DOI: 10.37943/22PSRO3633

Batyrkhan Omarov
PhD in Information System, Associate Professor, Department of Information System
batykhan@gmail.com; orcid.org/0000-0002-8341-7113
International Information Technology University, Kazakhstan

Rustam Abdrakhmanov
Candidate of Technical Science, Associate Professor, Department of Information Technologies
abdrakhmanov.rustam@iuth.edu.kz; orcid.org/0000-0002-5508-389X
International University of Tourism and Hospitality, Kazakhstan

Aigerim Toktarova
Master, Senior lecturer, Department of Computer Engineering
toktar.aigerim@list.ru; orcid.org/0000-0002-6265-9236
Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan

GLOVE-EMBEDDED ATTENTION BILSTM NETWORKS FOR ENHANCED MULTICLASSIFICATION OF TWEETS IN CYBERBULLYING DETECTION ON ONLINE CONTENT

Abstract: This paper offers a neural network method for social media cyberbullying detection and classification. The model uses GloVe-embedded BiLSTM networks with self-attention to recognize language and semantic patterns. The research uses advanced machine learning methods to fight cyberbullying and suggests ways to improve cyberbullying detection systems' precision and ethics. The proposed paradigm addresses several cyberbullying levels and forms, enabling targeted interventions and victim support.

GloVe implementations do semantic processing, BiLSTM networks sequentially learn, and self-attention mechanisms focus contextual analysis in the model. Word clouds show the abundance and relevance of phrases across several cyberbullying categories, revealing common themes and vocabulary. Tweet lengths, confusion matrix, training and validation loss and accuracy metrics, and ROC curves included in the dataset. The logistic regression model's ROC curve investigation shows substantial classification performance across multiple categories with AUC values between 0.905 and 0.997. The best model for age categorization has an AUC of 0.997, followed by religion (0.996) and ethnicity (0.993). Gender classification has an AUC of 0.979, whereas cyberbullying and non-cyberbullying have 0.921 and 0.905, respectively. The logistic regression model's ROC curve investigation shows substantial classification performance across multiple categories with AUC values between 0.905 and 0.997. The best model for age categorization has an AUC of 0.997, followed by religion (0.996) and ethnicity (0.993). Gender classification has an AUC of 0.979, whereas cyberbullying and non-cyberbullying have 0.921 and 0.905, respectively.

The study encourages AI technology for social good and emphasizes the need to improve categorization algorithms to handle cyberbullying language's complex changes. Expanding training datasets, exploring hybrid modeling methodologies, and creating AI application ethics must be future goals.

Keywords: Cyberbullying Detection; Deep Learning; Natural Language Processing; GloVe Embeddings; BiLSTM Networks; Self-Attention Mechanisms; Social Media.

Received: 15.12.2024 Accepted: 25.06.2025 Published: 30.06.2025

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

Introduction

The rapidly changing social media landscape has created the harmful issue of cyberbullying, which has become a significant public concern affecting people across multiple platforms. Cyberbullying, characterized by the use of online communication to intimidate or intimidate an individual, has shown significant psychological consequences for victims [1], [2]. The pervasiveness of social media magnifies the impact of bullying behaviors, necessitating the development of effective detection methods to reduce harm and ensure safe online settings.

The complexity of language, which includes slang, coded messages, and context-sensitive terms, poses significant challenges to automated cyberbullying detection [3]. Traditional approaches to cyberbullying detection often rely on keyword-based filtering and basic machine learning models that fail to understand the complex semantics of cyberbullying content [4]. Furthermore, the fluidity of language on social media, characterized by the continuous emergence of new slang and idioms, complicates the detection process, requiring models that can adapt and learn to changing patterns.

Recent advances in natural language processing (NLP) have facilitated the development of more sophisticated methodologies. A significant advance in this area is the use of bi-directional long short-term memory (BiLSTM) networks, which can understand context by examining the sequence of previous and subsequent states [5]. This property is particularly useful for understanding complex sentences, which are often used in cyberbullying.

The introduction of word embedding techniques such as global vectors for word representation (GloVe) significantly increases the complexity of NLP models. GloVe embeddings encapsulate both the syntactic characteristics of words and their semantic relationships, which are important for inferring the intent behind messages. These embeddings have been used effectively in many areas of sentiment analysis, demonstrating their ability to understand contextual meanings [6], [7].

However, while BiLSTM and GloVe embeddings significantly improve analytical capabilities, they do not inherently prioritize the most relevant segments of text data, which may be important for detecting nuanced incidents of cyberbullying. The attention mechanism alleviates this problem by allowing the model to focus on text segments that are more relevant to a particular task. This process has revolutionized the field of machine learning by improving the interpretability and performance of models in tasks that require understanding complex relationships in data [8].

This paper presents a novel strategy that combines BiLSTM networks embedded in GloVe with an attention mechanism for multi-classification of tweets, taking into account the limitations and capabilities of existing technology. This methodology aims to improve the detection of cyberbullying across different categories, including harassment, threats, and often nuanced and context-sensitive hate speech. The implementation of GloVe embeddings facilitates deep semantic understanding of the text, while the attention mechanism allows the model to highlight important elements of the textual material [9].

In addition, the multi-classification functionality of the proposed model addresses another important aspect of cyberbullying detection: the different degrees and manifestations of cyberbullying. By differentiating between different forms of cyberbullying, this approach facilitates the implementation of more targeted interventions, thereby enhancing support and protection for victims. This aspect of the research is important because it is better adapted to different cyberbullying situations and responds to the demand for effective responses [10].

This study is initially organized to demonstrate the theoretical framework and evolution of the proposed model, and then provides a comprehensive account of the methodology used to implement and evaluate it. The following sections present the results, analyzing the effectiveness of the model in detecting different types of cyberbullying. The paper concludes with a

DOI: 10.37943/22PSRO3633

critical review of the conclusions, implications for future research, and potential implementation of the model in a practical context.

This study significantly advances the field of cyberbullying detection by addressing shortcomings in existing strategies and applying advanced NLP techniques. It provides a comprehensive framework that improves detection accuracy and contributes to a broader initiative to eliminate cyberbullying on digital platforms.

