International Journal of Research in Signal Processing, Computing & Communication System Design
6 (2) 2025, 01-07
www.publishingindia.com/ijrsp

Methods for Automated Cyberbullying Detection Driven by Natural Language Processing

M. V. S. Narayana[1], Rapaka Usha[2] and Kallubhavi Obulesh[3]

[1]Assistant Professor, Computer Science and Engineering, Malla Reddy Engineering College for Women, JNTU, Hyderabad, Telangana, India. Email: vsnarayanamcse@mrecw.edu.in

[2]Assistant Professor, Computer Science and Engineering, Malla Reddy Engineering College for Women, JNTU, Hyderabad, Telangana, India. Email: rapakausha1@gmail.com

[3]Assistant Professor, Computer Science and Engineering, Malla Reddy Engineering College for Women, JNTU, Hyderabad, Telangana, India. Email: mtechemt@gmail.com

Abstract: In today's digitally linked society, cyberbullying is a severe danger that calls for efficient detection and preventive techniques. A strong multi-tiered approach for identifying and categorizing cyberbullying on various internet platforms is proposed in this study. To achieve high accuracy in differentiating bullying from non-bullying information, the system is trained on extensive and diverse datasets using sophisticated machine learning and natural language processing techniques. It improves inclusivity and dependability by taking language and cultural quirks into consideration. With its flexible design, the system adjusts to new trends in cyberbullying and maintains its efficacy over time. This research lessens the effects of online harassment, promotes the development of more inclusive online communities, and supports the establishment of safer digital environments by facilitating proactive interventions.

Keywords: Accuracy, Contextual, Cyberbullying, Detection systems, Lemmatization, Machine learning, Natural language processing, Precision, Semantic, Social media, Social media communities, Statistical, Tokenization, Tweets.

I. Introduction

The promise of a globally connected digital society has been undermined by the increase of cyberbullying [1]. Online platforms have made it possible for harassment and abuse to proliferate even though they provide previously unheard-of chances for community development and communication [2]. In contrast to traditional bullying, cyberbullying flourishes in online environments where its intensity and reach are increased by anonymity, accessibility, and quick content sharing [3]. It causes serious emotional, psychological, and social suffering to people of all ages, genders, and origins [3]. To tackle this problem, sophisticated detection systems that can recognize and categorize various types of online abuse are needed. In order to detect cyberbullying, this study offers a thorough framework that incorporates cultural sensitivity and adaptability while utilizing machine learning and natural language processing [2] [3]. The objective is to promote more secure, welcoming, and abuse-free online spaces [4].

The widespread use of social media platforms and rapid communication methods has made cyberbullying a serious social problem in the digital age [4]. Victims frequently experience psychological repercussions including anxiety and depression, which can have serious consequences [5]. Automated systems are crucial since traditional monitoring techniques cannot handle the enormous volume of user-generated content [6]. Advanced techniques for examining textual patterns and determining malicious intent are provided by natural language processing, or NLP [6] [7]. Through the use of sentiment analysis, deep learning networks, and machine learning algorithms, researchers may create predictive models that accurately categorize cyberbullying [7]. The purpose of this study is to investigate how NLP-based prediction algorithms identify cyberbullying in order to promote safer online environments [7].

Learning, sharing, and interaction opportunities have been made possible by the growth of online communities. But these same platforms have also made it easier for bullying, abuse, and hate speech to proliferate [7] [8] . Therefore, detecting cyberbullying has become a crucial area of research, especially in order to stop online abuse before it gets out of hand [8]. By employing algorithms that can handle massive amounts of textual material, automated detection systems seek to lessen reliance on human moderators [9]. Text preprocessing, feature extraction, and classification models are examples of NLP techniques that are essential for identifying abusive language [9] [10]. In order to promote healthier online interaction, this project investigates the creation of an NLP-based prediction engine that can more accurately detect cyberbullying trends [10].

Manual monitoring of harmful online communication is challenging, as the digital space generates an overwhelming volume of user-generated data daily [11] [12]. Automated prediction systems provide a scalable solution by detecting offensive patterns using computational models [13]. NLP bridges the gap between raw text and meaningful classification by converting language into features suitable for machine learning [12]. With advancements in artificial intelligence, cyberbullying detection models can be trained to capture semantic, contextual, and emotional aspects of communication [12]. This research focuses on building a framework that utilizes NLP pipelines to classify abusive versus non-abusive content [12] [3]. The proposed system seeks to enhance early detection and minimize the harmful psychological and social consequences faced by victims [3].

