Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 13, Issue. 2, February 2024, pg.1 – 12

Cyberbullying Detection: A Comparative Study of Classification Algorithms

Parthav Nuthalapati[1]; Srinivas Aditya Abbaraju[2]; G. Hemanth Varma[3]; Sitanath Biswas[4]

[1,2,3,4]Department of CSE – AIML & IOT, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering &Technology

[1] parthavnuthalapati2019@gmail.com; [2] asaditya2001@gmail.com; [3] ghemanthvarma2403@gmail.com; [4] sitanathbiswas2006@gmail.com

DOI: https://doi.org/10.47760/ijcsmc.2024.v13i02.001

Abstract— In the realm of social media, cyberbullying's pervasive impact raises urgent concerns about its emotional and psychological toll on victims. This study addresses the imperative of effectively detecting cyberbullying. By leveraging ML and DL techniques, we aim to develop reliable methods that accurately identify instances of cyberbullying in social media data, enhancing detection efficiency and accuracy. This facilitates timely intervention and support for affected individuals. In this comprehensive analysis of existing systems, various ML and DL models are extensively texted for cyberbullying detection. The evaluated models include Random Forest, XgBoost, Naive Bayes, SVM, CNN, RNN, and BERT. Pre-processed datasets are utilized to train and evaluate the models. To evaluate the ability of each model to reliably identify cyberbullying in social media data, performance metrics such as F1 score, recall, precision, and accuracy are used. The findings of this study demonstrate the efficacy of different ML and DL models in monitoring cyberbullying in social media data. Among the models evaluated, the BERT model exhibits exceptional performance, achieving the highest accuracy rates of 88.8% for binary classification and 86.6% for multiclass classification.

Keywords— Natural language processing (NLP), cyberbullying detection, deep learning, comparative analysis, text classification, machine learning

I. INTRODUCTION

In today's world, social media has become an essential part of our ubiquitous communication, allowing people of all ages and backgrounds to connect with each other. However, the widespread use of social media has also led to the emergence of a serious problem: cyberbullying.

The anonymity and reach of the internet provide bullies with the perfect platform to harass and intimidate people, regardless of their location or identity. Unfortunately, the COVID-19 pandemic has only worsened the situation. With schools closed and more people spending time online, cyberbullying has become more prevalent than ever. In fact, UNICEF has warned that the outbreak has resulted in a heightened risk of online harassment and bullying, making it more important than ever to take steps to combat this growing problem.

Cyberbullying is a pervasive issue that involves using digital platforms to harass, intimidate, or threaten individuals or groups. It takes many forms, including excluding or shaming someone online, spreading rumors or false information, sending mean or hurtful messages, and posting humiliating images or videos [1]. The subject of cyberbullying is closely linked to traditional bullying and constitutes a crucial field of inquiry.

The escalation of online harassment is truly alarming: a staggering 36.5% of middle and high school students report being subjected to online intimidation and threats, while 87% admit to witnessing this behaviour.

The repercussions of cyberbullying on its victims can encompass a wide range of negative outcomes, from subpar academic performance and emotional distress to attempts at taking one's own life. Teaching pupils about "internet street smarts", being vigilant for cautionary signs, and offering counselling services are the primary approaches used to prevent online harassment [2].

Although online platforms like Snapchat, Twitter, Facebook, Instagram, and others provide resources and guidelines on cyberbullying, they lack effective measures to combat it. According to studies, about 90% of cyberbullying cases remain unreported. As a result, the development of an efficient user-generated text detection system is imperative in the fight against cyberbullying. Despite the efforts of numerous organizations to raise awareness about cyberbullying, the frequency of online attacks is still on the rise [3].

Around three billion individuals make use of social networking platforms as a means of connecting and interacting with others. While social networking applications like Facebook have undeniable benefits for their users, they can also be utilized for harmful purposes. Using digital technology to harass an individual or a group is known as "online harassment" and is deemed a malignant use of technology. Cyberbullying has a more significant and long-lasting impact compared to traditional bullying, as it can rapidly spread to a vast number of people. Additionally, it can prove challenging or nearly impossible to remove harmful content from online platforms. Cyberbullying has not been shown to result in physical harm to its targets, but it has been associated with negative mental health outcomes such as feelings of hopelessness, poor self-worth, exhaustion, and even suicide efforts [4].

Over the past decade, cyberbullying has become increasingly common, particularly among young people and adolescents. A recent study found that cyberbullying affected 35% of children in India in 2022, while Brazil had a prevalence rate of 30%, followed by the United States at 15%, and the UK at 13%. These research findings suggest that the issue is rapidly growing and is not correlated with a country's level of development. Sweden, which is among the world's most advanced nations, experienced a substantial increase in cyberbullying between 2011 and 2022 [5].