The introduction investigates sophisticated NLP methodologies for cyberbullying detection, using BiLSTM networks, GloVe embeddings, and an attention mechanism to improve contextual comprehension. Enabling multi-classification of tweets enhances the precision in detecting harassment, threats, and hate speech, hence facilitating more effective interventions and promoting safer online environments across digital platforms.

Literature Review

The proliferation of online communications has led to the rise of cyberbullying, a modern form of harassment that has a significant impact on individuals and communities. As digital platforms evolve, so do the methods of abuse, requiring continuous improvements in detection and response tools. This research review examines the evolution and incorporation of machine learning methodologies in cyberbullying detection, with a particular focus on BiLSTM networks, GloVe implementations, and attention mechanisms.

A. Cyberbullying on Digital Platforms.

Cyberbullying is a multifaceted form of hostility that uses digital technology to inflict harm, often anonymously. Broadly speaking, it encompasses behaviors that include harassment, discrimination, impersonation, stalking, and cybertalking, each of which can vary in their presentation and severity. The anonymity offered by digital platforms can amplify the occurrence and intensity of these attacks, making conventional surveillance and policing approaches inadequate.

The consequences of cyberbullying are severe, including psychological distress, serious mental health disorders, and suicidal ideation. Addressing these concerns is essential, as research has linked cyberbullying to negative outcomes such as distress, anxiety, and decreased self-esteem. This sets an important context for the development of automated detection systems that can provide rapid intervention.

B. Machine learning for cyberbullying detection.

Machine learning offers potential methods for detecting cyberbullying. Early approaches used conventional classifiers such as Support Vector Machines (SVM) and Random Forests, which focused primarily on superficial text attributes such as keywords and grammatical structures. However, these models often struggle to understand the complexity of the language used in cyberbullying, especially contextual and semantic aspects that are not captured by simple keyword-based methods [11].

C. Progress through deep learning.

The advent of deep learning has brought about significant changes in cyberbullying detection. Unlike conventional models, deep learning architectures can distinguish complex patterns and relationships between data. Long short-term memory (LSTM) networks and their bidirectional variations (BiLSTM) have been shown to be particularly useful for processing sequential input such as text, taking into account past and future contexts [12]. This ability is crucial for understanding the underlying intent of words, which is often important for recognizing conversational dynamics and abusive situations.

D. Integration of lexical representations.

The development of complex neural networks coincided with the transformative impact of word embeddings on the NLP field. GloVe embeddings, which offer word representations

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

based on co-occurrence matrices, have played a key role in explaining semantic relationships between terms in text corpora [13]. These embeddings have shown significant effectiveness in improving the performance of deep learning models on NLP tasks by providing a more comprehensive, pre-trained context for word meanings. In the context of cyberbullying detection, GloVe can significantly improve the model's understanding, allowing it to more accurately distinguish between harmful and harmless interactions.

E. Attention mechanisms: a focused approach.

The introduction of attention mechanisms into BiLSTM networks represents an advance in modeling skill. Originally developed for machine translation tasks, attention mechanisms allow models to focus on specific segments of input data that are more relevant to the task. In the field of cyberbullying identification, this means that the model can pay more attention to segments of text that predict bullying behavior, thereby increasing detection accuracy [14].

Challenges in multi-class classification.

Despite these advances, multi-class classification of cyberbullying poses additional challenges. Cyberbullying can take many forms, and distinguishing between them (e.g., bullying and insult) requires a comprehensive understanding of the content and a complex interpretation of the context [15]. Research has shown that multiclass models face challenges due to overlapping categories and unbalanced distribution of data sets, with some types of bullying being less represented than others.

G. Current models and shortcomings.

Recent literature has examined different topologies of LSTM networks and attention processes to improve performance on specific NLP tasks. However, there is a significant lack of research addressing the complex needs of cyberbullying detection across multiple categories using these advanced methodologies. Furthermore, while there is considerable research on the effectiveness of GloVe implementations, there is limited understanding of their synergistic effects with BiLSTM and attention mechanisms in a unified model focused on cyberbullying detection.

H. Empirical studies and effectiveness.

Empirical studies evaluating the effectiveness of these combined models have shown encouraging results, with significant improvements in accuracy and recall compared to previous models. However, these studies often emphasize the need for a more complete dataset that accurately reflects the diverse and evolving characteristics of online communication [16]. Furthermore, the literature presents a significant discourse on the ethical implications and potential biases inherent in automated detection systems that must be addressed to ensure fairness and accuracy in detection.

I. Promising trajectories.

The literature suggests many avenues for improving cyberbullying detection. One avenue is to improve the interpretability of models, which is important for understanding model reasoning, especially in sensitive contexts such as cyberbullying. Another area of focus is the adaptation of models for real-time detection, which presents specific challenges in terms of scalability and responsiveness.

In addition, there is a growing demand for interdisciplinary research that combines perspectives from psychology, sociology, and computer science to develop comprehensive and context-sensitive models. This combination could lead to significant advances in effectively reducing cyberbullying through tailored treatments [17].

The literature highlights the complexity of cyberbullying detection and the promise of modern technological innovations to reduce it. Despite significant advances, persistent challenges require continued innovation and research within the discipline. The proposed work seeks to improve upon this ongoing initiative by using advanced NLP technology to create a

DOI: 10.37943/22PSRO3633

comprehensive cyberbullying detection system that advances academic knowledge and offers practical methods for addressing an important social problem.

This part examines the advancement of cyberbullying detection using machine learning, focusing on BiLSTM networks, GloVe embeddings, and attention processes. Conventional models such as SVM and Random Forest have limitations in contextual comprehension, but deep learning improves detection precision. GloVe embeddings enhance semantic comprehension, whereas attention processes sharpen concentration on essential text portions. Challenges in multi-class categorization, dataset imbalances, and ethical issues remain prevalent. Recent studies underscore the efficacy of these methodologies while advocating for enhanced interpretability, real-time flexibility, and transdisciplinary study.

