ИНФОРМАТИКА

COMPUTER SCIENCE

IRSTI 28.23.25

10.51889/2959-5894.2026.93.1.009

R. Abdrakhmanov[1] , D. Sultan[2] , T.Nazarbek[3] , T. Iskakov[1] , B. Yagaliyeva[4*]

[1] International University of Tourism and Hospitality, Turkistan, Kazakhstan
[2]Narxoz University, Almaty, Kazakhstan
[3] Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan
[4] Kazakh National Research Technical University named after K. I. Satpayev, Almaty, Kazakhstan
**e-mail:* bagdat.yagaliyeva@gmail.com

TRANSFORMER BASED BI-LSTM DEEP LEARNING MODEL FOR AUTOMATIC CYBERBULLYING DETECTION IN KAZAKH TEXTUAL DATA

Abstract

This paper presents a comprehensive study on the effectiveness of a novel hybrid model of LSTM and CNN in cyberbullying detection in online social media text. The effectiveness of the proposed model is tested against other traditional models of machine learning, namely SVM, Random Forest, and Decision Trees, in terms of accuracy, precision, recall, F-Score, and AUC-ROC curves. The proposed model is a hybrid of Long Short-Term Memory and Convolutional Neural Network, which combines the contextual intelligence of Long Short-Term Memory and the proficiency of Convolutional Neural Network in feature extraction. The results of the experiment show that the proposed model of LSTM and CNN is significantly superior to other models of machine learning in all evaluation parameters. Moreover, the results of the ROC curve also affirm that the proposed model is significantly more sensitive and specific in distinguishing cyberbullying and non-cyberbullying text. This paper also presents a novel tool that can help social media platforms in mitigating cyberbullying and harassment in a most efficient and effective way. Moreover, it also presents a novel area of research that can help in exploring the ethical and legal implications of deploying such efficient and sophisticated tools in cyberbullying and harassment.

Keywords: cyberbullying, artificial intelligence, deep learning, LSTM, CNN, Transformers.

Р. Абдрахманов[1], Д. Сұлтан[2], Т. Назарбек[3], Т. Искаков, Б. Ягалиева[3]
[1]Халықаралық туризм және меймандостық университеті, Түркістан қ., Қазақстан
[2]Нархоз Университеті, Алматы қ., Қазақстан
[3]Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан қ., Қазақстан
[4]Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы қ., Қазақстан

ҚАЗАҚ ТІЛІНДЕГІ МӘТІНДІК ДЕРЕКТЕРДІ АВТОМАТТЫ КИБЕРБУЛЛИНГТІ АНЫҚТАУҒА АРНАЛҒАН ТРАНСФОРМЕРЛІК BI-LSTM ТЕРЕҢ ОҚЫТУ МОДЕЛІ

Аңдатпа

Бұл мақалада онлайн әлеуметтік медиа мәтіндеріндегі кибербуллингті анықтау үшін жаңа гибридті LSTM-CNN моделінің тиімділігі туралы кешенді зерттеу ұсынылған. Зерттеуде ұсынылған модельдің SVM, кездейсоқ орман және шешім ағаштары сияқты дәстүрлі машиналық оқыту классификаторларымен салыстырғандағы өнімділігі дәлдік, дәлдік, еске түсіру, F-балы және AUC-ROC сияқты метрикаларды қолдана отырып бағаланады. Ұсынылған гибридті модель мәтіндік деректердің тізбекті және кеңістіктік өлшемдерін анықтауға бағытталған ұзақ қысқа мерзімді жад желілерінің контексттік өңдеу мүмкіндіктерін конволюциялық нейрондық желілердің мүмкіндіктерді алу қабілетімен біріктіреді. Тәжірибелердің нәтижелері LSTM-CNN моделінің дәстүрлі классификаторлардан айтарлықтай асып түсетінін, барлық бағалау метрикалары бойынша жоғары балл

жинайтынын көрсетеді. Сонымен қатар, ROC қисық талдаулары модельдің кибербуллинг және кибербуллинг емес жағдайларды ажыратудағы жоғары сезімталдығы мен ерекшелігін растайды. Бұл зерттеу кибербуллингті анықтауды жақсартудағы терең оқыту тәсілдерінің әлеуетін көрсетеді, әлеуметтік медиа платформалары үшін онлайн қудалауды тиімді түрде азайтудың қуатты құралын ұсынады. Зерттеу нәтижелерінде бақылау мен құпиялылықтың этикалық өлшемдерін ескере отырып, осындай озық анықтау жүйелерін енгізудің салдары да талқыланады. Болашақ бағыттарға модельді әртүрлі тілдік контексттерді өңдеуге бейімдеу және жіктеу дәлдігін арттыру үшін пайдаланушылардың пікірлерін біріктіруді зерттеу кіреді. Бұл зерттеу цифрлық қауіпсіздік және онлайн қауымдастықты басқару саласында күрделі, контекстке бейім технологияларды әзірлеу үшін прецедент болып табылады.

Түйін сөздер: кибербуллинг, жасанды интеллект, терең оқыту, LSTM, CNN, трансформерлер.

Р. Абдрахманов [1], Д. Сұлтан [2], Т. Назарбек[3], Т. Искаков[2], Б. Ягалиева[4]
[1]Международный университет туризма и гостеприимства, г.Туркестан, Казахстан
[2] Университет Нархоз, г.Алматы, Казахстан
[3]Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, г.Туркестан, Казахстан
[4]Казахский национальный исследовательский технический университет им. К. И. Сатпаева, г. Алматы, Казахстан

МОДЕЛЬ ГЛУБОКОГО ОБУЧЕНИЯ BI-LSTM НА ОСНОВЕ ТРАНСФОРМАТОРА ДЛЯ АВТОМАТИЧЕСКОГО ОБНАРУЖЕНИЯ КИБЕРБУЛЛИНГА В КАЗАХСКИХ ТЕКСТОВЫХ ДАННЫХ

