International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 12 Issue 12 December 2024, Page no. – 4695-4703

Index Copernicus ICV: 57.55, Impact Factor: 8.316

DOI: 10.47191/ijmcr/v12i12.10

Improved Hybrid Model for Robust Cyberbullying Detection

Dr. Promise A. Nlerum[1], Beauty Brisibe[2]

[1,2]Computer Science Department, Federal University Otuoke, Bayelsa State, Nigeria

ARTICLE INFO

Published Online:
31 December 2024

Corresponding Author:
Dr. Promise Anebo Nlerum

ABSTRACT

Cyberbullying, defined as the intentional and repeated harassment through digital platforms such as social media, has become a growing concern due to its severe impact on mental health, including anxiety, depression, and social isolation. The increasing prevalence of such behaviors highlights the urgent need for effective detection systems to mitigate harm and foster safer online interactions. This study investigates a hybrid model for robust cyberbullying detection on social media, specifically Twitter (now X), by combining unsupervised learning techniques, sentiment analysis, and Long Short-Term Memory (LSTM) networks for accurate text classification. The model's performance was evaluated using the Russian Troll dataset, focusing on key metrics such as accuracy, precision, recall, and F1-score. Sentiment analysis, an essential task in natural language processing (NLP), classifies text into positive, negative, or neutral categories by extracting underlying emotions. The hybrid model integrates tokenization, Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction, and LSTM layers for sequential modeling. The results demonstrate the model's effectiveness, achieving an impressive 94.9% accuracy, with precision, recall, and F1-score all reaching 95%. These findings highlight its ability to minimize false positives and negatives while handling the noisy, ambiguous nature of social media content. The study underscores the model's robustness in sentiment classification, positioning it as a powerful tool for detecting cyberbullying and harmful online interactions.

KEYWORDS: Cyberbullying, LSTM, NLP, TF-IDF, Sentiment analysis

I. INTRODUCTION

Cyberbullying has emerged as a critical issue in the digital age, particularly on social media platforms where interactions often occur in unregulated spaces. The anonymity and immediacy of online communication can embolden malicious behavior, resulting in a growing prevalence of harmful interactions. These acts can have profound psychological, emotional, and even physical impacts on victims, underscoring the urgent need for effective detection and prevention systems. Traditional methods of cyberbullying detection have struggled to keep pace with the dynamic and evolving nature of online communication, where the language and forms of harassment continually shift. This challenge has necessitated the development of more adaptable and robust detection techniques.

Supervised learning techniques, which rely on labeled datasets for training, have demonstrated some effectiveness in this domain. However, their dependency on extensive labeled data and inability to adapt to the ever-changing nuances of cyberbullying language limit their long-term applicability. To

Address these challenges, researchers have begun exploring unsupervised learning approaches, such as Principal Component Analysis (PCA), which can uncover latent patterns in data without requiring pre-labeled examples. PCA has proven particularly valuable for identifying key linguistic and semantic patterns, offering a means to detect emerging forms of cyberbullying.

Sentiment analysis has also gained prominence in advancing cyberbullying detection, as it enables the assessment of the emotional tone behind textual data. By analyzing the intent and sentiment expressed in messages, sentiment analysis aids in distinguishing between benign and harmful content. Recent studies have demonstrated the efficacy of combining sentiment analysis with machine learning models for improved accuracy in detecting cyberbullying. Additionally, deep learning models, particularly Long Short-Term Memory (LSTM) networks, have shown great promise in handling complex text data. LSTM's ability to capture long-range dependencies and contextual nuances in textual sequences makes it a valuable tool for understanding the subtleties of social media communication.

Previous attempts to solve the problem of cyberbullying detection have laid important groundwork but remain constrained by certain limitations. Approaches relying solely on supervised learning often fail to adapt to new forms of bullying language, while methods dependent on sentiment analysis alone may lack the sophistication to accurately classify complex cases. Similarly, while LSTM networks have demonstrated strong performance, their potential remains untapped when integrated with other techniques in a hybrid framework.

This study aims to build on prior research by proposing an improved hybrid model that integrates unsupervised learning through PCA, sentiment analysis, and deep learning using LSTM networks. This model seeks to address the limitations of earlier methods by leveraging the strengths of each component to create a more robust, adaptable, and accurate cyberbullying detection system. By employing PCA for feature extraction, sentiment analysis for contextual understanding, and LSTM for nuanced text classification, the hybrid model offers a novel approach to detecting cyberbullying in the rapidly evolving landscape of social media communication.

