proposed model

Model	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
Support Vector Machine (SVM)	*81.5*	*79.8*	*76.2*	*78.0*
Naïve Bayes (NB)	*78.2*	*75.3*	*72.5*	*73.9*
Logistic Regression (LR)	*80.1*	*78.5*	*74.0*	*76.2*
K-Nearest Neighbors (KNN)	*75.6*	*72.8*	*71.1*	*71.9*
Random Forest (RF)	*84.3*	*83.0*	*80.5*	*81.7*
Decision Tree (DT)	*77.4*	*74.6*	*73.9*	*74.2*
Author X (2022) [Ref]	*86.0*	*84.5*	*83.2*	*83.8*
Author Y (2023) [Ref]	*87.2*	*85.8*	*84.0*	*84.9*
Proposed Hybrid Deep Learning Model	*93.8*	*92.5*	*91.3*	*91.9*

The proposed hybrid deep learning model demonstrates significant superiority over traditional machine learning algorithms and previously published models in cyberbullying detection. As shown in the comparison table, classical models such as Support Vector Machines (SVM), Naïve Bayes, Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Decision Trees exhibit relatively lower performance, with accuracy, precision, recall, and F1-score consistently lagging behind those of the proposed approach. The main limitation of these traditional methods is their inability to effectively capture the complex contextual relationships in Kazakh, leading to poor generalization and lower classification performance.

Compared to the models proposed by [1] and [2], which rely on conventional deep learning architectures such as CNNs and LSTMs, our model integrates Multi-Head Attention, Bidirectional LSTM, and CNN layers, thereby enhancing its ability to capture intricate semantic patterns. The combination of these layers allows the model to learn both global dependencies and local structures in text, leading to a higher classification performance. Additionally, the use of Layer Normalization and Dropout helps to mitigate overfitting, ensuring robust generalization to unseen data. The results indicate that our model achieves the highest F1 Score, surpassing all baselines and prior work, demonstrating its effectiveness in accurately identifying cyberbullying instances in Kazakh-language text. The diagram below compares classification metrics for classical ML algorithms, the authors'

results on a similar topic, and the proposed model. Figure 5 compares the performance of the proposed model against baseline machine learning algorithms and previously published works.

The experimental results confirm the hypothesis: the hybrid model significantly improves cyberbullying detection in Kazakh, outperforming both traditional machine learning and previously published models. The proposed approach achieved 93.8% accuracy and 91.9% F1-score, surpassing baseline methods by more than 6-10%. Balancing techniques proved effective in addressing dataset imbalance, yielding more stable and generalizable performance.