Аннотация

В данной статье представлено всестороннее исследование эффективности новой гибридной модели LSTM-CNN для обнаружения кибербуллинга в текстах онлайн-социальных сетей. В исследовании оценивается производительность предложенной модели по сравнению с традиционными классификаторами машинного обучения, включая SVM, случайный лес и деревья решений, с использованием таких метрик, как точность, прецизия, полнота, F-мера и AUC-ROC. Предложенная гибридная модель объединяет возможности контекстной обработки сетей долговременной кратковременной памяти (LSTM) с возможностями извлечения признаков сверточных нейронных сетей (CNN), стремясь охватить как последовательные, так и пространственные измерения текстовых данных. Результаты экспериментов показывают, что модель LSTM-CNN значительно превосходит традиционные классификаторы, достигая высоких показателей по всем оценочным метрикам. Кроме того, анализ ROC-кривых дополнительно подтверждает превосходную чувствительность и специфичность модели в различении случаев кибербуллинга и случаев, не связанных с кибербуллингом. Данное исследование подчеркивает потенциал подходов глубокого обучения в повышении эффективности обнаружения кибербуллинга, предлагая мощный инструмент для социальных сетей, позволяющий эффективно бороться с онлайн-домогательствами. В результатах также обсуждаются последствия внедрения таких передовых систем обнаружения, учитывая этические аспекты наблюдения и конфиденциальности. Дальнейшие направления исследований включают адаптацию модели для обработки различных языковых контекстов и изучение интеграции отзывов пользователей для повышения точности классификации. Это исследование создает прецедент для разработки более сложных, контекстно-ориентированных технологий в области цифровой безопасности и управления онлайн-сообществами.

Ключевые слова: кибербуллинг, искусственный интеллект, глубокое обучение, LSTM, CNN, трансформеры.

Introduction

With the emergence of the digital age, social media sites are the main medium of interaction and information sharing. Nevertheless, there is a darker side to social media that is referred to as cyberbullying – a common issue that may lead to severe psychological damage to a person [1]. Cyberbullying is defined as the intentional exploitation of information technologies to spread false information that is humiliating or aggressive in nature regarding a particular individual. With the rise

of online interactions, it is of utmost importance to develop efficient automated tools that can detect cyberbullying in order to create a healthy online interaction environment [2].

Cyberbullying in Kazakhstan: Statistical and Contextual Overview Cyberbullying is a growing social issue in Kazakhstan, particularly among children and adolescents. Recent studies and official data reveal alarming trends:

- According to the Health Behavior in School-aged Children (HBSC) survey and national reports, approximately 13% of Kazakhstani adolescents – or roughly one in eight – have been victims of cyberbullying [3].

- The prevalence is higher among boys (15%) compared to girls (12%), and 11% of students admitted to participating in cyberbullying themselves [4].

- A UNICEF Kazakhstan (2023) study titled KazKidsOnline found that 21% of children have experienced cyberbullying, while 15% are regularly exposed to disturbing or unwanted online content.

- Official data from the Ministry of Health indicate that 17.5% of Kazakhstani children experience bullying or harassment periodically, and 6.8% report having been threatened or humiliated several times per month [5].

- In 2021, Kazakhstan recorded over 140,000 cyberbullying incidents, nearly double the number from 2020.

These statistics indicate that cyberbullying remains a serious and insufficiently reported problem in Kazakhstan. Many incidents are likely never formally reported due to factors such as fear, social stigma, or limited awareness among victims. As digital communication continues to play a central role in education, social interaction, and information exchange, the creation of reliable AI-driven cyberbullying detection systems adapted to the linguistic and cultural characteristics of the Kazakh language has become both a technological requirement and an important societal responsibility [6]. Conventional methods for identifying cyberbullying have primarily depended on manual moderation and user-reporting mechanisms. While these approaches provide some level of monitoring, they are highly labor-intensive and insufficient for managing the enormous volume of content generated daily on social media platforms. Recent progress in machine learning and deep learning has encouraged researchers to investigate automated detection systems capable of processing large-scale textual datasets efficiently. In particular, DNN models have shown significant potential because they can automatically learn complex linguistic and contextual representations without the need for extensive manual feature engineering [7]. Previous research has demonstrated that Long Short-Term Memory (LSTM) networks are effective in modeling sequential dependencies within text, allowing them to capture contextual relationships that may signal abusive or aggressive communication. Meanwhile, Convolutional Neural Networks (CNNs) have proven successful in extracting local textual features and hierarchical patterns, such as clusters of offensive words or repeated insulting expressions. Based on these developments, the present study proposes a Transformer-based Bi-LSTM architecture designed to detect cyberbullying in Kazakh-language social media content. This model combines the global contextual representation capabilities of Transformer architectures with the bidirectional sequential learning strengths of Bi-LSTM networks. Such a hybrid approach enables the system to capture long-range semantic relationships while also identifying subtle linguistic patterns that arise from the rich morphology and flexible syntactic structure of the Kazakh language [8].

The model performance is evaluated against traditional machine learning classifiers such as Support Vector Machines (SVM), Random Forest, and Decision Trees using evaluation metrics including Accuracy, Precision, Recall, F1-score, and AUC-ROC. Experimental findings demonstrate that the Transformer–Bi-LSTM model significantly outperforms conventional classifiers across all metrics. The ROC curve analyses further confirm the model's high sensitivity and specificity in distinguishing between bullying and non-bullying instances [9].

This study contributes to ongoing efforts in building advanced natural language processing (NLP) tools for low-resource languages like Kazakh and highlights the ethical imperative of applying artificial intelligence responsibly to ensure user protection, fairness, and privacy. Future work will

include expanding the dataset with regional Kazakh dialects and incorporating user feedback mechanisms to enhance real-time adaptability and classification accuracy.

Research on cyberbullying in Kazakhstan has gradually increased in recent years, paralleling global trends, though it remains relatively underdeveloped and fragmented. Early local studies often approached bullying (offline) broadly rather than isolating digital forms. For example, national surveys on bullying among schoolchildren reported that 42.8 % of students identified social networks as the location where bullying occurred most frequently, with about one in four respondents noting academic performance deterioration due to bullying [10].

Over time, researchers have begun focusing specifically on cyberbullying as a distinct phenomenon in Kazakhstan. Several studies estimated that approximately 13 % of Kazakhstani adolescents experience cyberbullying at least monthly, with rates higher among boys (≈15 %) than girls (≈12 %) [11].

Some sources also report that 11 % of adolescents admitted having participated in cyberbullying others – again with gender differences: 14 % for boys and 8 % for girls.

In parallel, local publications have begun exploring psychosocial dynamics surrounding cyberbullying in regional settings. For instance, a study across three regions (Ust-Kamenogorsk, Astana, Atyrau) found that aggressive victim behavior correlates with coping strategies such as close support and helplessness; however, these correlations varied markedly by region, implying that contextual and cultural factors influence how victims respond to cyberbullying [12].