Numerous studies have implemented machine learning methods to detect cyberbullying automatically [6, 7]. However, most research on this topic has been conducted in English. Most of the investigations in this field have utilized text-mining approaches similar to those employed in sentiment analysis research. Social media posts, due to their inherent characteristics, are subject to change and are influenced by the surrounding circumstances. Therefore, it would be inaccurate to view them as independent pieces of text [8]. In this research, we attempt to perform a comprehensive analysis of the already existing systems with our dataset. We aim to compare the effectiveness of different models, including BERT, RNN, CNN, Logistic Regression, SVM, Naive Bayes, XgBoost, and Random Forest. Random Forest, XgBoost, Multinomial Naive Bayes, SVM, Logistic Regression, CNN, RNN, and BERT.

This paper is organized as follows: Section-2 deals with the literature review, Section-3 explains the proposed system, Section-4 describes the dataset, metrics, and experimental analysis as a part of the experimental setup, and the details of the result analysis are demonstrated in Section-5. Finally, Section-6 gives a detailed conclusion about the proposed research. Section-7 provides proper acknowledgment to the people who contributed directly or indirectly to this research.

II. RELATED WORK

Cyberbullying is a growing concern in today's digital world, and various methods have been proposed to detect and combat the issue. Here are a few recent studies related to cyberbullying detection.

Firstly, in their 2018 paper, Al-Ajlan et al. [9] proposed the Optimized Twitter Cyberbullying Detection (OCDD) method, which employed deep learning and metaheuristic optimization algorithm for cyberbullying detection on Twitter. However, it required significant training data and computational resources. Training data was labeled by human intelligence, and word embeddings were generated using GloVe. The resulting word

embeddings were imputed to a CNN for classification. The authors claimed that OCDD outperformed previous approaches, achieving 81.7% accuracy. While CNN has been used in text mining, its application to cyberbullying detection is novel.

In their 2019 study, Mahlangu and colleagues [10] proposed a novel approach to detecting cyberbullying using deep learning techniques. Their method, which employed a stacked embedding approach utilizing both BERT and GloVe embeddings, achieved an impressive accuracy rate of 84.3%. This outperformed traditional ML algorithms like SVM and Logistic Regression. However, one potential limitation of this approach is that it may not perform as well when dealing with code-switching, which occurs when multiple languages are used in the same text data. Despite this limitation, Mahlangu et al.'s study provides valuable insights into the potential of deep learning techniques for detecting cyberbullying and highlights the importance of considering language use in such detection methods.

Vijay Banerjee and his team [11] presented a cyberbullying detection method based on deep neural networks in their 2019 study, which achieved an impressive accuracy of 93.97%. However, despite the system's success, several limitations were identified, including biased data, difficulties in capturing contextual understanding and language dynamics, high false positive rates, and limited generalizability. In response to the limitations of previous research, a new approach to cyberbullying detection is proposed in this paper. The proposed system utilizes a CNN algorithm, which operates through many layers to provide accurate classification, offering a more intelligent way of detecting cyberbullying in comparison to traditional classification algorithms.

Amrita Dewani et al. [12], in 2021, proposed a deep learning architecture to detect cyberbullying in Roman Urdu micro text. The authors addressed the research gap in cyberbullying detection in Roman Urdu by performing extensive preprocessing on the micro text data. This included creating a Roman Urdu slang-phrase dictionary, eliminating cyberbullying domain-specific stop words, and processing unstructured data to handle metadata, non-linguistic elements, and encoded text formats. The authors implemented CNN, RNN-BiLSTM, and RNNLSTM models, varying the number of model layers, epochs, and hyperparameter tuning for extensive experiments. The models' performance and efficiency were evaluated using different metrics, showing that RNN-LSTM achieved an F1 score of 0.7 and a validation accuracy of 85.5% for the aggression class, while RNN-BiLSTM achieved an F1 score of 0.67 and a validation accuracy of 85%.

Luo et al. [13] proposed GCA: BiGRU + CNN + ATTETION, a sentiment classification model for cyberbullying detection, in their 2021 work. The GCA model addresses the limitations of existing models by combining a BiGRU layer for global context, a CNN layer for local features, and an ATTETION mechanism layer for assigning weights to representative words. It was trained and tested on a Kaggle dataset and a social network emoji dataset, achieving an accuracy of 91.07%. However, the authors acknowledged limitations such as lack of explanation, a limited dataset, dataset bias, performance on new data, and computational intensity.

Chahat Raj et al. [14] proposed a novel neural network framework for cyberbullying detection in 2021. They conducted a comparative study on eleven classification methods, including shallow neural networks and traditional machine learning. The study explored feature extraction, word embedding, and algorithmic performance on real-world cyberbullying datasets. Results showed ATTETION models and bidirectional neural networks had high accuracy, with Bi-LSTM and BiGRU performing the best. Logistic Regression and TF-IDF achieved good results. GloVe worked well with neural networks. Limitations included data bias and limited generalization. The study emphasizes the importance of comparing methods and features for improved cyberbullying detection accuracy.