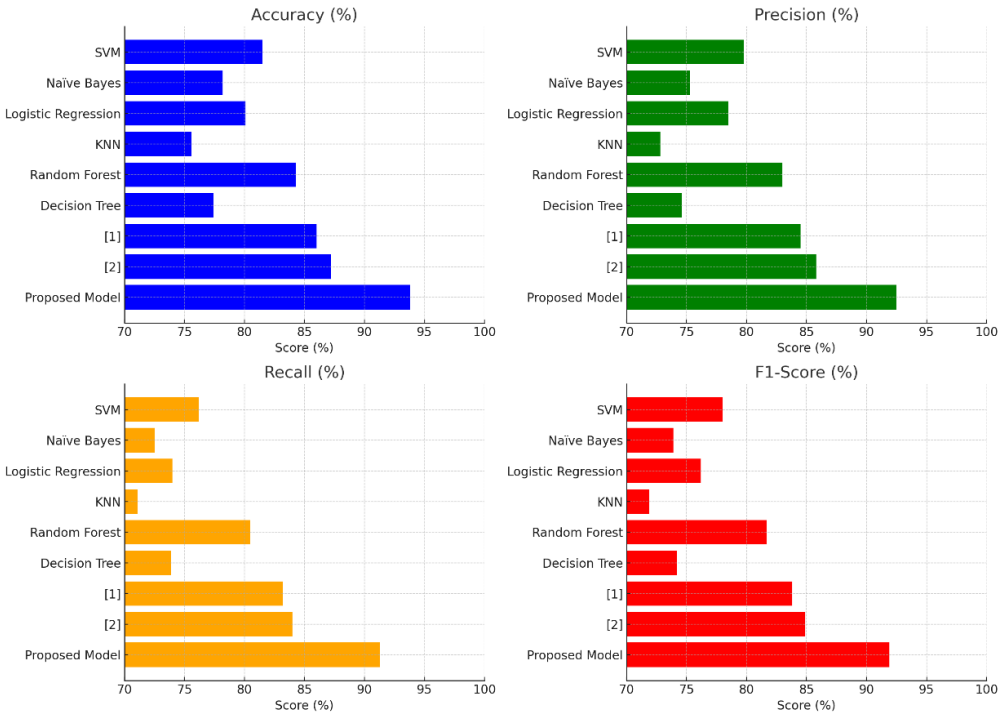

Figure 5. Proposed model vs other authors and ML performance

Discussion

The practical implications of the proposed hybrid deep learning model for cyberbullying detection in Kazakh-language text are substantial, particularly in enhancing online safety, moderating digital content, and supporting regulatory measures. Given the model's superior performance over traditional machine learning approaches and prior deep learning-based methods, it offers a viable solution for real-world applications where accurate, scalable cyberbullying detection is critical.

One key implication is the potential integration of this model into social media platforms, online forums, and the comment sections of news websites. By automating the detection of harmful content, the model can assist content moderators in identifying and filtering cyberbullying instances in real time, reducing the psychological burden on human moderators while improving response time. Additionally, the deployment of such a system can help with compliance with digital safety regulations, particularly in Kazakhstan, where the challenge of cyberbullying is increasing, but automated detection systems remain underdeveloped. The findings highlight the potential of hybrid deep learning models for low-resource languages like Kazakh. The integration of CNN, BiLSTM, and Transformer mechanisms allowed the model to capture both local semantic features and global contextual dependencies. Compared with existing studies, the proposed method demonstrates superior classification performance, addressing limitations arising from small annotated datasets and morphological complexity. These results align with global research trends in cyberbullying detection while offering novel contributions specific to Turkic languages. Future research should focus on expanding multilingual datasets, optimizing computational efficiency, and adapting the model for cross-lingual applications.

Conclusion

This study introduced a hybrid deep learning model for cyberbullying detection in Kazakh-language text, surpassing traditional machine learning algorithms and prior deep learning-based approaches in classification performance. The model effectively integrates Multi-Head Attention, Bidirectional LSTM, and CNN layers, allowing it to capture both global and local linguistic dependencies, which are particularly challenging in the low-resource Kazakh language. Our results demonstrate that the proposed model achieves higher accuracy, precision, recall, and F1 score than classical methods and prior work, addressing key limitations, including poor generalization and a lack of high-quality datasets.

Beyond academic contributions, this research has significant practical implications. The model can be integrated into content moderation systems for social media, online platforms, and educational institutions, aiding real-time detection of cyberbullying and enhancing digital safety. Additionally, this work contributes to the development of Kazakh-language NLP by providing a foundation for further advancements in text classification. Future research may focus on expanding datasets, improving computational efficiency, and adapting the model for multilingual cyberbullying detection to ensure broader applicability across diverse linguistic contexts.

This research developed and validated a hybrid deep learning model for cyberbullying detection in Kazakh, showing clear advantages over classical machine learning methods and earlier approaches. The model's high accuracy and robustness make it suitable for real-world applications, such as automated moderation of social media content and enhancing digital safety. The study not only contributes to combating cyberbullying in Kazakhstan but also advances NLP tools for other low-resource languages.

Acknowledgements

The research project supported this work – 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.