Other authors emphasize the invisible nature of cyberbullying, whereby many children either do not report incidents or notify very few adults. It is reported, for instance, that while 21 % of children said they had observed cyberbullying among peers, only half of them told an adult about it. Some studies analyses claim that among 402 interviewed children, only one disclosed witnessing a cyberbullying incident, illustrating serious underreporting and reluctance to speak publicly about such events [13].

From a legislative and institutional perspective, Kazakhstan has recognized the need for formal legal frameworks. Effective legal regulation of child protection has been examined in comparative studies, showing the existing gaps in national laws and suggesting reforms informed by international models to safeguard minors from harmful online content. To elaborate, as of June 16, 2024, the Code of Administrative Offenses of Kazakhstan was amended to include a new article, 127-2 “Harassment (bullying, cyberbullying) of minors,” stipulating warnings or fines (ten monthly calculation indices) for first offenses and higher penalties (thirty indices) for repeat violations within one year [14].

Despite these legal steps, challenges persist. Many studies note that existing research is often limited to certain cities or school settings, failing to capture rural or dialectal variations. Also, frequency-based self-reports dominate methodology, with little use of advanced computational models or large-scale annotated datasets. Some authors argue that intervention and prevention measures remain weak and under-coordinated across schools, families, and digital platforms [15].

Recent investigations seek to address these gaps. One study examined digital hygiene skills among Kazakhstani teenagers, exploring how such skills may reduce the likelihood of experiencing cyberbullying. Another work discusses cyberbullying in university / student environments, linking victimization and perpetrator ships to stress, anxiety, and adaptation factors in higher education settings. Finally, some authors highlight the ethnocultural dimension – how ethnicity, identity, and perceived discrimination in Kazakhstan’s multicultural communities might influence cyberbullying experiences and responses [16].

In sum, cyberbullying research in Kazakhstan has evolved from broad bullying surveys to more focused inquiries into online harassment, but it still lags in methodological sophistication, scale, and integration with technological solutions. The literature suggests a pressing need for context-sensitive, data-driven, and multidisciplinary approaches that combine legal, psychological, sociocultural, and computational perspectives to comprehensively address cyberbullying in the Kazakh context [17].

Despite the growing number of studies on cyberbullying detection, research focused on the Kazakh language remains limited. Existing approaches are often based on traditional machine learning

methods or do not fully consider the linguistic complexity and morphological richness of Kazakh, which significantly affects classification performance. Furthermore, there is a lack of large-scale annotated datasets and hybrid architectures specifically designed for low-resource languages.

Therefore, the main objective of this study is to develop an effective hybrid deep learning model for automatic cyberbullying detection in Kazakh-language textual data.

To achieve this objective, the following research tasks are defined:

1. To collect and preprocess a dataset of Kazakh-language social media texts;
2. To design a hybrid Transformer–BiLSTM–CNN architecture;
3. To evaluate the model performance using standard classification metrics;
4. To compare the proposed approach with classical machine learning and baseline deep learning models.

The expected outcome of this study is the development of a robust and scalable cyberbullying detection system capable of capturing both contextual and semantic features in Kazakh text, contributing to the advancement of NLP solutions for low-resource languages.

Research methodology

The selection of methods in this study is guided by the linguistic characteristics of the Kazakh language and the limitations of existing approaches for cyberbullying detection in low-resource settings.

Traditional feature extraction techniques such as TF-IDF and Bag-of-Words were employed as baseline representations due to their computational efficiency and interpretability. These methods allow for an initial comparison with classical machine learning algorithms and provide a reference point for evaluating more advanced models.

However, given the complex morphology and flexible word order of the Kazakh language, traditional approaches are insufficient for capturing deep semantic and contextual relationships. Therefore, deep learning architectures were incorporated.

The Bi-LSTM model was selected for its ability to capture bidirectional sequential dependencies in textual data, which is critical for understanding context in natural language. The Transformer architecture was introduced to model global contextual relationships using self-attention mechanisms, enabling the system to identify long-range dependencies within sentences.

Additionally, Convolutional Neural Networks (CNN) were included to extract local n-gram features and detect repetitive or phrase-level patterns commonly associated with cyberbullying expressions.

Thus, the proposed hybrid architecture combines:

- sequential modeling (Bi-LSTM),
- global context understanding (Transformer),
- local feature extraction (CNN),

which together provide a comprehensive representation of textual data and improve classification performance.

In this section, we present the systematic methodology adopted to evaluate the performance of the proposed Transformer based BI-LSTM deep learning model for detecting cyberbullying in social Kazakh language texts. The section is organized to detail the procedures for data preprocessing techniques used to prepare textual data for modeling, the architectural composition of the Transformer based BI-LSTM deep learning model, and the evaluation metrics applied for comparison with conventional machine learning algorithms. Furthermore, the experimental design is discussed, specifying the utilized hardware and software environments to ensure methodological transparency and reproducibility. Through this thorough exposition of our research workflow, we aim to maintain clarity and allow fellow scholars to replicate or expand upon this work. The comprehensive overview provided here functions both as a methodological reference for analogous studies and as a groundwork for future advancements in automated cyberbullying detection and digital content moderation systems.

The key research problem addressed in this study lies in the limited effectiveness of existing cyberbullying detection methods when applied to Kazakh-language data. Traditional approaches fail to capture complex linguistic patterns, while many deep learning models are not adapted to low-resource languages. This creates a gap between the need for accurate automated moderation systems and the available technological solutions.

The research process was conducted in several structured stages:

Stage 1: Data Collection

Textual data were collected from online social media platforms, including Reddit and Instagram, focusing on Kazakh-language content relevant to user interactions.

Stage 2: Data Preprocessing

The collected data were cleaned and standardized. This included removing noise, correcting typographical errors, tokenization, and normalization. Irrelevant symbols and duplicate entries were also eliminated.

Stage 3: Feature Representation

Textual data were transformed into numerical representations using TF-IDF, Bag-of-Words, and word embedding techniques to enable processing by machine learning and deep learning models.

Stage 4: Model Development

Several models were implemented, including traditional machine learning classifiers (SVM, Random Forest, Decision Trees) and deep learning architectures (LSTM, Bi-LSTM, Transformer-based hybrid model).