Roy et al. [15], in 2022, introduced a deep transfer learning model for detecting image-based cyberbullying with an 89% accuracy rate. They investigated various transfer learning and deep learning frameworks to identify the most suitable model for predicting cyberbullying in images on social media platforms. The deep learning-based 2D CNN achieved 69.60% accuracy, while the transfer learning models VGG16 and InceptionV3 achieved a higher accuracy of 89%. This proposed system effectively detects most image-based cyberbullying posts. However, the model has some shortcomings. It does not detect textual cyberbullying, and it does not consider text-image combinations in cyberbullying posts.

In a recent study by Mitushi Raj et al. [16] in 2022, a cyberbullying detection system based on a deep learning framework was proposed. Despite achieving an accuracy of 89.5%, the system was found to have limitations such as language constraints, limited scope, and a lack of contextual understanding. The authors addressed these limitations by proposing a model for automatically detecting cyberbullying text in multilingual data. Their study highlights the importance of controlling social media content in multiple languages and

protecting users from the negative impacts of toxic comments. The authors tested various models of neural networks and found that the CNN-BiLSTM network had the best accuracy, This was due to its ability to learn long-term dependencies and global features, which the CNN alone could not achieve with only local characteristics from word n-grams.

In their paper, Accurate Cyberbullying Detection and Prevention on Social Media, Perera et al. [17], 2021 addressed the challenge of detecting cyberbullying on social media platforms, which exploit the dissemination of hatred. Existing solutions lacked accessibility and struggled with the evolving language. The authors proposed an automated system using supervised ML techniques like logistic regression and support vector machines. They considered attributes like malicious intentions, repetitive patterns, and abusive language to efficiently detect and prevent cyberbullying. The system classified cyberbullying into thematic categories, including racism, sexuality, physical harm, and profanity. By integrating feature extraction techniques, the system showed enhanced accuracy and was evaluated using F1-score, precision, and recall metrics.

Desai et al. [18], in their latest study on cyberbullying detection on social media, using ML in 2021, introduced a novel approach to cyberbullying detection by proposing a model that incorporated a wide range of crucial features. They utilized the bidirectional deep learning model BERT to implement selected features, aiming to enhance the effectiveness of cyberbullying detection. Their comprehensive approach considered various aspects of cyberbullying content, enabling more accurate identification and classification. By leveraging BERT's capabilities, the study contributed to mitigating the adverse impact of cyberbullying on individuals' well-being.

TABLE I
A JIST OF LITERATURE REVIEW

Year	Author	Advantages	Drawbacks	Research Gap
2022	Mitushi Raj et al. [1]	Achieved an accuracy of 89.5%, considered bilingual data, and detected cyberbullying text in multiple languages	Language constraints, limited scope and lack of contextual understanding	How to improve the accuracy of the model for detecting cyberbullying text in multiple languages
2022	Roy et al. [2]	Achieved an accuracy rate of 89%, effectively detects most image-based cyberbullying posts	Does not consider textual cyberbullying detection and does not consider text-image combinations in cyberbullying posts	How to improve the model to detect textual cyberbullying and text-image combinations in cyberbullying posts
2021	Desai et al. [3]	Incorporated a wide range of crucial features, enhanced the effectiveness of cyberbullying detection, contributed to mitigating the adverse impact of cyberbullying on individuals' well-being	Limited dataset and dataset bias	How to improve the explanation of the model, how to collect a more comprehensive dataset, and how to address dataset bias
2021	Perera et al. [4]	Enhanced accuracy by integrating feature extraction techniques, consideration of attributes like malicious intentions, repetitive patterns, and abusive language	Data bias and limited generalization	How to reduce data bias and how to improve the generalization of the model