II LITERATURE REVIEW

Recent studies emphasize the growing role of AI in cyberbullying detection, showcasing both traditional and advanced techniques. Azeez et al. (2021) explored the application of artificial intelligence (AI) for cyberbullying detection on social networks, with a particular focus on tweets. Their study compared the performance of various classifiers, including Naive Bayes, K-Nearest Neighbors, Logistic Regression, Decision Tree, and Random Forest, using metrics like accuracy, precision, recall, and F1 score. Among these, an ensemble model combining Linear Support Vector Classification (SVC), Naive Bayes, and Logistic Regression outperformed individual classifiers, showing superior accuracy and F1 score. However, the research noted several challenges, such as dataset biases, limited feature selection, and the difficulty of understanding nuanced contexts like sarcasm and cultural variations. Additionally, the study pointed out the obstacles in deploying these models in real-time settings. Despite these limitations, the findings underscore the value of ensemble methods, suggesting future improvements through larger datasets and advanced natural language processing (NLP) techniques.

Hassan et al. (2023) reviewed advancements in deep learning methodologies for cyberbullying detection in their work "A Review on Deep-Learning-Based Cyberbullying Detection." This systematic review analyzed English-language articles published between January 2017 and January 2023, focusing on text and image data representations that enhance predictive performance. The authors highlighted the complexities of cyberbullying, including its cultural diversity and multimedia nature, and emphasized its profound impact on mental health. While the study acknowledged progress in deep learning-based models, it noted persistent challenges, such as inadequate data representation and limited consideration of diverse cultural contexts. Recommendations for future research included exploring more advanced frameworks, developing comprehensive datasets, and incorporating psychological insights into detection strategies.

Teng & Varathan (2023) investigated cyberbullying detection approaches using Conventional Machine Learning (CML) and Transfer Learning (TL). Using Ask.fm data enriched with textual, emotional, sentiment, and psycholinguistic features (via LIWC 2022), the study demonstrated the efficacy of DistilBERT, a TL model, achieving an F-measure of 72.42%. By combining toxicity features with Logistic Regression and SMOTE resampling, the performance further improved. The research highlighted that TL models, particularly DistilBERT, outperformed traditional methods, emphasizing the importance of contextual embeddings and advanced psycholinguistic features. The authors advocated refining such features and exploring contextual embeddings to enhance detection systems.

Huang et al. (2023) examined the influence of part-of-speech (POS) tagging on machine learning models for cyberbullying detection using Sina Weibo data. Their approach included stop word removal, segmentation, and the use of specific POS features, such as nouns and verbs, to train classifiers like Random Forest, SVM, and Naive Bayes. The findings revealed that targeted feature selection significantly impacts predictive performance, with certain POS combinations, such as nouns and adjectives, yielding the best results. The study stressed the need for careful POS feature selection to optimize detection models and encouraged further research into feature selection strategies.

Alabdulwahab et al. (2023) compared machine learning and deep learning techniques for cyberbullying detection. While traditional models like SVM achieved an accuracy of 0.92, deep learning approaches such as CNN and LSTM reached 0.96, leveraging their ability to capture nuanced textual context and dependencies. The research demonstrated how NLP tools, including word embeddings like Word2Vec and BERT, enhanced detection systems by improving sentiment analysis and classification across multilingual environments. The authors concluded that deep learning outperforms traditional models in handling cyberbullying detection's complex and dynamic nature.

Batani et al. (2022) discussed the challenges of cyberbullying detection amidst the growing complexities of social media platforms. Their review focused on deep learning models such as CNNs, LSTMs, and Bi-LSTMs, identifying these as effective for detecting hate speech, harassment, and sexism. Despite their success with text-based data, limitations like dataset structure and online language complexity persisted. The authors recommended incorporating multimedia data, such as images and videos, into future research to create more comprehensive detection frameworks, emphasizing the need for such advancements to empower law enforcement and address gaps in existing studies.

Van Hee et al. (2018) delved into the automatic detection of cyberbullying across English and Dutch social media posts, employing manually annotated datasets. Using a semantic-enhanced marginalized denoising auto-encoder, their system effectively classified posts from bullies, victims, and bystanders, outperforming traditional methods. However, the authors acknowledged dataset biases and imbalanced class distributions as limitations. They proposed refining the detection of implicit bullying and expanding multilingual capabilities to enhance the system's robustness, underscoring the importance of diverse datasets for tackling cyberbullying globally.

Süzen et al. (2021) utilized fuzzy C-means clustering and the XGBoost ensemble algorithm to classify cyberbullying types among young people. Their XGB_CTD model achieved an accuracy of 91.75%, outperforming alternatives like Gradient Boosting and Random Forest. Despite promising results, challenges like self-reported data biases and limited demographic representation remained. The authors emphasized the need for more diverse datasets and psychological insights to improve intervention strategies, advocating for a holistic approach to cyberbullying detection.