Methods and Materials

The dataset utilized in this study was acquired from Twitter through the Twitter API over a three-month duration in early 2024. A total of 40,000 tweets were collected, consisting of 20,000 categorized as hate speech and 20,000 as neutral. Expert raters conducted manual annotation in accordance with comprehensive guidelines, guaranteeing superior labeling quality and inter-annotator concordance.

Before model training, the text data underwent preprocessing, which included lowercasing, elimination of special characters, tokenization, and padding to maintain consistent input length. The dataset was subsequently divided into training (80%) and testing (20%) subsets via stratified sampling to maintain label distribution across the sets.

The model's efficacy was assessed utilizing conventional classification metrics, encompassing accuracy, precision, recall, F1-score, and AUC-ROC. All scores were calculated based on the predictions from the test set. Confusion matrices were employed to ascertain true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), which constituted the foundation for all metric computations.

In the Table 1 depicts the complex architecture of a neural network for detecting cyberbullying in social media texts. The network is organized in the following order.

Table 1. Scatter Plot for Hate Speech vs Offensive Language

Layer Name	Input Shape	Output Shape	Layer Type
embedding_input	(None, 56)	(None, 56)	Input Layer
embedding	(None, 56)	(None, 56, 200)	Embedding Layer
self_attention_1	(None, 56, 200)	(None, 56, 200)	Self-Attention
bidirectional_lstm	(None, 56, 200)	(None, 56, 28)	Bidirectional LSTM
self_attention_2	(None, 56, 128)	(None, 56, 128)	Self-Attention
dense	(None, 56, 128)	(None, 56, 32)	Dense
dense_1	(None, 56, 32)	(None, 56, 1)	Dense
flatten	(None, 56, 1)	(None, 56)	Flatten
dense_2	(None, 56)	(None, 8)	Dense
dense_3	(None, 8)	(None, 5)	Dense (Output Layer)

Input layer: The input layer receives a sequence of text data, each sequence represented by a total length of 56 tokens.

The input layer transforms the input tokens into a 200-dimensional vector space using pre-trained GloVe embeddings. This layer is essential for transforming discrete text input into continuous vectors that encapsulate semantic relationships between words.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

$$X = \{p_1, p_2, ..., p_n\} \quad (1)$$

Where X represents the complete input sequence. Each p_i represents the i-th token, word, or vector in the series. n represents the length of the sequence, namely the quantity of words or tokens in the input.

The initial self-attention mechanism processes the 200-dimensional input vectors after the embedding stage. This layer allows the model to assess the importance of different terms in the text, improving its ability to focus on relevant semantic attributes for cyberbullying detection.

Bidirectional long short-term memory layer: While paying attention, a 128-unit bidirectional long short-term memory (BiLSTM) layer studies the text data in both forward and backward directions. This bidirectionality allows the model to assimilate the context of previous and subsequent stages, which facilitates a thorough understanding of the sequence.

Second self-attention layer: Following the BiLSTM, an additional self-attention layer evaluates the 128-dimensional output. This layer, akin to the original attention mechanism, augments the model's capacity to concentrate on semantically pertinent characteristics inside the sequence. The attention mechanism is calculated via the scaled dot-product attention formula:

$$Q = XW^Q,\ K = XW^K,\ V = XW^V \quad (2)$$

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

In this context, Q, K and V represent the query, key, and value matrices obtained from the BiLSTM output representations, whereas d_k kdenotes the dimension of the key vectors, which is typically 128. This approach enables the model to dynamically assess the significance of various tokens throughout the final classification decision-making process.

The output of the second attention mechanism is processed by a dense layer of 32 units, which converts the extracted features into a compact representation.

Output dense layer: The subsequent dense layer reduces the dimensionality to a single unit for each sequence step, thus preparing the data for final categorization.

Smoothing layer: This layer transforms the two-dimensional data from the previous dense layer output into a one-dimensional array suitable for final processing, transforming the output into a singular vector.

Final dense layers: The smoothed vector is further processed by an additional dense layer of 8 units, followed by a final dense layer with 5 units corresponding to the categorization categories. The final layer uses a softmax activation function to generate probabilities for each category, which provides a basis for multi-class classification of text into multiple types of cyberbullying.

This design takes advantage of the synergistic benefits of GloVe embeddings, self-attention processes, and BiLSTM units to address the complex challenges of identifying subtle and context-sensitive language that represents cyberbullying behavior. The systematic implementation of these technologies ensures that each element contributes to a more detailed understanding of the input data, thereby improving the overall predictive accuracy of the model [18].

DOI: 10.37943/22PSRO3633

The Fig. 1 shows a bar chart depicting the distribution of tweets across several categories of cyberbullying, along with non-cyberbullying tweets, in the dataset used to train the detection algorithm [19]. The categories shown consist of the categories "ethnicity", "age", "faith", "gender", and "non-cyberbullying", each of which is marked with a unique color. The "non-cyberbullying" category shows the highest frequency, containing 7,268 tweets, indicating a significant prevalence of neutral or non-abusive content in the sample. Within the cyberbullying categories, "religion" ranks highest with 7,937 tweets, followed by "age" with 7,865 tweets, followed by "gender" with 7,448 tweets, and "nationality" with the lowest frequency with 7,683 tweets. This distribution reflects the frequency of cyberbullying related to religion and gender issues in the collected data, suggesting that these domains may require targeted attention in detection initiatives. The approximately equal distribution across categories reinforces the model's requirement for a diverse array of instances to effectively learn and accurately categorize different manifestations of cyberbullying [20].