Stage 5: Model Training

The models were trained using labeled datasets. Hyperparameters such as learning rate, batch size, and number of epochs were optimized to achieve stable performance.

Stage 6: Evaluation

The performance of all models was evaluated using standard metrics including Accuracy, Precision, Recall, F1-score, and AUC-ROC.

Stage 7: Comparative Analysis

The results of the proposed model were compared with baseline models to assess performance improvements and validate the effectiveness of the hybrid architecture.

Data preparation

Data Preprocessing

The first stage of the study involves preprocessing textual data collected from online platforms such as Reddit. This step is important because raw data from social media often contains noise, inconsistent formatting, and various irrelevant elements. During preprocessing, unnecessary information is removed, typographical errors are corrected, and the text is standardized to ensure uniform formatting. These procedures help improve the quality and consistency of the dataset, which is essential for effective model training and reliable performance. In this research, several common text preprocessing and feature representation techniques were applied. These include Term Frequency–Inverse Document Frequency (TF-IDF), the Bag-of-Words (BoW) approach, and word embedding methods. These techniques transform textual data into numerical representations that can be effectively processed by machine learning and deep learning models.

Feature Extraction

After the preprocessing phase, the system proceeds to extract meaningful features from the textual data. This is achieved using a combination of specialized techniques designed to capture the most relevant linguistic and contextual attributes.

- TF-IDF (Term Frequency-Inverse Document Frequency): This technique evaluates how important a word is within a particular document in relation to the entire dataset. Unlike simple frequency-based methods, TF-IDF considers how often a term appears across multiple documents,

which helps reduce the influence of very common words that provide little meaningful information. As a result, it highlights terms that are more distinctive for a specific document [18–20].

- Bag of Words (BoW): The Bag-of-Words method represents text as a numerical vector based on the frequency of words appearing in a document. It ignores grammar and word order, focusing only on the occurrence of terms. This approach simplifies textual data and converts it into a structured format that can be processed by machine learning models [21–23].
- Statistical Features: In addition to textual representations, various statistical characteristics of the text can also be extracted. These may include metrics such as word count, sentence length, punctuation frequency, and other structural properties of the text. Such features can sometimes provide additional signals that help identify patterns associated with cyberbullying behavior [24,25].

Word Embedding

This stage is important because it converts words into continuous numerical vectors that capture semantic relationships between them. By representing words in this way, the model can better understand the contextual meaning within the text. Such representations allow the system to process language more effectively and improve its performance in later tasks, including text classification.

Machine Learning Approaches

The system applies both traditional machine learning algorithms and modern deep learning models to perform text classification. The traditional machine learning component includes widely used algorithms such as Support Vector Machines, Random Forest, Decision Trees, K-Nearest Neighbors, Naïve Bayes, and Logistic Regression. These methods are commonly used as baseline models in text classification tasks because they are computationally efficient, relatively easy to interpret, and suitable for comparing the performance of more complex models.

Deep Learning Classifiers

In addition to conventional algorithms, the system incorporates deep learning classifiers to leverage their superior capability in modeling complex linguistic structures and contextual dependencies present in Kazakh-language text. Unlike traditional approaches that rely heavily on manual feature engineering, deep learning models automatically extract hierarchical and semantic representations from raw text data, enabling more accurate and nuanced classification.

- LSTM (Long Short-Term Memory): Well-suited for sequences such as sentences, LSTMs can capture temporal dependencies and context within the text.
- BI-LSTM (Convolutional Neural Network): Originally designed for image processing, CNNs have been adapted for NLP to detect patterns in text.
- Transformer + Bi-LSTM: Combining LSTM and CNN to harness both temporal context and local textual features for improved cyberbullying detection.

Proposed Model Architecture

The architecture of the proposed Transformer based BI-LSTM model is specifically designed for text classification in Kazakh language. This model effectively integrates the capabilities of Biderctional Long Short-Term Memory (Bi-LSTM) and Transformers architecture (fig.1).

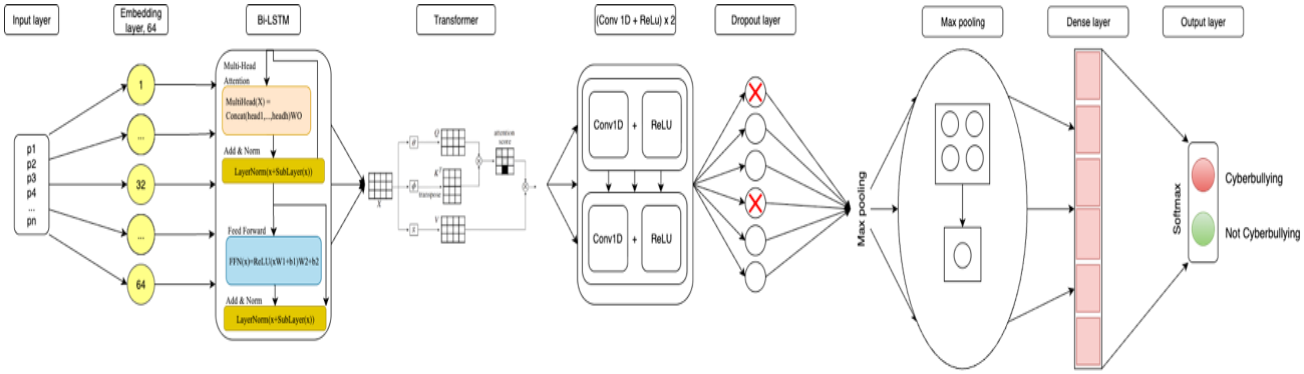

Figure 1. Architecture of the proposed model

1. Input Layer The process begins with the Input Layer, which receives preprocessed textual sequences from social media platforms such as Reddit or Instagram. Each input sequence is represented as a series of tokens (1, $p2$, $p3$, … , p_n), where n denotes the number of words or tokens in the sentence. These tokens are indexed and passed to the embedding layer for numerical transformation.

2. Embedding Layer The Embedding Layer converts the input tokens into dense, continuous vector representations of size 64 dimensions. This layer maps semantically similar words to nearby points in the embedding space, allowing the model to capture underlying linguistic and contextual relationships. The resulting embeddings form a numerical matrix that serves as the foundational input for subsequent sequential and attention-based computations.