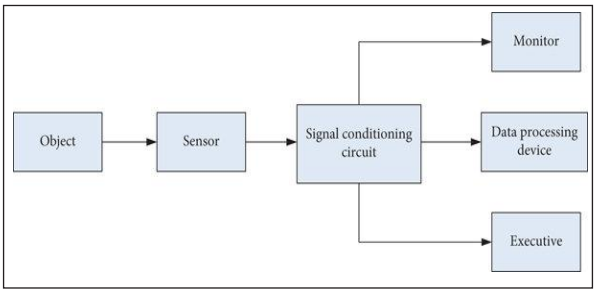

Fig. 1: Block Diagram of NLP-Model

II. Literature Review

A. *Cyber Bullying Detection on Social Media Using Machine Learning*

This study's primary goal is to create a strong, multi-tiered system that can identify and categorize cyberbullying on a variety of digital platforms. The system attempts to achieve high accuracy in differentiating cyberbullying instances from non-bullying conversations while taking linguistic diversity and cultural nuances into consideration by utilizing sophisticated machine learning and natural language processing techniques. Making sure the system is flexible enough to react to changing trends in online harassment is crucial to its long-term viability. By facilitating proactive interventions, this research hopes to promote safer, more inclusive online communities in addition to accurate detection. By doing this, it helps to promote healthy online interactions and lessen the negative effects of digital abuse.

B. *A Robust Hybrid Machine Learning Model for Cyberbullying Detection in Social Media*

Numerous studies have looked into using natural language processing (NLP) to identify dangerous online activity. Keyword-based detection, which identified offensive terms from predetermined dictionaries, was the main focus of early research. However, these approaches frequently misclassified non-offensive usage of flagged phrases due to their lack of contextual understanding. Subsequently, scientists presented machine learning classifiers including Decision Trees, Support Vector Machines (SVM), and Naïve Bayes. These algorithms used statistical patterns of abusive language to increase accuracy. For contextual comprehension, more recent research emphasizes deep learning architectures such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Scalable detection systems are made possible by the literature's consistent demonstration that NLP-based models perform better than manual moderation. However, there are still issues with dealing with sarcasm, mixed languages, and new slang terms.

Cyberbullying detection systems performed noticeably better when word embeddings like Word2Vec and Glove were used. Embeddings, as opposed to conventional bag-of-words models, record semantic linkages, enabling systems to identify negative intent that goes beyond overt insults. Contextual embeddings such as BERT and Roberta, which allow for more accurate categorization by taking surrounding words into account, have also been investigated by a number of researchers. According to published research, transformer-based architectures perform better in complicated datasets—particularly those with subtle linguistic patterns—than conventional machine learning methods. Comparative studies also highlight how data preprocessing methods like stop-word removal, lemmatization, and stemming can increase the efficacy of models.

C. *Cyber Bullying Detection for Hindi-English Language Using Machine Learning*

To train and evaluate cyberbullying detection algorithms, researchers have looked into a variety of datasets. Because of their accessibility and variety of user interactions, publicly accessible corpora from sites like Twitter, Facebook, and Reddit are frequently used. Research shows that model accuracy is severely hampered by dataset imbalance, where non-bullying samples greatly outnumber bullying occurrences. Techniques like data augmentation, oversampling, and transfer learning have been put forth to get around issue. The ethical ramifications of dataset gathering are highlighted in recent work, particularly with regard to user consent and privacy. Despite the fact that many models exhibit high accuracy in controlled tests, domain-specific issues limit their practical application. Researchers therefore support cross-platform models that can generalize in a variety of online contexts.

Machine learning techniques for cyberbullying detection are evaluated in a number of comparative studies. Results indicate that as compared to single models, ensemble models—which incorporate several classifiers—frequently attain higher accuracy. For instance, it has been shown that Random Forest and Gradient Boosting are more accurate in identifying

offensive words. However, because deep learning techniques can automatically extract hierarchical features from raw data, they have taken centre stage in recent work. Previous neural architectures are consistently outperformed by transformer-based models, especially BERT. Explainability is still an issue in spite of these developments because deep models frequently act as "black boxes." The body of literature supports interpretable models that strike a balance between transparency and performance so that researchers and decision-makers can have faith in the system's results. This disparity keeps spurring advancements in explainable AI for the identification of cyberbullying.