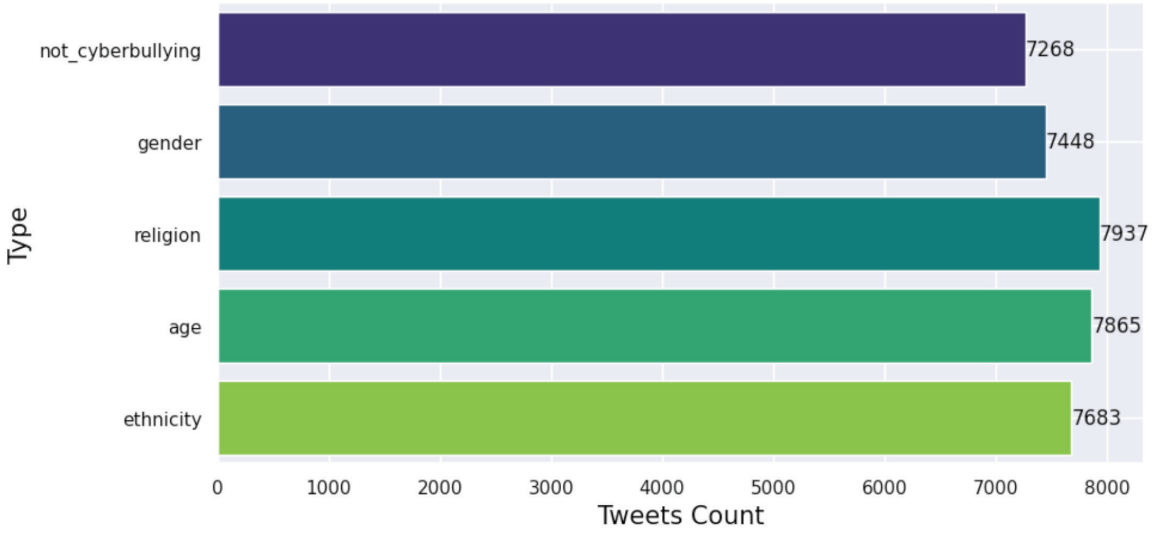

Figure 1. Distribution of Tweet Categories in Cyberbullying Dataset

The Fig. 2 presents a boxplot of the distribution of word lengths across several cyberbullying categories and non-cyberbullying tweets. Each category—ethnicity, age, religion, gender, and non-cyberbullying—is represented by separate boxes that show the median, interquartile range, and extremes of word lengths across tweets. The "religion" category shows the widest interquartile range, indicating significant variability in tweet length for this type of cyberbullying, which may reflect different expressions and contexts within this category. The "ethnicity" category shows the smallest interquartile range and median, indicating concise tweet content. The "non-cyberbullying" category, represented in contrasting color with a single outlier, shows a relatively narrow distribution, indicating uniformity in word length across tweets that are considered non-cyberbullying. This boxplot is critical for understanding the textual attributes of tweets in these categories, which can guide preprocessing and modeling efforts to detect cyberbullying.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

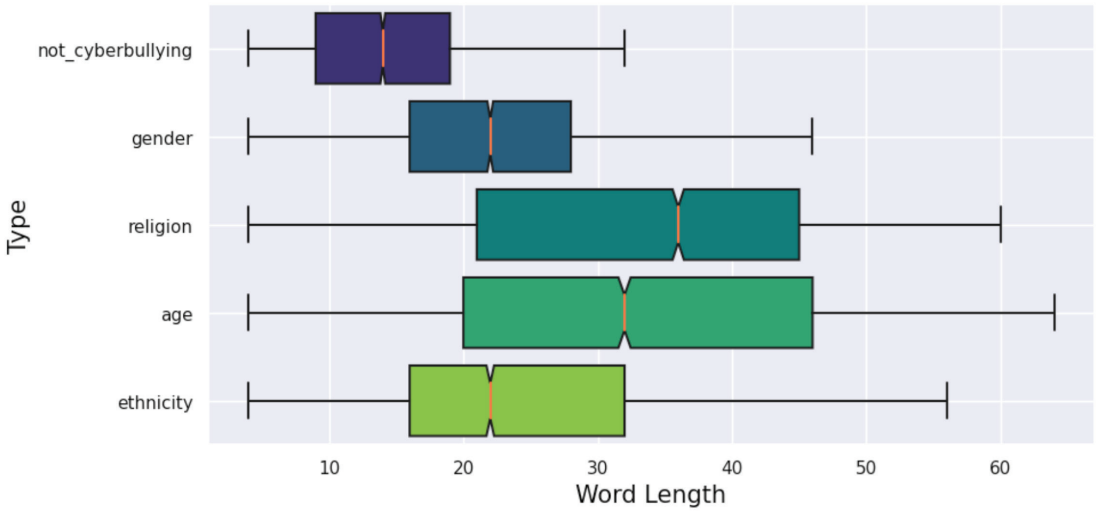

Figure 2. Word Length Distribution by Cyberbullying Categories

The Fig. 3 presents a scatterplot analysis, along with density distributions, showing the correlation between the average word length and total word count in tweets categorized by different forms of cyberbullying, labeled as types 0 to 4. The scatterplots (dots), indicate an increase in the prevalence of shorter and longer words, indicating a higher prevalence of total word count for that category of cyberbullying. The scatterplots of types 1 to 4, labeled with blue, green, purple, and light blue dots, respectively, show different patterns in word usage and length, with type 1 showing a trend toward higher average word length relative to the other types. The marginal histograms above and to the right of the scatterplot provide additional density information regarding total word count and average word length, with visible peaks indicating data point concentrations in certain regions. This visualization makes it easier to understand the textual dynamics across multiple categories of cyberbullying, highlighting potential variations in linguistic style that are important for developing specific detection algorithms.

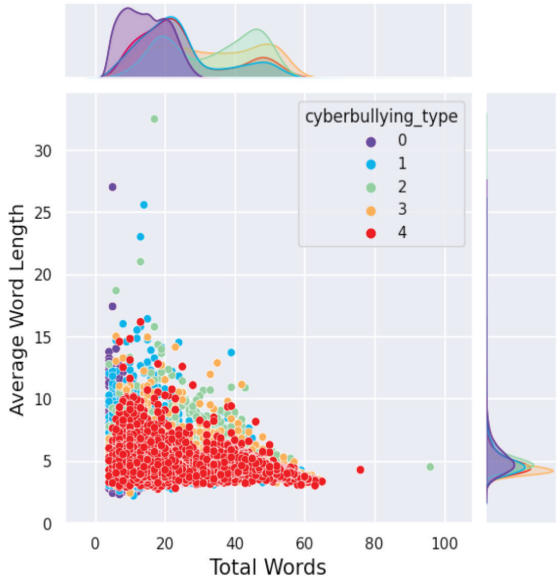

Figure 3. Scatterplot and Density Distributions of Average Word Length Versus Total Words by Cyberbullying Type

DOI: 10.37943/22PSRO3633

The Materials and Methods section of this paper describes the detailed methodology for developing and evaluating a neural network model for cyberbullying detection. The model uses GloVe implementations for semantic processing, combines the sequential learning capabilities of BiLSTM networks, and employs self-focus mechanisms for targeted contextual analysis. A dataset containing various labeled instances of cyberbullying across multiple categories was used for training and validation. The model's performance was rigorously evaluated using common metrics such as ROC curves and confusion matrices to assess its accuracy and recall across different cyberbullying scenarios. This methodology provides a solid foundation for testing the effectiveness of modern machine learning algorithms in understanding and addressing the challenges of cyberbullying in digital interactions.