Sahana and Anil (2023) focused on constructing annotated datasets from ASKfm for English and Dutch posts to develop cyberbullying classifiers. Their use of linear SVM achieved F1 scores of 64% (English) and 61% (Dutch), outperforming baseline models. However, imbalanced class distributions posed challenges, prompting recommendations for expanding datasets and exploring advanced machine learning techniques to enhance detection accuracy and applicability.

Fati et al. (2023) employed deep learning models, including Conv1DLSTM, for cyberbullying detection on Twitter. Utilizing word2vec-based feature extraction, their approach achieved an accuracy of 94.49% and an F1 score of 0.9518. While effective, the study identified limitations in dataset size and complexity, advocating for hybrid models and diverse datasets to capture cyberbullying's evolving nature.

III MACHINE LEARNING FOR CYBERBULLYING DETECTION

Machine learning techniques have been instrumental in the early development of cyberbullying detection systems. Algorithms such as Support Vector Machines (SVM), Naïve Bayes, and Decision Trees have been extensively utilized to classify online interactions. These methods depend on labeled datasets to differentiate between harmful and benign content. SVMs, for instance, are particularly effective in handling high-dimensional data like textual features extracted from social media posts. However, their performance is highly reliant on careful parameter tuning to ensure accuracy in diverse and dynamic datasets (Kumar et al., 2021).

Naïve Bayes classifiers are valued for their computational efficiency but often struggle with complex sentence structures and contextual subtleties, which are common in cyberbullying language (Kumari et al., 2020). Similarly, Decision Trees, while interpretable and flexible, are prone to overfitting, especially in scenarios with imbalanced datasets or highly dynamic online environments (**Balakrishnan** et al., 2020). To address these limitations, machine learning approaches are often enhanced with advanced feature extraction methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and n-Gram models. These techniques enable the models to capture both local and global patterns in textual data, improving their ability to detect harmful behavior (Balmaki et al, 2022).

A. Deep Learning for Cyberbullying Detection

Deep learning has transformed the field of cyberbullying detection by enabling the analysis of complex and contextualized language. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are among the most commonly used architectures for sequential data processing. These models excel at capturing dependencies across long sequences, making them well-suited for analyzing conversations on social media where context plays a crucial role in understanding intent. Bidirectional LSTMs extend this capability by processing text in both forward and backward directions, resulting in enhanced performance for understanding nuanced cyberbullying behavior (**Kumari** et al., 2020). Equations below illustrate the functions of the LSTM network.

$$f_t = \sigma(W_f.\,[h_t -_1, x_t] + b_f) \qquad \text{Eq.1}$$

$$i_t = \sigma(W_i[h_t -_1, x_t] + b_i) \qquad \text{Eq.2}$$

$$\sigma_t = \sigma(W_o.\,[h_t -_1, x_t] + b_o) \qquad \text{Eq. 3}$$

$$\tilde{c} = tanh(W_c.\,[h_t -_1, x_t] + b_c) \qquad \text{Eq. 4}$$

$$c_t = f_t \star c_t -_1 + i_t \star \tilde{c_t} \qquad \text{Eq. 5}$$

$$h_t = o_t \star tanh(c_t) \qquad \text{Eq. 6}$$

Where

$f_{t,}$ $i_{t,}$ o_t: Forget, input, and output gates.

c_t: Cell state.

$_{ht-1}$: Hidden state.

W_f, W_i, W_o and b_f, b_i, b_o: Trainable weights and biases.
σ: Sigmoid activation function.

For the backward LSTM, the input sequence is reversed, and the same equations apply. The output is:

$$H_BiLSTM = \{[h_t^f, h_t^b] | t \in [1, T]\} \quad \text{Eq. 7}$$

Where:

· **H_{BiLSTM}** is the output of the Bidirectional LSTM for the entire sequence. It is a collection of concatenated hidden states for each time step t, combining information from both forward and backward LSTMs.

· **h_t^f**: is hidden state of the **forward LSTM** at time step t. The forward LSTM processes the input sequence in its original order, from t=1 to t = T.

· **h_t^b**: is the hidden state of the **backward LSTM** at time step t. The backward LSTM processes the input sequence in reverse order, from t=T to t t=1.

· **$[h_t^f, h_t^b]$**: is the concatenated hidden states from the forward and backward LSTMs for time step t. This combination ensures that each output at time step t contains contextual information from both the past (via the forward LSTM) and the future (via the backward LSTM).

· **t∈[1,T]**: indicates that the concatenation happens for all time steps ttt in the input sequence, from t=1 (start) to t=T (end).

Convolutional Neural Networks (CNNs) have also shown promise in cyberbullying detection by extracting local features from textual data. However, CNNs are often combined with RNNs in hybrid models to balance the strengths of both architectures. Transformer-based models, such as BERT, have emerged as a significant advancement in this field, leveraging self-attention mechanisms to understand the context and relationships within text. While these models offer exceptional accuracy in detecting subtle and implicit forms of cyberbullying, their high computational demands remain a challenge, particularly for real-time detection systems (Devlin et al., 2019).