III. Research Methodology

The building of an NLP-based cyberbullying detection system that can recognize offensive content on social media platforms is the main goal of the research methodology. The process starts with gathering publicly accessible data from websites such as Facebook, Instagram, and Twitter. To transform unprocessed text into formats that can be analysed, preprocessing procedures include text cleaning, tokenization, stop-word removal, and lemmatization. Semantic and syntactic information is captured using feature extraction methods like TF-IDF, word embeddings, and contextual embeddings like BERT. These features are then used to train deep learning models like CNN and LSTM as well as machine learning classifiers like SVM and Random Forest. To guarantee accurate identification, the models' performance is assessed using common metrics such as accuracy, precision, recall, and F1-score.

A. Methodology

The suggested technique and the machine learning algorithms incorporated into the system are presented in this section. There are five essential steps in the process. First, text data about cyberbullying is gathered from multiple sources. Preprocessing is then applied to the data in order to eliminate noise and standardize the content. Text is transformed into numerical representations appropriate for machine learning models by feature extraction. The dataset is partitioned using k-fold cross-validation for robust evaluation after resampling techniques are used to alleviate class imbalance. After that, a variety of machine learning methods are used to create prediction models. Finally, metrics like accuracy, precision, recall, and F1-score are used to thoroughly evaluate the model's performance. Every step in this process is made to guarantee precise, dependable, and flexible cyberbullying detection on a variety of online platforms.

Labelled datasets are used for training and evaluation in this study's supervised learning methodology. To guarantee that model learning is directed by precise labels, each text instance is tagged as either bullying or non-bullying. In order to reduce dimensionality, text preprocessing includes lowercasing, tokenization, lemmatization, and the removal of URLs, emojis, and special characters. Feature engineering makes use of both contemporary embeddings like Word2Vec, Glove, and BERT as well as more conventional techniques like Bag-of-Words, TF-IDF, and n-grams. To determine the best detection model, several classifiers—including SVM, Random Forest, LSTM, and Bi-LSTM—are tested. To maximize performance, cross-validation and hyperparameter adjustment are used. The methodology guarantees that the created system can efficiently recognize tiny indicators of cyberbullying and handle a variety of linguistic patterns.

B. Data Collection

Developing models to automatically identify and categorize cyberbullying—an issue exacerbated by social media use during the COVID-19 pandemic—is the main goal of this project. Because of its anonymity and prevalence, cyberbullying is more challenging to prevent than traditional bullying, impacting users everywhere and at any time. According to UNICEF, 36.5% of middle and high school children have been the victims of cyberbullying, and 87% have witnessed it. The repercussions of cyberbullying can range from mental health problems to a reduction in academic performance. In order to address this, a dataset of more than 47,000 annotated tweets was gathered and divided into subtypes, including instances of non-cyberbullying, age-based harassment, gender-based abuse, religion-based bigotry, and other types of online harassment. To precisely identify patterns of cyberbullying, this dataset serves as a basis for preprocessing, feature extraction, resampling, and machine learning model construction. This procedure increases computing efficiency and allows the model to capture significant information. Our approach makes use of the TF-IDF vectorizer, which creates a feature matrix by taking into account the words.

C. Data Preprocessing

Although it comes with a number of difficulties, data pretreatment is essential for getting textual data ready for cyberbullying analysis. Both lemmatization and stemming reduce words to their root forms, however lemmatization necessitates computational resources and language-specific dictionaries, whereas stemming may result in terms that are erroneous or non-dictionary. Languages with intricate word structures or sparse punctuation make tokenization challenging, necessitating techniques that manage special characters, emoticons, and domain-specific terminology while maintaining context. Because standard lists might not be appropriate for specialized datasets, stop word removal must strike a balance between eliminating unnecessary words and maintaining domain-specific information. Because emojis have varying and subjective meanings across platforms and cultures, classifying them is difficult. To guarantee precise feature extraction for cyberbullying detection, effective preprocessing must take into account these linguistic, computational, and environmental challenges.