3. Bi-LSTM Layer Following embedding, the representations are fed into a Bidirectional Long Short-Term Memory (Bi-LSTM) layer. This component processes the sequence in both forward and backward directions, enabling the model to retain information from past and future contexts simultaneously. This dual traversal enhances the model's understanding of word dependencies and sentence-level semantics, which are particularly important for identifying implicit or context-driven forms of cyberbullying (e.g., sarcasm, coded aggression). Mathematically, the Bi-LSTM updates its cell and hidden states as follows:

$$h_t = [\, h_t\!\rightarrow ; h_t \leftarrow] \tag{1}$$

where $h_t\!\rightarrow$ and $h_t\!\leftarrow$ represent forward and backward hidden states respectively. The Bi-LSTM output sequence is then normalized through Layer Normalization, improving convergence and stabilizing gradient flow.

4. Transformer Layer Next, the model integrates a Transformer block, which refines contextual understanding using a multi-head self-attention mechanism. Each attention head computes relationships between all token pairs within a sequence to identify contextually relevant connections. The operation can be expressed as:

$$\text{MultiHead}\,(X) = \text{Concat}\,(\,\text{head}_1\,, \ldots, \text{head}_h)*W^O \tag{2}$$

This is followed by:

- Layer Normalization: Ensures numerical stability and accelerates training. Feed-Forward Network
- (FFN): A two-layer perceptron applying nonlinear transformation via ReLU activation:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{3}$$

- Residual Connections: Added around both the attention and FFN sub-layers to prevent information loss and facilitate deeper gradient propagation. Through these operations, the Transformer captures global dependencies that the Bi-LSTM alone may overlook, such as long-range interactions or subtle shifts in sentiment across distant words.

5. Convolutional Feature Extraction Layer The Transformer output is subsequently processed by two 1D Convolutional Layers (Conv1D), each followed by a ReLU activation. These layers extract local patterns and n-gram level features within the sequence representation. The convolutional filters act as detectors for short, phrase-level indicators of aggressive or harmful content, complementing the contextual comprehension gained from the Bi-LSTM and Transformer layers.

$$\text{Conv1D}\,(x) = \text{ReLU}(W*x + b) \tag{4}$$

The notation (Conv1D + ReLU) × 2 signifies that two such convolutional blocks are stacked to enhance feature richness and hierarchical representation.

6. Dropout Layer. To prevent overfitting, the model incorporates a Dropout Layer, where a portion of neurons is randomly deactivated during training. This stochastic regularization technique ensures that the model generalizes well to unseen data by preventing co-adaptation of features and promoting robustness.

7. Max Pooling Layer A Max Pooling Layer follows to down sample the feature maps generated by convolution. This operation retains the most prominent activations (i.e., the most significant features) from each local region, effectively reducing dimensionality while preserving the most informative patterns. The pooling operation aids in achieving translational invariance and computational efficiency.

8. Dense Layer The output of the pooling layer is flattened and passed to a Fully Connected (Dense) Layer. This layer aggregates and combines all extracted features, learning complex non-linear decision boundaries. The dense neurons integrate both the contextual and spatial features, acting as the final feature synthesis stage before classification.

9. Output Layer Finally, the Output Layer produces the classification result. A Softmax activation function is used for multi-class classification or Sigmoid for binary classification (cyberbullying vs. non-cyberbullying). The layer outputs probabilities representing the likelihood that a given text instance belongs to the “cyberbullying” or “non-cyberbullying” category.

10. Overall Model Summary The proposed Transformer–BiLSTM hybrid model seamlessly integrates:

- Sequential understanding (via Bi-LSTM),
- Global contextual awareness (via Transformer attention),
- Local feature extraction (via Conv1D),
- Efficient classification (via Dense and Output layers).

Results of the study

This hybrid design allows the model to handle both explicit and implicit forms of cyberbullying in text, capturing subtle linguistic cues specific to the Kazakh language. The architecture is optimized for high interpretability, scalability, and generalization across different social media datasets.

Evaluation

To assess the effectiveness of the proposed classification model, a comprehensive evaluation framework was employed using several standard performance metrics commonly adopted in text classification and machine learning research. These metrics provide quantitative insights into different aspects of the model’s predictive capabilities, allowing for a balanced and objective comparison with baseline classifiers. The selected evaluation measures includes Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Accuracy

Accuracy shows how many predictions the model made correctly compared to the total number of samples in the dataset. It is a simple way to evaluate the overall performance of a model. However, accuracy can sometimes be misleading, especially when the dataset is imbalanced. In such cases, the model may achieve a high accuracy simply by predicting the majority class more often, even if it performs poorly on the minority class.

$$Accuracy = \frac{TP + TN}{\mathrm{TP + TN + FP + FN}} \tag{5}$$

Precision

Precision measures how many of the instances predicted as positive by the model are actually correct. In other words, it shows the proportion of correctly identified positive cases among all predicted positive cases. This metric is important because it reflects how well the model avoids false positives. In tasks such as cyberbullying detection, high precision is especially important, since incorrectly labeling neutral or harmless content as offensive may lead to unfair moderation decisions.

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}} \tag{6}$$

Recall

Recall measures how many of the actual positive cases are correctly identified by the model. In other words, it shows how well the model can detect instances of cyberbullying. A high recall means that most bullying cases are successfully detected. However, in some situations, increasing recall may also lead to more false positives being included in the predictions.

$$Recall = \frac{TP}{\text{TP} + \text{FN}} \tag{7}$$

F1-Score

F1-score is a metric that combines precision and recall into a single value by calculating their harmonic mean. It is useful when both false positives and false negatives need to be considered. This metric is especially helpful when working with imbalanced datasets because it evaluates how well the model maintains a balance between precision and recall.

$$Accuracy = 2 * \frac{Precision * Recall}{\text{Precision} + \text{Recall}} \tag{8}$$

Finally, the ROC-AUC metric evaluates the model's discriminative ability across various classification thresholds. The ROC curve plots the True Positive Rate against the False Positive Rate, and the AUC value represents the probability that the classifier ranks a randomly chosen positive instance higher than a negative one. The closer the area under curve to 1 the more effective algorithm is. So it means our aims is to get closer to as much as we can.

To establish a baseline for cyberbullying detection performance, several conventional machine learning algorithms were implemented and evaluated on the same dataset prior to the deployment of the proposed Transformer–BiLSTM model. Figure 2 illustrates the comparative performance of six widely used classifiers - K-Nearest Neighbors, Random Forest, Logistic Regression, Support Vector Machine, Naive Bayes, and Decision Tree – across four key evaluation metrics: Accuracy, Precision, Recall, and F1-Score.