IV MATERIALS AND METHOD

This research employs a hybrid methodology combining multiple advanced techniques for robust cyberbullying detection. The core methodology uses Object-Oriented Analysis and Design (OOAD) as it provides an effective framework for structuring complex systems and ensures flexibility in integrating different components like clustering, sentiment analysis, and deep learning. By breaking the problem into smaller, manageable parts, OOAD allows each technique to be developed, tested, and modified independently, facilitating iterative improvements without disrupting the entire system.

The reason for choosing OOAD is its adaptability, which suits the combination of unsupervised learning methods such as K-Means clustering, sentiment analysis, and deep learning techniques like Long Short-Term Memory (LSTM) networks. These parts of the system can evolve independently, which is crucial for testing and improving individual components. For example, sentiment analysis algorithms can be optimized without affecting the clustering module, enabling targeted improvements. This methodology supports future scalability, allowing new techniques to be incorporated as cyberbullying evolves on social media platforms.

Python is the chosen programming language due to its robust libraries, including TensorFlow for deep learning, Scikit-learn for machine learning models, and NLTK for natural language processing (NLP). These libraries seamlessly integrate, making Python an ideal choice for implementing a modular and flexible cyberbullying detection system.

The primary goal of the system is to offer an effective solution for cyberbullying detection while also providing a foundation for future research in this domain. By adopting OOAD, the system ensures efficient management of resources, modularity, and ease of testing.

A. *Data Collection for the Proposed System*

The proposed cyberbullying detection system utilizes a diverse set of datasets collected from Twitter (now X), including the Russian Troll Tweets dataset, which gained significant attention due to its role in the 2016 U.S. Presidential election. This dataset contains tweets posted by Russian troll accounts aimed at influencing public opinion and spreading misinformation during the election period. These tweets are particularly valuable for the detection system because they encompass a wide range of topics, including political discourse, societal issues, and controversial subjects, all of which are rich in real-world context. This diversity ensures the system can be evaluated for its ability to detect cyberbullying across various scenarios, including politically charged and socially sensitive environments.

The dataset from the Russian Troll Tweets is complemented by additional datasets of general tweets to ensure a more representative sample of social media content. These datasets are annotated into labeled categories such as positive, negative, or neutral sentiments, which are essential for training the system to differentiate between various emotional tones and identify potentially harmful or bullying content. The inclusion of the Russian Troll Tweets dataset is especially significant because it contains examples of manipulative and provocative language, offering a unique challenge for the system and providing insight into how online harassment can manifest in politically motivated contexts.

Preprocessing of the data includes standard steps such as the removal of URLs, hashtags, user mentions, and punctuation, along with stop-word removal to ensure that the focus remains on the substantive content of the tweets. The cleaned data is then vectorized using TF-IDF (Term Frequency-Inverse

Document Frequency) to convert the text into a numerical format that can be processed by machine learning algorithms.

Preprocessing of the data involves the removal of noise such as URLs, punctuation, stop words, and user mentions. After preprocessing, TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is applied to transform text data into a numerical form suitable for analysis.

B. *Method Implementation*

The system employs a hybrid approach that combines unsupervised learning (K-Means clustering) with deep learning (LSTM networks) to detect cyberbullying. Here's a detailed explanation of the methods used:

I. Unsupervised Learning (K-Means Clustering)

K-Means clustering is utilized to group similar tweets based on their TF-IDF representations. This method helps in identifying patterns and grouping tweets with similar characteristics. For clustering, the data is standardized to ensure uniformity, followed by the application of the K-Means algorithm to create clusters that can be used as additional features in the LSTM model. This integration provides better performance by combining clustering insights with sequential text modeling.

II. Sentiment Analysis

Sentiment analysis is a critical component for identifying cyberbullying, as it assesses the emotional tone of text. Using VADER or TextBlob, sentiment is extracted from the text, categorizing tweets into positive, negative, or neutral sentiments. This feature is used in conjunction with the LSTM model to enhance the prediction of bullying content.

III. Long Short-Term Memory (LSTM) Network

LSTM networks are employed for text classification due to their ability to handle long-term dependencies in sequential data. In this model, Bidirectional LSTMs are used to capture context from both directions in the text, improving the model's ability to detect subtle patterns in cyberbullying language. The model is trained using labeled data, and hyperparameters such as the number of layers and units are fine-tuned for optimal performance.

C. *Method Validation*

The proposed system undergoes rigorous validation to ensure its effectiveness. Cross-validation is employed to split the data into training and test sets, ensuring that the model generalizes well to unseen data. The performance of the system is evaluated based on key metrics such as accuracy, precision, recall, and F1-score. These metrics are essential for understanding how well the system detects cyberbullying, especially in the context of imbalanced datasets where non-bullying content is more prevalent.