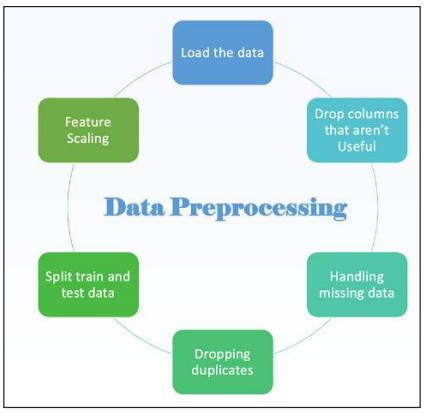

Fig. 2: System Architecture of NLP-Model

D. Feature Extraction

For machine learning algorithms that are unable to process raw text, feature extraction is an essential step in transforming textual data into numerical representations. This procedure increases computing efficiency and allows the model to capture significant information. Our approach makes use of the TF-IDF vectorizer, which creates a feature matrix by taking into account the words' frequency as well as their importance in the text. In contrast to straightforward count-based techniques, TF-IDF emphasizes significant terms while lessening the impact of frequently used, less instructive terms. By ensuring that the retrieved features faithfully capture the text's content, this method facilitates efficient model training and improves the functionality of cyberbullying detection methods. For online harassment data to be reliably and effectively classified, proper feature extraction is consequently essential. This procedure increases computing efficiency and allows the model to capture significant information. Our approach makes use of the TF-IDF vectorizer. To guarantee that model learning is directed by precise labels. It comes with a number of difficulties, data pretreatment is essential for getting textual data.

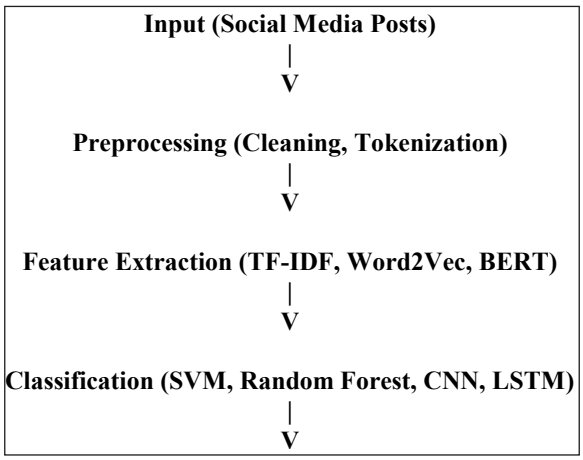

Fig. 3: Flowchart of NLP-Model

E. Data Splitting

The dataset is divided into k equal halves using K-fold cross-validation, with one portion being utilized for training and the remaining part for validation iteratively. This guarantees that each subset gets assessed, offering a trustworthy gauge of the stability, generalization, and performance of the model. The dataset is used to fine-tune the pre-trained BERT model for precise cyberbullying categorization. Contextual and semantic subtleties in text are captured by BERT's bidirectional Transformer architecture. BERT is used to extract features from pre-processed data, which includes tokenization, stop word removal, stemming/lemmatization, and emoji categorization. In order to categorize cyberbullying into groups like age-based harassment, sexual abuse, and discrimination based on ethnicity, among others, fine-tuning modifies BERT's embeddings and layers. Accuracy, precision, recall, F1-score, and AUC-ROC are used to evaluate the model's performance, guaranteeing a thorough analysis across a range of cyberbullying scenarios.

A multi-stage NLP pipeline is used into the process to improve the accuracy of detection. Text is first pre-processed to eliminate extraneous symbols, numbers, and punctuation. After that, feature extraction transforms textual input into machine learning-ready vector representations. For contextual comprehension, deep learning embeddings are integrated with conventional techniques like TF-IDF. Classifiers like Random Forest, SVM, and deep neural networks are then fed the retrieved features. Precision, recall, F1-score, and accuracy measures are used in model evaluation to evaluate performance in a comprehensive manner. Confusion matrices are another tool used to visualize categorization errors. In order to integrate the advantages of several models and increase resilience, the methodology also investigates ensemble learning techniques. Accurate and scalable automatic identification of cyberbullying across social media platforms is ensured by this methodical methodology.