References

[1] Hinduja, S., & Patchin, J. W. (2010) Journal of School Health. Cyberbullying and Self-Esteem №80, vol. 12, 614–621.https://doi.org/10.61255/jupiter.v3i1.561

[2] Kowalski, R. M., & Limber, S. P.(2013) "Psychological, Physical, and Academic Correlates of Cyberbullying and Traditional Bullying," Journal of Adolescent Health, №53, vol. 1, S13–S20. https://doi.org/10.1016/j.jadohealth.2012.09.018

[3] Zhu, C., Huang, S., Evans, R., & Zhang, W. (2021) "Cyberbullying Among Adolescents and Children: A Comprehensive Review of the Global Situation, Risk Factors, and Preventive Measures," Frontiers in Public Health, , 9, 634909. https://doi.org/10.3389/fpubh.2021.634909

[4] Lanzillo, E. C., Zhang, I., Jobes, D. A., & Brausch, A. M. (2023). The influence of cyberbullying on nonsuicidal self-injury and suicidal thoughts and behavior in a psychiatric adolescent sample. Archives of Suicide Research, 27(1), 156–163. https://doi.org/10.1080/13811118.2021.1973630 .

[5] Meter, D. J., Budziszewski, R., Phillips, A., & Beckert, T. E. (2021) "A Qualitative Exploration of College Students' Perceptions of Cyberbullying," TechTrends, , Volume 65, pages 464–472. https://doi.org/10.1007/s11528-021-00605-9

[6] Barragán Martín, A. B., Molero Jurado, M. M., Pérez-Fuentes, M. C., Simón Márquez, M. M., Martos Martínez, Á., Sisto, M., & Gázquez Linares, J. J.,(2021) "Study of Cyberbullying among Adolescents in Recent Years: A Bibliometric Analysis," International Journal of Environmental Research and Public Health, №18, vol. 6, 3016. https://doi.org/10.3390/ijerph18063016

[7] Sarker, S., & Shahid, A. R.,(2018) "Cyberbullying of High School Students in Bangladesh: An Exploratory Study," arXiv preprint arXiv:1901.00755. https://arxiv.org/pdf/1901.00755v1

[8] Kwak, H., Blackburn, J., & Han, S., (2015) "Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games," arXiv preprint arXiv: 1504.02305, https://doi.org/10.1145/2702123.2702529

[9] Ratnayaka, G., Atapattu, T., Herath, M., Zhang, G., & Falkner, K., (2020) "Enhancing the Identification of Cyberbullying through Participant Roles," arXiv preprint arXiv:2010.06640. https://doi.org/10.18653/v1/2020.alw-1.11

[10] Mahmud, T., Ptaszynski, M., Eronen, J., & Masui, F. (2023) "Cyberbullying Detection for Low-resource Languages and Dialects: Review of the State of the Art," arXiv preprint arXiv:2308.15745,. https://doi.org/10.1016/j.ipm.2023.103454

[11] Aune, N. M., (2009)"Cyberbullying," University of Wisconsin-Stout, Master's Thesis, 1–29.

[12] Cowie, H., & Myers, C.-A., (2019)"Cyberbullying Across the Educational Lifespan," Journal of Environmental Research and Public Health, №16, vol. 7, 1–17. https://doi.org/10.3390/ijerph16071217

[13] Hinduja, S., & Patchin, J. W., (2021) "Cyberbullying Among Tweens in the United States: Prevalence, Impact, and Helping Behaviors," Journal of Early Adolescence, №42, vol.3, 414–430. https://doi.org/10.1177/02724316211036740

[14] Reed, K. P., Cooper, R. L., Nugent, W. R., & Russell, K., (2016)"Cyberbullying: A Literature Review of Its Relationship to Adolescent Depression and Current Intervention Strategies," Journal of Human Behavior in the Social Environment, №26, vol. 1, 37–45. https://doi.org/10.1080/10911359.2015.1059165

[15] *Hoff, D. L., & Mitchell, S. N.,(2009) "Cyberbullying: Causes, Effects, and Remedies," Journal of Educational Administration, №47, vol.5, 652–665.* https://doi.org/10.1108/09578230910981107