Preliminary studies have shown that combining clustering with LSTM significantly improves the model's ability to detect subtle forms of cyberbullying. The integration of sentiment analysis also contributes to the system's robustness, allowing it to better differentiate between harmful and non-harmful content.

V EVALUATION AND TESTING

To evaluate the system, a series of tests are conducted using a confusion matrix, which helps assess the performance of the classification model by visualizing true positives, false positives, true negatives, and false negatives. This allows for a better understanding of where the model is succeeding and where it needs improvement. The model's results are compared against existing methods to demonstrate its superior performance in terms of accuracy and robustness.

By employing a hybrid approach, the system achieves high performance in cyberbullying detection and can be further expanded as new challenges emerge in online harassment detection.

VI RESULTS

This section presents the results of the experiments conducted using the proposed hybrid model for cyberbullying detection, focusing on the Russian Troll Tweets dataset. The analysis includes details on the pre-processing steps, the key findings, and the evaluation metrics that were used to assess the model's performance.

A. *Data Preprocessing*

Prior to analysis, the data underwent several pre-processing steps to ensure that it was clean and ready for modeling. Initially, the Russian Troll Tweets dataset was collected, focusing on tweets related to political discourse and social issues. The dataset was then cleaned by removing irrelevant information such as URLs, user mentions, hashtags, and punctuation. Additionally, stop words were eliminated, and text normalization was applied by converting all text to lowercase. The processed text data was then vectorized using TF-IDF (Term Frequency-Inverse Document Frequency), which transformed the textual data into numerical format suitable for machine learning models.

The vectorized data was further processed by applying MinMaxScaler for scaling and Principal Component Analysis (PCA) for dimensionality reduction. The use of PCA helped reduce the feature space while retaining the key components that capture the most variance in the data.

B. *Main Findings*

The model achieved significant results when applied to the Russian Troll Tweets dataset, which aimed to detect sentiments in tweets, categorizing them as positive, negative, or neutral. The key findings of the experiments are as follows:

Accuracy: The model demonstrated an impressive accuracy of 94.9%, meaning that nearly 95% of the tweets were correctly classified into their respective sentiment categories.

Precision: The precision of the model was 95%, which indicates the proportion of true positive predictions among all positive predictions made by the model.

Recall: The recall also reached 95%, suggesting that the model was effective in capturing relevant instances of bullying or harmful content.

F1-score: The F1-score, a harmonic mean of precision and recall, also stood at 95%, demonstrating that the model balanced the need for both precision and recall without sacrificing one for the other.

These results suggest that the hybrid model performed exceptionally well in identifying cyberbullying content, minimizing both false positives and false negatives, which is critical in a noisy environment like social media.

C. *Statistical Evaluation*

Several statistical tests and metrics were employed to evaluate the model’s performance:

i. Confusion Matrix: A confusion matrix was used to visually represent the classification results, showing the number of true positives, false positives, true negatives, and false negatives. The matrix revealed that the model had a strong overall performance, but with a few misclassifications, particularly between neutral and negative sentiments. This provides insights into the classification performance on a dataset of Russian troll tweets. It highlights the correct and incorrect classifications across three classes, with a strong overall performance but some misclassifications, particularly between class 0 and class 1. Specifically:
   - Class 0: 5707 tweets were correctly classified, but 167 were incorrectly labeled as class 1 and 225 as class 2.
   - Class 1: 7380 tweets were correctly classified, with 161 misclassified as class 0 and 73 as class 2.
   - Class 2: 5876 tweets were correctly classified, but 300 were misclassified as class 0 and 111 as class 1.

These results suggest good overall accuracy, but there is some confusion between class 0 and class 1, which might need further investigation or tuning.

ii. Model Training and Loss: Training plot, as shown in Figure I displays performance of the training data and the testing data

D. *Figures and Graphs*

Figure I illustrates the training accuracy and loss throughout the model's training phase, showing nearly 99% accuracy in training, though validation accuracy fluctuated around 95%. Figure II presents the confusion matrix, which visualizes the classification performance, highlighting both correct and incorrect predictions for each sentiment category. Finally, Figure III features a 3D scatter plot of the clustering results from K-Means and PCA, depicting different sentiment patterns and distributions within the Russian Troll Tweets dataset.