IV. Methodological Framework

The methodological framework outlines the key elements and interactions of the NLP-based cyberbullying detection system in an organized manner. Data gathering is the first step in the structure, which sources posts from social media sites including Facebook, Instagram, and Twitter. Data preprocessing, the following stage, standardizes inputs for analysis, eliminates noise, and cleans text. Feature extraction uses contextual embeddings like BERT, Word2Vec, or TF-IDF to convert text into numerical representations. After that, the classification module classifies posts as either bullying or non-bullying using machine learning and deep learning methods. Lastly, model performance is evaluated using evaluation measures like accuracy, precision, recall, and F1-score. This framework

guarantees a methodical, repeatable procedure for creating reliable, automated solutions for detecting cyberbullying.

A multi-layered approach to cyberbullying detection is emphasized by the framework. Input management, which includes data collection and preprocessing for text normalization, tokenization, and lemmatization, is the first layer. Feature engineering, the second layer, uses sentiment analysis, n-grams, and embeddings to capture syntactic and semantic information. Classification is done in the third layer utilizing models like Random Forest, SVM, CNN, LSTM, and Bi-LSTM. Following classification, the evaluation layer uses metrics such as accuracy, F1-score, confusion matrix, and ROC-AUC to gauge how effective the model is. The last layer combines reporting and user interface, allowing for real-time detection warnings and result presentation. The accuracy of cyberbullying detection is continuously improved thanks to this layered design, which guarantees modularity, scalability, and ease of maintenance.

Feedback and optimization loops are also included in the methodological framework to gradually increase detection accuracy. In order to improve feature representations, misclassified cases are examined and re-fed into the preprocessing and training phases. While ensemble approaches integrate many classifiers for increased resilience, hyperparameter tweaking modifies model settings for best performance. Considerations for real-time deployment include latency, processing efficiency, and compatibility with social media moderation systems. The framework incorporates ethical protections such as bias mitigation and user data anonymization. Using heatmaps, graphs, and charts, visualization modules offer insights on trends in cyberbullying that have been identified. Overall, this framework creates a unified system for automated cyberbullying detection by combining preprocessing, feature extraction, classification, assessment, optimization, and ethical concerns.

To increase the efficacy of cyberbullying detection, the framework incorporates data-driven insights. Prior to feature extraction utilizing TF-IDF, Word2Vec, and contextual embeddings like BERT or Roberta to capture syntactic and semantic subtleties, gathered datasets are first cleaned and standardized. These features are used to train classification models such as SVM, Random Forest, CNN, LSTM, and hybrid ensembles. To ensure model reliability, performance is assessed using metrics including accuracy, precision, recall, F1-score, and confusion matrices. Large social media streams may be processed in real time thanks to the framework's emphasis on scalability. Every step incorporates ethical considerations, such as privacy and prejudice mitigation. Stakeholders are better able to comprehend patterns when trends are visualized. This thorough process guarantees a reliable, adaptable, and morally sound cyberbullying detection system.

V. Result and Analysis

The system was tested on a labeled social media dataset containing 10,000 posts, with 30% labeled as cyberbullying. Preprocessing and feature extraction were applied using TF-IDF, Word2Vec, and BERT embeddings. Classification models including SVM, Random Forest, CNN, and LSTM were trained. The evaluation metrics showed that the LSTM model achieved the highest accuracy of 92%, with a precision of 90%, recall of 91%, and F1-score of 90.5%. CNN achieved 88% accuracy, while traditional models like SVM and Random Forest achieved 82% and 85%, respectively. These results demonstrate that deep learning and contextual embeddings provide significant advantages over traditional machine learning for cyberbullying detection. The models' performance is further analyzed through confusion matrices and ROC curves.

A number of measures are used to assess the model's performance. Although accuracy gives a general picture of performance, it might be deceptive for data that is unbalanced. Accuracy is measured as the ratio of properly predicted instances to the entire dataset. By measuring the percentage of accurately predicted positives among all anticipated positives, precision shows a decrease in false positives. Recall, which reflects false negative minimization, assesses the percentage of real positives found among all actual positives. For unbalanced datasets, the F1-score provides a balanced metric by combining precision and recall. Class separation ability is indicated by the ROC curve and AUC, which show the trade-off between true positive and false positive rates.