The materials and methods part of article constructs a neural network for the detection of cyberbullying using GloVe embeddings, BiLSTM networks, and self-attention processes. The model analyzes text input via many phases: embedding transformation, self-attention evaluation, bidirectional LSTM processing, supplementary attention mechanisms, and thick layers for classification.

Results

The Fig. 4 shows a set of word clouds that depict the frequency and importance of keywords in several categories related to cyberbullying. The categories include gender, religion, age, ethnicity, and non-bullying, and a graphical representation of the most frequently used words in each category. The size of each word in the cloud indicates its frequency, with larger terms indicating greater use in the dataset. The gender-related word cloud includes words such as "gay," "sexist," and "rape," indicating that conversations or cyberbullying incidents in this domain often involve sexual orientation and gender-based discrimination. The "religion" cloud emphasizes the words "Islam," "Muslim," and "terrorist," indicating that religious identity and related prejudices are concentrated. The age category emphasizes words such as "school," "bully," and "child," indicating the prevalence of cyberbullying among young people, especially in educational settings. Cyberbullying based on ethnicity is distinct from blatant racist slurs, as illustrated in the word cloud that represents online racial discrimination. The "Non-bullying" cloud is a collection of neutral conversational terms that are not related to harassment, such as "people," "think," and "go." Finally, the "All Tweets" cloud combines components from all categories, providing a comprehensive perspective on the most common phrases across all conversations, including bullying and non-bullying settings. These visualizations serve as a powerful tool for quickly understanding the thematic and linguistic components prevalent in each category, helping to formulate targeted strategies to monitor and mitigate cyberbullying on social media platforms.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

Figure 4. Word Clouds Representing Keyword Distributions Across Different Cyberbullying Categories

The Fig. 5 shows a bar chart depicting the frequency distribution of the ten most common words identified in the cyberbullying dataset. The terms are ordered by frequency, with “school” occurring 8,960 times, followed by “blah” at 5,668 instances, and “like” at 5,449 instances. Other notable terms include “girl,” “nigger,” “joke,” “dumb,” “high,” “Muslim,” and “bully,” plotted in descending order of frequency. Each bar is color-coded from dark purple to light green to clearly distinguish each word’s occurrence. This distribution is critical to understanding the common themes and terminology used in cyberbullying scenarios, providing insights into areas that may require more attention in detection and prevention strategies. The graph serves as both a quantitative assessment of word usage and a qualitative representation of the themes and perspectives that are common in discussions of cyberbullying.

DOI: 10.37943/22PSRO3633

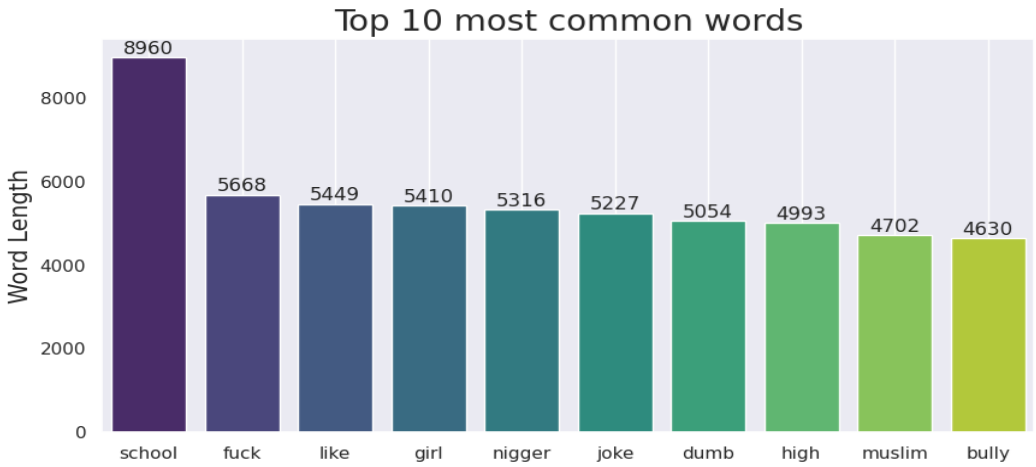

Figure 5. Frequency Distribution of Top 10 Most Common Words in Cyberbullying Dataset

The Fig. 6 depicts a histogram showing the distribution of tweet lengths in the dataset, with a focus on tweets containing a higher number of words, namely between 11 and 100. The x-axis represents the number of words in a tweet, while the y-axis represents the frequency of tweets with each word count. The histogram shows a decreasing trend in tweet frequency as word count increases, reflecting a typical feature of social media communication that favors brevity. The bars indicate a significant concentration of tweets in the lower word count range, with the highest frequency being 11 words per tweet, for a total of 2,643 tweets. There is a significant decrease in frequency as word count increases, indicating that tweets containing extended text are less common. For word counts between 11 and 30, the count generally remains constant, but drops significantly beyond this range. The spikes in random frequencies occur at certain high word counts, namely 74, 86, and 105 words, although these events are quite rare relative to the overall trend. This distribution is important for understanding the average length of tweets in the dataset, which can impact the design and effectiveness of text analysis algorithms designed to efficiently process and classify such data.