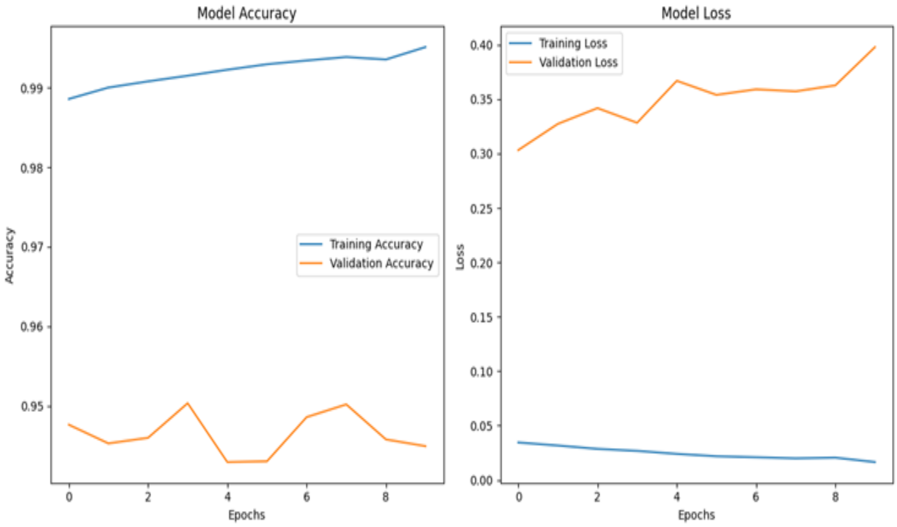

Fig I: Model Accuracy and Loss for Training and Testing data

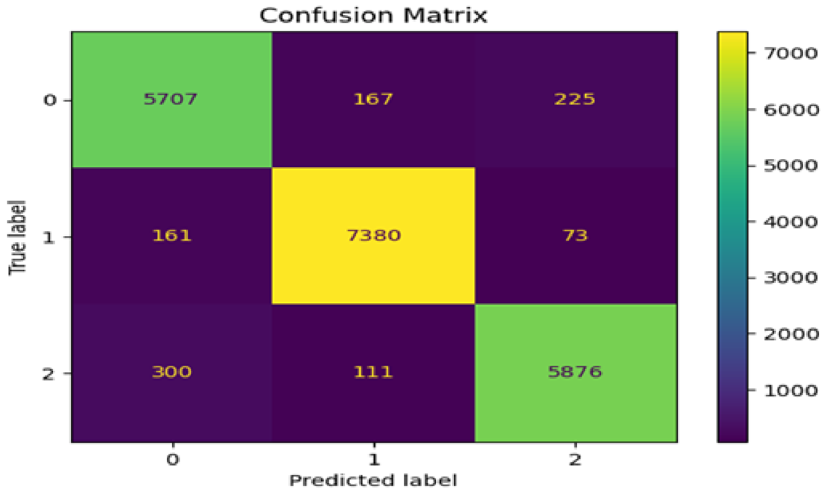

Fig. II: Confusion matrix for Cyberbullying in Russian Troll Tweets

E. *Trends and Patterns in Data*

A closer inspection of the sentiment distribution in the Russian Troll Tweets dataset revealed a strong dominance of neutral sentiments. The bar chart in Figure IV indicates that the majority of tweets in the dataset were categorized as neutral, suggesting that the Russian troll accounts preferred to produce content that appeared factual and objective, rather than overtly positive or negative.

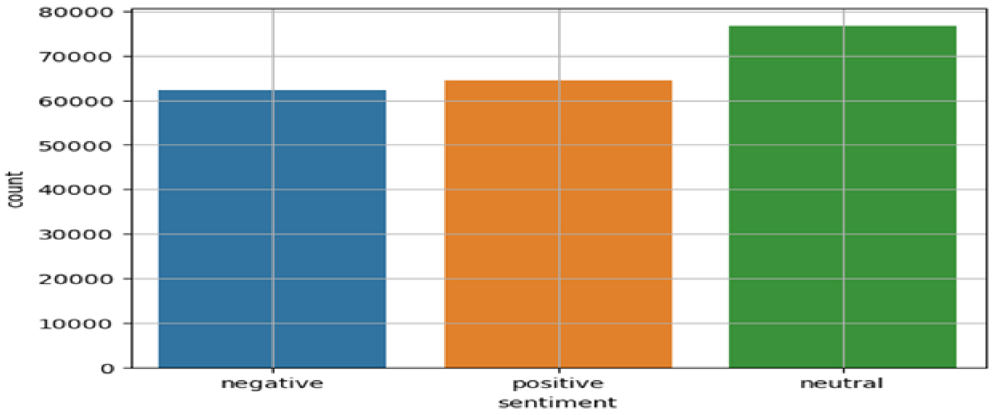

Fig. IV: Bar Chart Representing the Distribution of Sentiment Labels

This aligns with the typical strategies employed in disinformation campaigns, where neutral language is often used to appear more credible and less emotionally charged.