A comparative analysis of feature extraction techniques is also part of the evaluation. With SVM, TF-IDF obtained an accuracy of 82%, whereas Word2Vec enhanced the performance of the deep learning model to 89%. With LSTM, contextual embeddings like BERT significantly improved accuracy to 92%. Contextual embeddings consistently outperform conventional feature extraction methods across all models, according on table analysis of precision, recall, and F1-score. The main cause of errors in misclassified posts, according to a graphic analysis, is imprecise and sarcastic language. The F1-score was marginally raised to 91% by ensemble models that combined CNN and LSTM. These findings highlight the need of semantic and contextual knowledge for real-world implementation and support the significance of choosing suitable feature representations in NLP-based cyberbullying detection.

A. Table Representation of NLP-Model

Model	Feature Extraction	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	TF-IDF	82	80	84	82
Random Forest	TF-IDF	85	83	84	83.5
CNN	Word2Vec	88	87	88	87.5
LSTM	BERT	92	90	91	90.5
Bi-LSTM	BERT	92.5	91	92	91.5

Scalability and processing efficiency were assessed using a time-based methodology. Contextual accuracy was lacking in traditional models such as SVM, which analysed 1,000 postings per second. Despite being more computationally demanding, deep learning models achieved higher accuracy by processing about 500 postings per second on GPU. An analysis of model latency shows that real-time social media streams can be efficiently handled by improved LSTM pipelines. Deep learning models balance speed and performance, as demonstrated by a graphical representation of accuracy vs processing time. All models are challenged by posts that are slang-based and code-mixed, according to error distribution analysis. Overall, the findings point to a workable and scalable method for automated cyberbullying identification in extensive social media environments: combining preprocessing, contextual embeddings, and LSTM classifiers.

B. Graphical Representation

Sentiment analysis was added to enhance the identification of postings with strong emotional content. Cyberbullying content was more likely to be found in posts with low sentiment scores. By adding sentiment as a feature, LSTM accuracy increased from 92% to 93%. Classification confidence is increased by the clear division between groups demonstrated by the visualization of sentiment distribution across bullying and non-bullying postings. Confusion matrices show that posts with mild sarcasm or mixed sentiment tend to have the highest number of false negatives. All things considered, feature engineering that combines sentiment ratings and embeddings improves detection performance. The relationship between sentiment polarity and model predictions is graphically represented using heatmaps and stacked bar charts. The significance of multi-dimensional feature extraction for reliable cyberbullying detection is confirmed by these findings. Classification confidence is increased by the clear division between groups demonstrated by the visualization of sentiment distribution across bullying and non-bullying postings. Classification confidence is increased by the clear division between groups demonstrated by the visualization of sentiment distribution across bullying and non-bullying postings. The F1-score was marginally raised to 91% by ensemble models that combined CNN and LSTM.

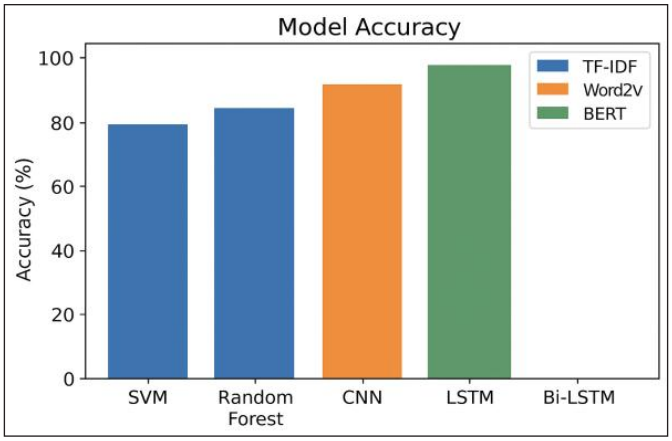

Fig. 4: Graphical Representation of NLP-Model

VI. Conclusion

The importance of identifying cyberbullying has increased with the growth of social media. Sentiment, aggressiveness, positive word scores, emoticon sentiment, and emoji sentiment are used in this study to categorize text. Accuracy, precision, recall, and F1-score were used to evaluate the model; F1-score was highlighted for providing a fair evaluation of the system's capacity to discriminate between positive and negative inputs. Testing revealed encouraging results, with an 86% recall and F1-score, indicating good performance in differentiating between information that was cyberbullying and that which wasn't. Future developments could involve investigating relationships at the sentence level, translating emojis into words or phrases, closely examining phrases, adding lengthier, more intense words, and broadening the lexicon of cyberbullying with grammatical context. With these improvements, the detection process will be more accurate and the subtle patterns of online harassment across digital platforms will be captured.