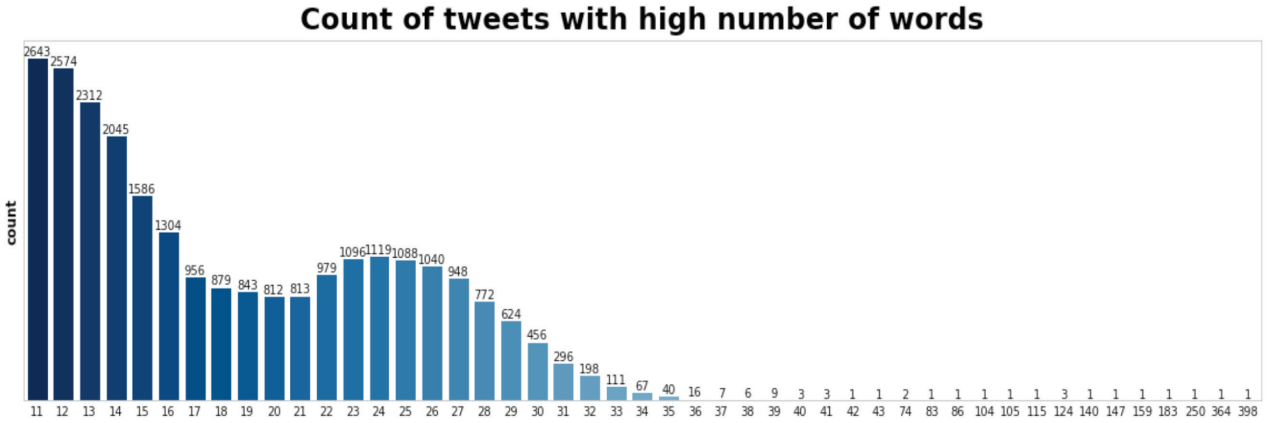

Figure 6. Distribution of Tweet Lengths in the Dataset

The Fig. 7 depicts the confusion matrix of the cyberbullying classification model, which provides a comprehensive overview of the model's performance across different categories of cyberbullying, including religion, age, ethnicity, gender, other cyberbullying, and no-bullying. The matrix is structured with test categories on the vertical axis and predicted categories on the horizontal axis. Each cell in the matrix represents the number of predictions made by the model, with diagonal cells representing the number of correct predictions for each category,

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

from top left to bottom right. The model shows robust predictive accuracy for specific categories, with religion, age, and ethnicity showing high true positive rates of 1541, 1528, and 1523, respectively. However, the matrix also highlights instances of misclassification, particularly in the context of gender and additional forms of cyberbullying. The gender category was sometimes confused with other types of cyberbullying, as indicated by the off-diagonal value of 120. In addition, a significant number of cases classified as other cyberbullying were erroneously identified as non-bullying, as indicated by the number 344. This visualization is important for identifying the strengths and limitations of the model, facilitating improvements aimed at increasing its classification accuracy across different types of cyberbullying.

Figure 7. Frequency Distribution of Top 10 Most Common Words in Cyberbullying Dataset

The Fig. 8 shows the evolution of training and validation loss and accuracy for a machine learning model over multiple training epochs. The graph shows four separate lines representing training loss (green), training accuracy (blue), validation loss (orange), and validation accuracy (red) over epochs 0 to 12. This visualization style is an estimate of the model's performance and overfitting during the training period. The training loss shows a significant decrease from epoch 0 to around epoch 2, indicating a rapid improvement in the model's learning and adaptation to the training data. The training accuracy also increases significantly in the early epochs, stabilizing at a high level over the duration of the training procedure. In contrast, the validation loss initially decreases but soon stabilizes and remains fairly constant, indicating that the model does not show a comparable rate of improvement on the unseen validation data despite the improvement in performance on the training set. The validation accuracy remains largely constant during the training process, indicating that the model's ability to generalize to new data does not improve significantly during the initial training period. The gap between training and validation measures may indicate an overfitting problem, where the model overlearns the intricacies of the training data, reducing its effectiveness on new, unknown data.

DOI: 10.37943/22PSRO3633

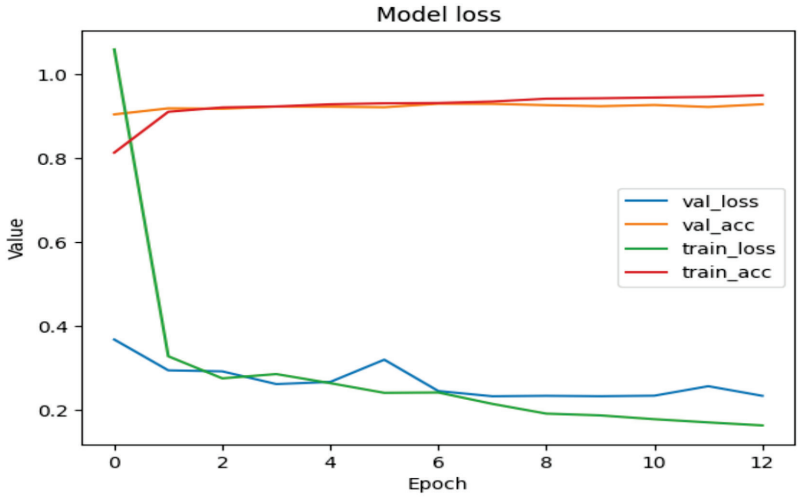

Figure 8. Training and Validation Loss and Accuracy Over Epochs

The Fig. 9 shows the Receiver Operating Characteristic (ROC) curves for a logistic regression model applied to multiple abuse categories. The ROC curve graphically depicts the diagnostic performance of a binary classifier system as its discrimination threshold is adjusted. This metric represents the true positive rate (TPR) relative to the false positive rate (FPR) for age, ethnicity, gender, religion, and additional types of cyberbullying, without cyberbullying. Each category is highlighted in a specific color, and the area under the curve (AUC) for each category is shown, indicating the ability of the model to discriminate across classes. The curves show robust performance across most categories, with significant increases in AUC for age (0.997) and religion (0.996), indicating that the model is particularly adept at accurately identifying true cases of cyberbullying associated with these factors, while also reducing false positives. The ethnicity and gender categories show particularly good discrimination capabilities, with AUCs of 0.993 and 0.979, respectively. The categories of no_cyberbullying and other_cyberbullying show lower AUC values of 0.905 and 0.921, respectively, indicating somewhat less effective but still commendable performance in differentiating these conditions. The ROC curve provides important insights into the effectiveness of the logistic regression model, highlighting its strengths in certain categories of cyberbullying detection, and identifying areas for potential improvement.