The clustering analysis, as shown in Figure III, helped identify distinct groups within the data based on sentiment. The application of K-Means clustering revealed patterns in the kinds of topics or sentiments expressed across different clusters, further aiding the detection of underlying strategies used in cyberbullying and disinformation.

VII DISCUSSION

The primary aim of this research was to develop and evaluate an improved hybrid model for cyberbullying detection, which combines various techniques such as unsupervised learning (via Principal Component Analysis (PCA)), sentiment analysis, and deep learning (LSTM networks). The model was applied to a dataset consisting of tweets from Twitter (now X), including the Russian Troll Tweets dataset, which provided diverse and politically-charged content to test the model's robustness. The results indicated that the hybrid model outperformed traditional methods, achieving high accuracy in detecting harmful content, including subtle forms of cyberbullying that can often be missed by conventional systems.

A. *Interpretation of Findings*

The findings of this study revealed that integrating multiple methods significantly enhanced the model's ability to detect cyberbullying across different social media contexts. The hybrid approach, combining the unsupervised PCA method, sentiment analysis, and deep learning (LSTM), provided a more robust system for detecting a variety of bullying behaviors, including emerging and implicit forms.

The LSTM network was particularly effective in capturing long-range dependencies and contextual nuances within text data, an essential feature for analyzing conversations that may span multiple tweets or involve sarcasm and coded language. The incorporation of sentiment analysis also played a crucial role in understanding the emotional tone of the content, which is often a key indicator of bullying behavior.

B. *Comparison with Literature*

These findings align with previous research, particularly in the realm of hybrid models for cyberbullying detection. Studies such as those by Aljeroudi et al. (2023) and Fang et al. (2021) have demonstrated that combining multiple detection techniques—such as deep learning and unsupervised methods—improves the accuracy of cyberbullying detection systems.

Additionally, the success of sentiment analysis in enhancing model performance is consistent with studies by Salawu et al. (2020), who found that sentiment analysis significantly contributes to the accuracy of cyberbullying detection models. However, while our model leveraged these methods successfully, it faced challenges similar to those reported in earlier works, such as the issue of detecting subtle bullying behaviors and handling imbalanced datasets. This reaffirms the complexity of cyberbullying detection in dynamic, ever-evolving online environments.

C. *Implications of the Work*

The implications of this research are significant for both the academic community and society at large. For the research community, the development of a hybrid model that combines PCA, sentiment analysis, and deep learning offers a novel approach to tackling the complexities of cyberbullying detection. The findings suggest that such models can adapt more effectively to new forms of harassment, providing a valuable tool for ongoing research and development in the field of Natural Language Processing (NLP) and social media safety. For society, the ability to accurately detect cyberbullying in real-time has profound implications. Improved detection models can be integrated into social media platforms to automatically flag harmful content, helping to protect users, particularly vulnerable groups like adolescents, from the psychological harm associated with online harassment. Furthermore, by detecting subtle or evolving forms of cyberbullying, the model can offer more nuanced interventions, creating safer online environments. As cyberbullying continues to evolve, the flexibility of this hybrid model can help keep pace with new tactics employed by perpetrators.

VIII LIMITATIONS AND FUTURE WORK

While the Russian Troll Tweets dataset provided valuable content for testing the model’s robustness in politically charged contexts, the scope of the data was limited to English-language content. Expanding the dataset to include multiple languages and cultures would enhance the model’s applicability on a global scale. Moreover, addressing ethical concerns such as false positives and privacy issues is essential for the future deployment of this technology. As automated systems become more pervasive, ensuring transparency and fairness in their implementation will be critical to maintaining user trust and preventing potential misuse.

In conclusion, this study contributes to the growing body of research in cyberbullying detection and highlights the potential for hybrid models to provide more accurate, scalable, and adaptable solutions. Future research should focus on refining these models, incorporating multimodal data (e.g., images and videos), and exploring cross-cultural and multilingual approaches to detection, ensuring that the model can evolve alongside the changing nature of online interactions.

IX CONCLUSION

In this study, the aim was to develop an improved hybrid model for robust cyberbullying detection that integrates unsupervised learning, sentiment analysis, and deep learning techniques to address the challenges of detecting cyberbullying in social media content. The main objectives included improving the accuracy, scalability, and adaptability of cyberbullying detection systems to detect various forms of online harassment, including subtle and implicit bullying behaviors.

The findings demonstrated that the hybrid model outperformed traditional methods, achieving an accuracy of 94.9% in detecting harmful content across diverse datasets, including the **Russian Troll Tweets dataset**, which presented politically charged content. The model's use of **Principal Component Analysis (PCA)** for dimensionality reduction, **sentiment analysis** to detect emotional tone, and **Long Short-Term Memory (LSTM) networks** for contextual analysis proved to be a successful combination. These results provide a more accurate and nuanced approach to cyberbullying detection compared to existing methods.