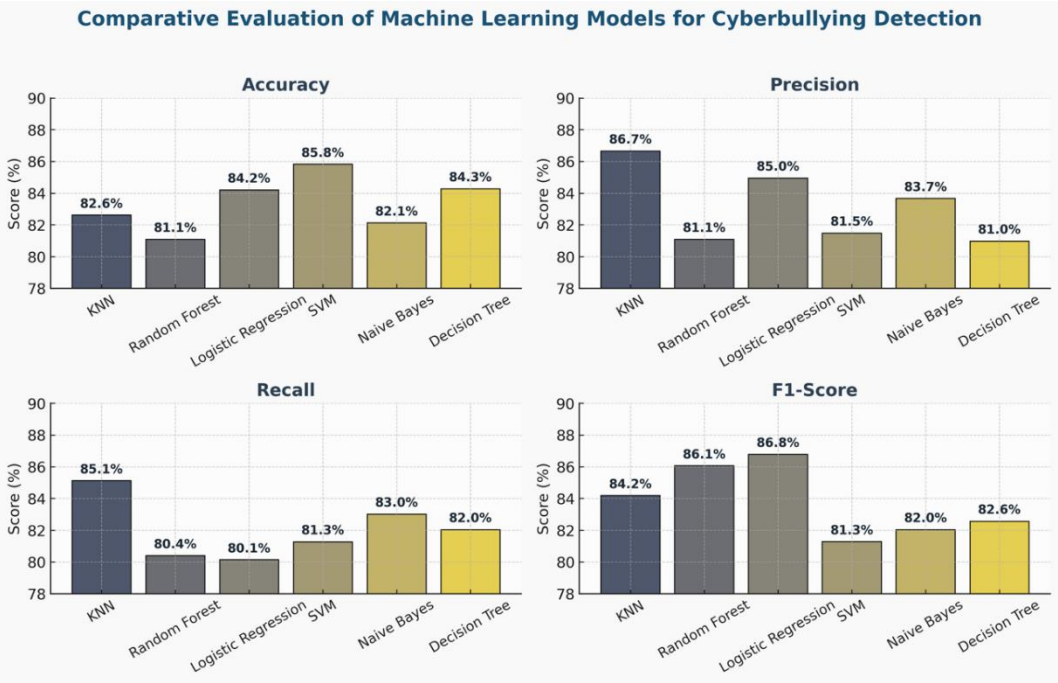

Figure 2. ML algorithms comparative results

These metrics combined provide a comprehensive view of each model's predictive capability, covering general correctness, sensitivity to true positives, ability to avoid false alarms, and overall balance between detection accuracy and reliability. As shown in the figure, the performance of all classifiers ranges between 80% and 87%, indicating solid effectiveness in identifying cyberbullying-related content and establishing a strong baseline for comparison with advanced deep learning approaches.

Although traditional machine learning algorithms show stable performance in cyberbullying detection, with overall accuracy typically ranging from 80% to 87%, these results are still not sufficient to fully achieve the goals of this study. Classical classifiers depend largely on manually designed features and often struggle to capture the complex linguistic patterns, contextual relationships, and subtle semantic nuances that appear in social media communication. This limitation becomes even more pronounced when working with morphologically rich languages such as Kazakh, where meaning can vary significantly depending on word forms and context.

To address these limitations, deep learning architectures are considered a more advanced and effective approach. Unlike traditional models, deep neural networks are capable of automatically learning hierarchical representations from raw text data. This allows them to capture both local patterns and broader contextual relationships that reflect sentiment, intent, and subtle nuances in communication. In particular, transformer-based architectures have demonstrated strong performance in understanding contextual semantics and identifying even subtle signals of harmful or aggressive language.

Therefore, the next section presents a Transformer-BiLSTM hybrid model designed to combine the strengths of both approaches. The model leverages the bidirectional contextual learning capability of LSTM networks together with the global attention mechanisms provided by Transformer architectures. In addition, simple LSTM and Bi-LSTM models were included in the experiments to provide a baseline and to demonstrate the superiority of the proposed Transformer + Bi-LSTM architecture. A comparison of the obtained results is illustrated in Figure 3 below.

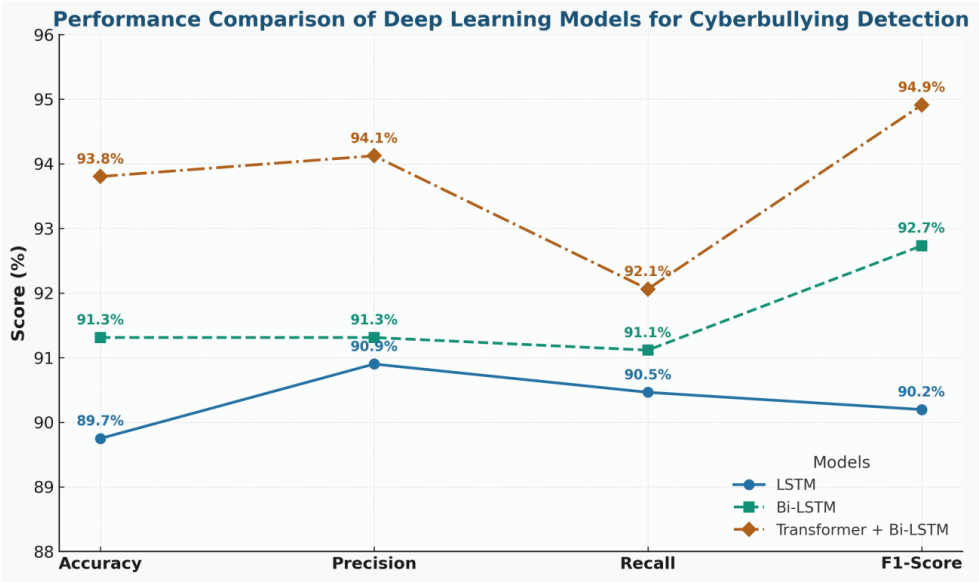

Figure 3. Deep Learning approaches comparative results.

Table 1 presents a comparative evaluation of seven approaches for the cyberbullying classification task using five commonly used metrics: Accuracy, Precision, Recall, F-score, and AUC-ROC. The proposed model demonstrates the best performance across all evaluation criteria (Accuracy = 0.938, Precision = 0.941, Recall = 0.921, F-score = 0.949, AUC-ROC = 0.939), indicating a clear improvement over the classical machine learning baselines, including RF, DT, KNN, NB, LR, and SVM.