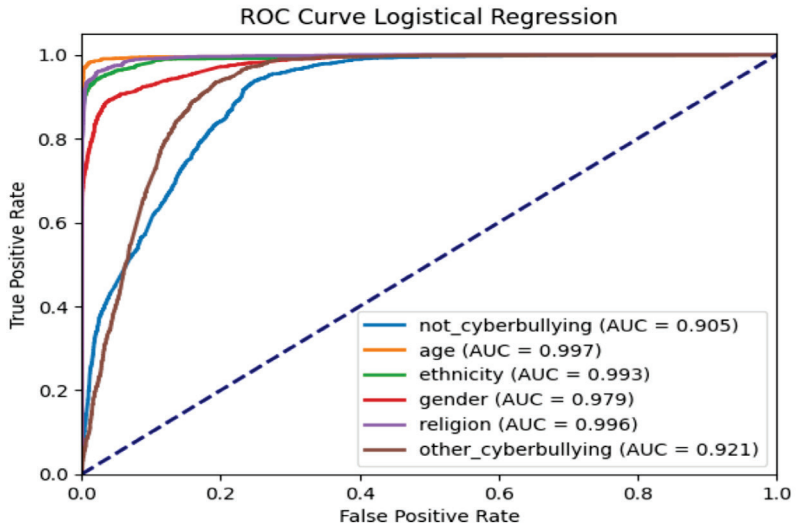

Figure 9. ROC Curves for Logistic Regression Model Across Different Cyberbullying Categories

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

The evaluation results of the neural network model show high proficiency in recognizing and classifying different forms of cyberbullying with considerable accuracy. The model showed particular skill in identifying cyberbullying based on religion and age, as evidenced by high AUC values approaching 1.0. However, it struggled with more complex categories such as gender and different forms of cyberbullying, where AUC values dropped significantly and misclassifications were common. The data show that while the model is very effective in specific settings, its performance varies across several categories, reflecting the complexity of the language used in cyberbullying. The training and validation measurements showed an initial rapid improvement in model accuracy, followed by a plateau in validation accuracy, indicating overfitting to the training data. This discrepancy suggests the need for further refinement of the model to improve its generalization skills in different real-world context.

Discussion

This research illustrates the efficacy of deep learning techniques, particularly BiLSTM networks integrated with GloVe embeddings and self-supervised mechanisms, in identifying cyberbullying across many categories. The model demonstrated exceptional efficacy in detecting religion- and age-related cyberbullying, with AUC scores of 1.0, signifying a robust capacity to discern distinct discriminatory patterns. Nonetheless, categorization performance deteriorated in more nuanced categories, such as gender bullying and other subtle forms, where the model shown heightened misclassification. The results demonstrate that although the model effectively identifies explicit discriminatory material, it encounters difficulties with overlapping or context-dependent signals that are more subtle.

A notable problem faced was overfitting, as indicated by the disparity between training and validation accuracy. This indicates restricted generalization abilities and highlights the want for enhanced regularization methods, data augmentation tactics, and a broader variety of training data to bolster resilience.

To mitigate these constraints, it is essential to juxtapose the existing BiLSTM-based methodology with more sophisticated transformer models like DistilBERT and RoBERTa. These models provide enhanced contextual comprehension owing to their attention-based architecture, facilitating a more refined interpretation of language. Initial comparisons in related research indicate that transformers frequently surpass conventional RNN-based models in text classification tasks, particularly in addressing nuanced semantic distinctions and contextual overlap. Integrating such models into forthcoming research could markedly diminish classification errors, particularly in intricate categories like gender bullying.

In addition to performance factors, the study recognizes the ethical ramifications of implementing automated cyberbullying detection systems. Although AI provides scalable instruments for the surveillance and identification of detrimental information, it must be created with responsibility to avert bias and guarantee equity. False positives may unjustly suppress innocuous content, whilst they permit the continuation of detrimental material. The incorporation of explainable AI (XAI) methodologies will be essential for enhancing transparency, fostering user confidence, and guaranteeing accountability in practical applications.

This study advances cyberbullying detection with contemporary natural language processing techniques. Nonetheless, ongoing enhancement is essential to augment classification precision, reduce prejudice, and adhere to ethical standards. Subsequent research should investigate the incorporation of transformer-based models like RoBERTa and DistilBERT, use hybrid modeling approaches, and create adaptive frameworks capable of evolving with the fluid dynamics of online conversation.

Conclusion

This study significantly improved our understanding of the application of deep learning techniques to detect and classify cyberbullying in various social media environments. The study successfully used an advanced neural network model that combines BiLSTM networks embedded in GloVe with self-attention mechanisms to accurately identify specific categories of cyberbullying based on religion and age, achieving high accuracy, as indicated by AUC values of around 1.0. However, challenges remain in accurately categorizing more complex classifications, such as gender and different forms of cyberbullying, where linguistic nuances and topic overlaps hinder the predictive effectiveness of the model. The appearance of overfitting, as indicated by the gap between training accuracy and validation scores, highlights the complexities of natural language modeling and the need for tactics that improve model generalization. This study advances AI technology for social good and highlights the critical need to continuously improve categorization algorithms to address the dynamic and complex nature of the language used in cyberbullying. Future initiatives should prioritize expanding training datasets, exploring hybrid modeling methodologies, and formulating ethical rules to regulate AI applications that ensure responsible and effective use of these technologies to combat cyberbullying and improve online social interactions.