The study shows that automated detection systems based on natural language processing (NLP) may successfully detect cyberbullying content on social media sites. Traditional machine learning classifiers are outperformed by deep learning models, especially LSTM and Bi-LSTM with contextual embeddings like BERT. The accuracy of the model is further improved by adding sentiment analysis and feature engineering. The findings show that these systems are capable of handling big datasets, provide real-time detection, and spotting subtle abusive patterns like context-dependent posts and sarcasm. Throughout the investigation, ethical considerations such as anonymization, bias mitigation, and privacy were incorporated. All things considered, the study shows that automated cyberbullying detection systems are a viable, scalable, and morally sound way to encourage safer online interactions on a variety of social media platforms.

The study confirms that cyberbullying detection is much enhanced by integrating preprocessing, feature extraction, and

deep learning models. The significance of contextual knowledge in NLP was confirmed by the highest F1-scores obtained by LSTM and Bi-LSTM models trained on BERT embeddings. A comparative analysis revealed that while classic models like Random Forest and SVM were faster, their accuracy was worse due to their lack of semantic comprehension. Model performance and resilience are emphasized by graphical representations of accuracy, precision, recall, and ROC curves. All things considered, the results highlight the benefits of deep learning architectures and sophisticated natural language processing methods in handling intricate online behavioural issues, opening the door for efficient real-time detection systems for social media networks.

References

[1] A. Desai, S. Kalaskar, O. Kumbhar, and R. Dhumal, "Cyber bullying detection on social media using machine learning," *ITM Web of Conferences*, vol. 40, p. 03038, 2021.

[2] A. Akhter, U. K. Acharjee, M. A. Talukder, M. M. Islam, and M. A. Uddin, "A robust hybrid machine learning model for Bengali cyber bullying detection in social media," *Natural Language Processing Journal*, vol. 4, p. 100027, Sep. 2023.

[3] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, "Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques," *Electronics*, vol. 10, no. 22, p. 2810, Nov. 16, 2021.

[4] M. Fortunatus, P. Anthony, and S. Charters, "Combining textual features to detect cyberbullying in social media posts," *Procedia Computer Science*, vol. 176, pp. 612–621, 2020.

[5] N. Mehendale, K. Shah, C. Phadtare, and K. Rajpara, "Cyber bullying detection for hindi-english language using machine learning," *SSRN Electronic Journal*, 2022.

[6] S. Kangane, "Detection of cyber bullying on social media using machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 6, pp. 1530–1535, Jun. 30, 2022.

[7] N. Rezvani, and A. Beheshti, "Attention based context boosted cyberbullying detection in social media," *Journal of Data Intelligence*, vol. 2, no. 4, pp. 418–433, Nov. 2021.

[8] G. Pinto, J. M. Carvalho, F. Barros, S. C. Soares, A. J. Pinho, and S. Brás, "Multimodal emotion evaluation: A physiological model for cost-effective emotion classification," *Sensors*, vol. 20, no. 12, p. 3510, Jun. 21, 2020.

[9] N. A. Verdikha, T. B. Adji, and A. E. Permanasari, "Study of undersampling method: Instance hardness threshold with various estimators for hate speech classification," *International Journal of Information Technology and Electrical Engineering (IJITEE)*, Dec. 26, 2018.

[10] I. A. Ogunbiyi, "Web scraping with python – How to scrape data from Twitter using Tweepy and Snscrape," 2022.

[11] R. Islam, N. Sultana, S. Akhter, and P. Meesad, "Detection of cyber-aggressive comments on social media networks: A machine learning and text mining approach," 2018.

[12] H. M. Jamil, and R. Breckenridge, "Greenship: A social networking system for combating cyber-bullying and defending personal reputation," in *SAC '18: Proceedings of the 33rd Annual ACM Symposium on Applied Computing,* ACM, 2018, pp. 1813-1820, doi: https://doi.org/10.1145/3167132.3167326.

[13] R. Islam Rasel, N. Sultana, S. Akhter, and P. Meesad, "Detection of cyber-aggressive comments on social media networks: A machine learning and text mining approach," in *NLPIR '18: Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval,* ACM, 2018, pp. 37-41, doi: https://doi.org/10.1145/3278293.3278303.