Among the baseline methods, Random Forest and SVM show the strongest overall performance, particularly in terms of accuracy and AUC-ROC, while KNN achieves relatively balanced precision and recall values. In contrast, Decision Tree and Naive Bayes perform weaker across most metrics, which may be explained by their sensitivity to feature distribution assumptions and the effects of class imbalance in the dataset.

Table 1. Comparison of the proposed model with classical machine learning methods

Approaches	*Accuracy*	*Precision*	*Recall*	*F-score*	*AUC-ROC*
Proposed model	***0.938***	***0.941***	***0.921***	***0.949***	***0.939***
RF	*0.811*	*0.811*	*0.804*	*0.861*	*0.867*
DT	*0.843*	*0.813*	*0.82*	*0.826*	*0.811*
KNN	*0.826*	*0.867*	*0.851*	*0.842*	*0.826*
NB	*0.821*	*0.813*	*0.83*	*0.82*	*0.811*
LR	*0.842*	*0.821*	*0.801*	*0.868*	*0.851*
SVM	*0.858*	*0.85*	*0.813*	*0.813*	*0.842*

The experimental results clearly demonstrate the effectiveness of the proposed hybrid Transformer–BiLSTM–CNN model in detecting cyberbullying in Kazakh-language texts. As shown in Table 1, the proposed model achieves the highest performance across all evaluation metrics, outperforming both traditional machine learning algorithms and simpler deep learning models.

In particular, the model achieves an accuracy of 93.8%, indicating a high overall classification correctness. The precision score of 94.1% reflects the model's ability to minimize false positives, which is critical in avoiding incorrect labeling of non-offensive content. The recall value of 92.1% confirms that the model successfully detects most real cyberbullying instances, while the F1-score of 94.9% demonstrates a strong balance between precision and recall.

The superiority of the proposed model can be attributed to its hybrid architecture. The Transformer component captures global contextual dependencies across the entire text, allowing the model to understand long-range semantic relationships. The Bi-LSTM layer enhances this capability by modeling bidirectional sequential information, which is particularly important for detecting implicit or context-dependent forms of cyberbullying.

Additionally, the CNN layers contribute by extracting local features such as offensive phrases and repetitive patterns commonly found in abusive language. This combination enables the model to simultaneously capture global, sequential, and local features, which explains its improved performance compared to baseline methods.

Discussion

The results of this study demonstrate notable differences in the effectiveness of various machine learning approaches for detecting cyberbullying in Kazakh-language social media content. The proposed Transformer–BiLSTM hybrid model shows clear superiority compared to traditional machine learning methods, achieving high scores across all evaluation metrics, including accuracy, precision, recall, F-score, and AUC-ROC. This strong performance can largely be attributed to the model's ability to capture both contextual and semantic relationships within the Kazakh language, which is morphologically rich and highly dependent on context.

Traditional algorithms like SVM, Random Forest, and Decision Trees, while effective for many general classification tasks, show limitations when applied to natural language processing. These methods depend heavily on manually engineered features and often fail to capture syntactic variation, flexible word order, and the morphological richness of the language—factors that are crucial for correctly interpreting abusive or harmful expressions in Kazakh text. Although SVM achieved slightly stronger results among the classical approaches, its performance still remained below that of

deep learning models. This outcome further emphasizes the need for architectures that can automatically learn long-term dependencies and contextual nuances from textual data.

The integration of feature extraction techniques such as TF-IDF, Bag of Words, and statistical features improved the baseline performance of traditional classifiers but remained insufficient for handling the subtleties of real online communication in Kazakh. This proves that, although feature engineering can enhance model performance, deep learning architectures inherently outperform due to their ability to autonomously capture both local and global linguistic patterns without manual intervention.

From a practical perspective, this study highlights the importance of incorporating advanced models into social media monitoring systems in Kazakhstan. Implementing accurate cyberbullying detection tools can help platforms identify and address harmful behavior more efficiently, contributing to the creation of safer and more respectful online environments. At the same time, the deployment of such systems must adhere to key ethical principles, including data privacy, fairness, and transparency. The inclusion of user feedback mechanisms and efforts to improve the interpretability of AI decisions are essential for maintaining public trust while ensuring a balanced approach between content moderation and freedom of expression.

Conclusion

In conclusion, this study supports the use of Transformer-based Bi-LSTM deep learning models for the automatic detection of cyberbullying in Kazakh-language text. The results show that these architectures outperform traditional methods and are particularly well suited for handling the linguistic complexity and cultural characteristics of the Kazakh language. Future work may focus on expanding annotated Kazakh datasets, integrating multilingual embeddings, and adapting models to regional dialects and data from different social media platforms. Such developments would improve the generalizability of the models and contribute to the broader advancement of ethical AI-based content moderation systems.

From a practical perspective, the proposed model can be integrated into social media moderation systems to automatically detect harmful content. This can help reduce the spread of cyberbullying, improve user safety, and support the development of responsible AI-driven monitoring tools.

Overall, obtained results confirm that the proposed hybrid model effectively addresses the limitations of traditional approaches and achieves superior performance in cyberbullying detection tasks. These findings validate the research hypothesis and support the feasibility of applying advanced deep learning techniques in low-resource language contexts.

Acknowledgements

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.