Acknowledgment

This study was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant No. AP23488900- Automatic detection of cyberbullying among young people in social networks using artificial intelligence)

References

[1] Atif, A., Zafar, A., Wasim, M., Waheed, T., Ali, A., Ali, H., & Shah, Z. (2024). Cyberbullying Detection and Abuser Profile Identification on Social Media for Roman Urdu. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3445288

[2] Talpur, K. R., Yuhaniz, S. S., & Amir, N. N. B. (2020). Cyberbullying detection: Current trends and future directions. Journal of Theoretical and Applied Information Technology, 98(16), 3197-3208. https://core.ac.uk/download/pdf/425547762.pdf

[3] Ahmadinejad, M., Shahriar, N., & Fan, L. (2023). Self-Training for Cyberbully Detection: Achieving High Accuracy with a Balanced Multi-Class Dataset (Doctoral dissertation, PhD thesis, Faculty of Graduate Studies and Research, University of Regina). https://www.proquest.com/openview/2e6b484d78e3a1fe0486ec1217dd574c/1?pq-origsite=gscholar&cbl=18750&diss=y

[4] Rao, M.P., Kota, N., Nidumukkala, D., Madoori, M., & Ali, D. (2024, April). Enhancing Online Safety: Cyberbullying Detection with Random Forest Classification. In 2024 10th International Conference on Communication and Signal Processing (ICCSP) (pp. 389-393). IEEE. https://doi.org/10.1109/ICCSP60870.2024.10543598

[5] Kaarthika, R., & Hemamalini, R. (2024, July). Enhancing Cyberbullying Detection Through Keyword Filtering: A Comparative Study of ML and DL Approaches. In 2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT) (pp. 1-6). IEEE. https://doi.org/10.1109/IConSCEPT61884.2024.10627823

[6] Saeid, A., Kanojia, D., & Neri, F. (2024, June). Decoding Cyberbullying on Social Media: A Machine Learning Exploration. In 2024 IEEE Conference on Artificial Intelligence (CAI) (pp. 425-428). IEEE. https://doi.org/10.1109/CAI59869.2024.00084

[7] Dharani, M., & Sathya, S. (2024). Deep Learning Algorithms with Adam Optimization for Detecting of Cyberbullying Comments. Nanotechnology Perceptions, 627-639. https://nano-ntp.com/index.php/nano/article/download/746/676/1257

[8] Sultan, D., Omarov, B., Kozhamkulova, Z., Kazbekova, G., Alimzhanova, L., Dautbayeva, A., ... & Abdrakhmanov, R. (2023). A Review of Machine Learning Techniques in Cyberbullying Detection. Computers, Materials & Continua, 74(3). https://doi.org/10.32604/cmc.2023.033682

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 22, JUNE 2025

[9] Kumar, C., Kumar, K.A., Gupta, S., & Sardar, T.H. (2024, March). Cyberbullying detection based on the fusion of DistilBERT and SIMHASH Technique. In 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA) (pp. 1-4). IEEE. https://doi.org/10.1109/AIMLA59606.2024.10531427

[10] Hoque, M.N., Chakraborty, P., & Seddiqui, M.H. The Challenges and Approaches during the Detection of Cyberbullying Text for Low-resource Language: A Literature. https://doi.org/10.37936/ecti-cit.2023172.248039

[11] Saranyanath, K.P., Shi, W., & Corriveau, J.P. (2022, September). Cyberbullying Detection using Ensemble Method. In CS & IT Conference Proceedings (Vol. 12, No. 15). CS & IT Conference Proceedings. https://doi.org/10.22215/etd/2022-15070

[12] Sari, T.I., Ardilla, Z.N., Hayatin, N., & Maskat, R. (2022). Abusive comment identification on Indonesian social media data using hybrid deep learning. IAES International Journal of Artificial Intelligence, 11(3), 895-904. https://doi.org/10.11591/ijai.v11.i3.pp895-904

[13] Liu, M. (2023, July). A Creativity Survey of Cyberbullying Classification Based on Social Network Analysis. In Proceedings of the 2nd International Conference on Mathematical Statistics and Economic Analysis, MSEA 2023, May 26–28, 2023, Nanjing, China. https://doi.org/10.4108/eai.26-5-2023.2334259

[14] Bhamidi, M., Nandyala, M., Dayalan, R., Karthik, N., & Vani, V. (2024, February). COOL: Classification of Online Offensive Language Using Machine Learning and Deep Learning. In International Conference on Computational Intelligence in Data Science (pp. 87-97). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-69982-5_7

[15] Mohite, S.S., Attar, V., & Kalamkar, S. (2022, October). Shaming tweets detection on Twitter using Machine learning Algorithms. In 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT) (pp. 1-6). IEEE. https://doi.org/10.1109/GCAT55367.2022.9972100

[16] Ismail, A.A., & Yusoff, M. (2022). An efficient hybrid LSTM-CNN and CNN-LSTM with GloVe for text multi-class sentiment classification in gender violence. International Journal of Advanced Computer Science and Applications, 13(9). https://doi.org/10.14569/IJACSA.2022.0130999

[17] Ibrahim, Y.M., Essameldin, R., & Darwish, S.M. (2024). An Adaptive Hate Speech Detection Approach Using Neutrosophic Neural Networks for Social Media Forensics. Computers, Materials & Continua, 79(1). https://doi.org/10.32604/cmc.2024.047840

[18] Koshiry, A. M. E., Eliwa, E. H. I., Abd El-Hafeez, T., & Omar, A. (2023). Arabic toxic tweet classification: leveraging the arabert model. Big Data and Cognitive Computing, 7(4), 170. https://doi.org/10.3390/bdcc7040170

[19] Sharma, D.K., Singh, B., Agarwal, S., Pachauri, N., Alhussan, A.A., & Abdallah, H.A. (2023). Sarcasm detection over social media platforms using hybrid ensemble model with fuzzy logic. Electronics, 12(4), 937. https://doi.org/10.3390/electronics12040937

[20] Slobodzian, V., Molchanova, M., Kovalchuk, O., Sobko, O., Mazurets, O., Barmak, O., & Krak, I. (2022, September). An Approach Based on the Visualization Model for the Ukrainian Web Content Classification. In 2022 12th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 400-405). IEEE. https://doi.org/10.1109/ACIT54803.2022.9913162