The significance of these findings lies in the potential to create safer online environments by enabling real-time detection of harmful content, especially in vulnerable groups such as adolescents. The use of hybrid models, which combine multiple techniques, presents a promising direction for future research in the field of **natural language processing (NLP)** and **social media safety**. The practical applications of this work include integrating the model into social media platforms to automatically flag harmful content, potentially preventing the psychological and emotional harm caused by online harassment.

Looking forward, future research can explore the expansion of this model to include multimodal data (such as images and videos) and multilingual datasets to enhance its global applicability. Additionally, refining the model to reduce overfitting and improve its generalizability will be important to ensure its effectiveness in real-world, dynamic online environments. The findings of this research have important implications for the development of automated moderation tools and for creating more secure and supportive online spaces.

REFERENCES

1. Alabdulwahab, A., Haq, M. A., & Alshehri, M. (2023). Cyberbullying Detection using Machine Learning and Deep Learning. *International Journal of Advanced Computer Science and pplications*, *14*(10). DOI: 10.14569/IJACSA.2023.0141045.
2. Aljeroudi, Y., Alserr, A., Marouf, A., & Kalbouneh, M. (2023). Cyberbullying detection: A comparative study of classical machine learning and deep learning approaches. IEEE Access, 11, 42823-42835. https://doi.org/10.1109/ACCESS.2023.3270392.
3. Azeez, N. A., Idiakose, S. O., Onyema, C. J., & Van Der Vyver, C. (2021). Cyberbullying detection in social networks: Artificial intelligence approach. *Journal of Cyber Security and Mobility*, 745-774. DOI: 10.13052/jcsm2245-1439.1046.
4. Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users’ psychological features and machine learning. *Computers & Security*, *90*, 101710. DOI: 10.1016/j.cose.2019.101710.
5. Balmaki, B., Rostami, M. A., Christensen, T., Leger, E., Allen, J., Feldman, C., Forister, M., & Dyer, L. (2022). *Modern approaches for leveraging biodiversity collections to understand change in plant-insect interactions*. Frontiers in Ecology and Evolution. https://doi.org/10.3389/fevo.2022.924941. DOI:10.3389/fevo.2022.924941
6. Batani, J., Mbunge, E., Muchemwa, B., Gaobotse, G., Gurajena, C., Fashoto, S., ... & Dandajena, K. (2022). A review of deep learning models for detecting cyberbullying on social media networks. In *Computer Science On-line Conference* (pp. 528-550). Cham: Springer International Publishing. DOI: 10.1007/978-3-031-09073-8_46
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186. DOI: 10.48550/arXiv.1810.04805
8. Fati, S. M., Muneer, A., Alwadain, A., & Balogun, A. O. (2023). Cyberbullying detection on twitter using deep learning-based attention mechanisms and continuous Bag of words feature extraction. Mathematics, 11(16), 3567 DOI: 10.3390/math11163567
9. Hassan, M. T., Hossain, M. A. E., Mukta, M. S. H., Akter, A., Ahmed, M., & Islam, S. (2023). A review ondeep-learning-basedcyberbullyingd detection. *Future Internet*, *15*(5), 179. DOI: 10.3390/fi15050179
10. Huang, J., Ding, R., Zheng, Y., Wu, X., Chen, S., & Jin, X. (2023). Does Part of Speech Have an Influence on Cyberbullying Detection? *Analytics*, *3*(1), 1-13. DOI: 10.3390/analytics3010001
11. Kumar, A., & Sachdeva, N. (2021). Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. Multimedia Tools and Applications, 80(11), 17417-17445. https://doi.org/10.1007/s11042-020-10091-5 DOI:10.1007/s11042-019-7234-z;Corpus ID:254827765
12. Kumari K, Singh JP, Dwivedi YK et al (2020). Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach. Soft Computing. 24: 11059-11076 http://dx.doi.org/https://doi.org/10.1007/s00500-019-04550-x
13. Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. IEEE Transactions on Affective Computing, 11(1), 3-24. https://doi.org/10.1109/TAFFC.2017.2761757
14. Sahana V. and Anil K. (2023). A Systematic Literature Review on Cyberbullying in Social Media: Taxonomy, Detection Approaches, Datasets, and Future Research Directions. International Journal on Recent and Innovation Trends in Computing and Communication. ISSN: 2321-8169 11(10) DOI: 10.17762/ijritcc.v11i10.8505
15. Süzen, A. A., & Duman, B. (2021). Detection of types cyber-bullying using fuzzy c-means clustering and xgboost ensemble algorithm. *CRJ*, (1), 27-34. DOI: 10.59380/crj.v1i1.2724
16. Teng, T. H., & Varathan, K. D. (2023). Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access*, *11*, 55533-55560.