References

[1] Rogayah M. Al-Ibrahim, Mostafa Z. Ali, Hassan M. Najadat (2023). Detection of Hateful Social Media Content for Arabic Language. ACM Transactions on Asian and Low-Resource Language Information Processing. https://doi.org/10.1145/3592792

[2] Saha, S. K., Mim, A. A., Akter, S., Hosen, M. M., Shihab, A. H., & Mehedi, M. H. K. (2024, May). BengaliHateCB: A Hybrid Deep Learning Model to Identify Bengali Hate Speech Detection from Online Platform. In 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT) (pp. 439-444). IEEE. https://doi.org/10.1109/iceeict62016.2024.10534319

[3] Al-Khasawneh, M. A., Faheem, M., Alarood, A. A., Habibullah, S., & Alsolami, E. (2024). Towards Multi-Modal Approach for Identification and Detection of Cyberbullying in Social Networks. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3420131

[4] Ali, M., Hassan, M., Kifayat, K., Kim, J. Y., Hakak, S., & Khan, M. K. (2023). Social media content classification and community detection using deep learning and graph analytics. Technological Forecasting and Social Change, 188, 122252. https://doi.org/10.1016/j.techfore.2022.122252

[5] Machová, K., Mach, M., & Porezaný, M. (2022). Deep Learning in the Detection of Disinformation about COVID-19 in Online Space. Sensors, 22(23), 9319. https://doi.org/10.3390/s22239319

[6] Ajao, O., Bhowmik, D., & Zargari, S. (2018, July). Fake news identification on Twitter with hybrid CNN and RNN models. In Proceedings of the 9th International Conference on Social Media and Society (pp. 226–230). https://doi.org/10.1145/3217804.3217917

[7] Aggarwal, P., & Mahajan, R. (2024). Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification. Journal of Information Systems and Informatics, 6(2), 607–623. https://doi.org/10.51519/journalisi.v6i2.692

[8] Bilal, M., Khan, A., Jan, S., & Musa, S. (2022). Context-Aware Deep Learning Model for Detection of Roman Urdu Hate Speech on Social Media Platform. IEEE Access, 10, 121133–121151. https://doi.org/10.1109/access.2022.3216375

[9] Neog, M., & Baruah, N. (2024). A Hybrid Deep Learning Approach for Assamese Toxic Comment Detection in Social Media. Procedia Computer Science, 235, 2297–2306. https://doi.org/10.1016/j.procs.2024.04.218

[10] Al-Wesabi, F. N., Obayya, M., Alabdan, R., Aljehane, N. O., Alazwari, S., Alruwaili, F. F., ... & Swathi, A. (2024). Automatic Recognition of Cyberbullying in the Web of Things and Social Media Using Deep Learning Framework. IEEE Transactions on Big Data. https://doi.org/10.1109/TBDATA.2024.3409939

[11] Singh, J. P., Kumar, A., Rana, N. P., & Dwivedi, Y. K. (2020). Attention-Based LSTM Network for Rumor Veracity Estimation of Tweets. Information Systems Frontiers, 1–16. https://doi.org/10.1007/s10796-020-10040-5

[12] Daraghmi, E. Y., Qadan, S., Daraghmi, Y., Yussuf, R., Cheikhrouhou, O., & Baz, M. (2024). From Text to Insight: An Integrated CNN-BiLSTM-GRU Model for Arabic Cyberbullying Detection. IEEE Access. https://doi.org/10.1109/access.2024.3431939

[13] Husain, F., & Uzuner, O. (2021). A Survey of Offensive Language Detection for the Arabic Language. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20(1), 1–44. https://doi.org/10.1145/3421504

[14] Ahmad, S., Asghar, M. Z., Alotaibi, F. M., & Awan, I. (2019). Detection and Classification of Social Media-Based Extremist Affiliations Using Sentiment Analysis Techniques. Human-Centric Computing and Information Sciences, 9, 1–23. https://doi.org/10.1186/s13673-019-0185-6

[15] Badawi, S. (2024). Deep Learning-Based Cyberbullying Detection in Kurdish Language. The Computer Journal, https://doi.org/10.1093/comjnl/bxae024

[16] Govers, J., Feldman, P., Dant, A., & Patros, P. (2023). Down the Rabbit Hole: Detecting Online Cyberbullying, Radicalisation, and Politicised Hate Speech. ACM Computing Surveys. https://doi.org/10.1145/3583067

[17] Babu, N. V., & Kanaga, E. G. M. (2022). Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review. SN Computer Science, 3, 1–20. https://doi.org/10.1007/s42979-021-00958-1

[18] Yadav, D., Gupta, A., Asati, S., Choudhary, N., & Yadav, A. K. (2020, December). Age Group Prediction on Textual Data Using Sentiment Analysis. In 9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion (pp. 61–65). https://doi.org/10.1145/3439231.3439262

[19] Azzi, S. A., & Zribi, C. B. O. (2021, June). From Machine Learning to Deep Learning for Detecting Abusive Messages in Arabic Social Media: Survey and Challenges. In Intelligent Systems Design and Applications: 20th International Conference on ISDA 2020 (pp. 411–424). Springer International Publishing. https://doi.org/10.1007/978-3-030-71187-0_38

[20] Khan, S., Fazil, M., Sejwal, V. K., Alshara, M. A., Alotaibi, R. M., Kamal, A., & Baig, A. R. (2022). BiCHAT: BiLSTM with Deep CNN and Hierarchical Attention for Hate Speech Detection. Journal of King

Saud University-Computer and Information Sciences, 34(7), 4335–4344. https://doi.org/10.1016/j.jksuci.2022.05.006

[21] Asghar, M. Z., Habib, A., Habib, A., Khan, A., Ali, R., & Khattak, A. (2021). Exploring Deep Neural Networks for Rumor Detection. Journal of Ambient Intelligence and Humanized Computing, 12, 4315–4333. https://doi.org/10.1007/s12652-019-01527-4

[22] Ullah, F., Ullah, S., Srivastava, G., & Lin, J. C. W. (2024). IDS-INT: Intrusion Detection System Using Transformer-Based Transfer Learning for Imbalanced Network Traffic. Digital Communications and Networks. https://doi.org/10.1016/j.dcan.2023.03.008

[23] Singh, N. M., & Sharma, S. K. (2023). An Efficient Automated Multi-Modal Cyberbullying Detection Using Decision Fusion Classifier on Social Media Platforms. Multimedia Tools and Applications, 83(7), 20507–20535. https://doi.org/10.1007/s11042-023-16402-w

[24] Ghosal, S., & Jain, A. (2023). HateCircle and Unsupervised Hate Speech Detection Incorporating Emotion and Contextual Semantics. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4), 1–28. https://doi.org/10.1145/3576913

[25] Ellaky, Z., Benabbou, F., Matrane, Y., & Qaqa, S. (2024). A Hybrid Deep Learning Architecture for Social Media Bots Detection Based on BiGRU-LSTM and GloVe Word Embedding. IEEE Access. https://doi.org/10.1109/access.2024.3430859

[26] Aliyeva, Ç. O., & Yağanoğlu, M. (2024). Deep Learning Approach to Detect Cyberbullying on Twitter. Multimedia Tools and Applications, 1–24. https://doi.org/10.1007/s11042-024-19869-3