2021	Chahat Raj et al. [5]	High accuracy, good results with Logistic Regression and TF-IDF, and GloVe worked well with neural networks	Data bias and limited generalization	How to compare methods and features for improved cyberbullying detection accuracy
2021	Luo et al. [6]	Achieved an accuracy of 91.07%, combined a BiGRU layer for global context, a CNN layer for local features, and an ATTENTION mechanism layer for assigning weights to representative words	Limited dataset, dataset bias, performance on new data, and computational intensity	How to collect a more comprehensive dataset, how to address dataset bias, how to improve the performance of the model on new data, and how to reduce computational intensity
2021	Amrita Dewani et al. [7]	Performed extensive preprocessing on the micro text data, and evaluated the models' efficiency and performance using different metrics	Limited dataset	How to collect a more comprehensive dataset
2019	Vijay Banerjee et al. [8]	Achieved an impressive accuracy of 93.97%, utilizes a CNN algorithm, offers a more intelligent way of detecting cyberbullying in comparison to traditional classification algorithms	Biased data, difficulties in capturing contextual understanding and language dynamics, high false positive rates, and limited generalizability	How to address biased data, how to improve the contextual understanding and language dynamics of the model, how to reduce false positive rates, and how to improve generalizability of the model
2019	Mahlangu et al. [9]	Employed a stacked embedding approach utilizing both BERT and GloVe embeddings, achieved an impressive accuracy rate of 84.3%, outperformed traditional ML algorithms like Support Vector Machines (SVM) and Logistic Regression	May not perform as well when dealing with code-switching	How to improve the performance of the model when dealing with code-switching
2018	Al-Ajlan et al. [10]	Employed deep learning and meta-heuristic optimization algorithm for cyberbullying detection on Twitter, achieved an accuracy of 81.7%, CNN has been used in text mining, but its application to cyberbullying detection is novel	Requires significant training data and computational resources	How to reduce the need for significant training data and computational resources

The tabular representation of the currently existing systems can be found in Table I. The research conducted so far has yielded promising results, but there is still ample room for improvement in terms of model effectiveness, generalizability, and scalability. Further research is needed to achieve more reliable and effective cyberbullying detection models. By utilizing advanced techniques and implementing a continuous development methodology, researchers can create precise, effective, and scalable models that can aid in preventing and addressing the harmful impacts of cyberbullying. Such models have the potential to play a crucial role in addressing the issue of cyberbullying and mitigating its harmful effects.

III. PROPOSED SYSTEM

The proposed system for detecting cyberbullying in social media data is designed to achieve high accuracy in identifying different types of cyberbullying. The system utilizes pre-processing techniques with the NLTK module to accomplish this, including stop word removal, stemming, and lemmatization of the text data. The pre-processed data is then classified into one of the six categories under the “cyberbullying_type” column, including not_cyberbullying, gender, religion, other_cyberbullying, age, and ethnicity. This categorization aids in a better

understanding of the different types of cyberbullying present in the dataset, which can aid in the development of focused preventative efforts. The proposed system uses a cyberbullying detection model that is illustrated in Fig. 1. The figure shows the different components and layers of the model, which are used to accurately classify cyberbullying instances in social media data.

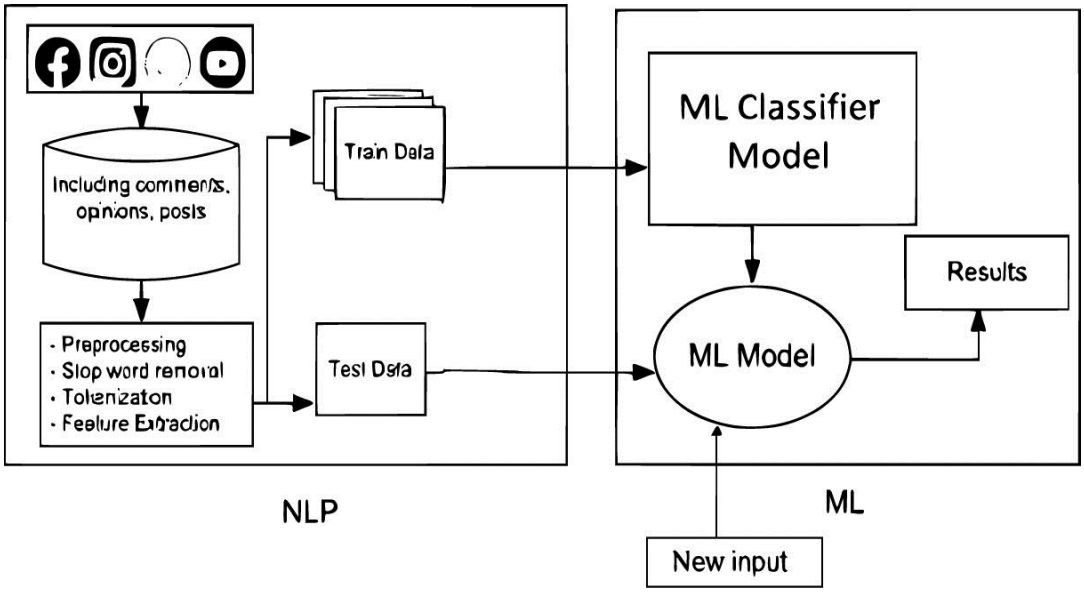

Fig. 1 Architecture of Proposed System

We have employed text classification task-appropriate machine learning techniques to accomplish the categorization. It includes a comprehensive comparative analysis of existing models, including BERT, RNN (Bi-LSTM), CNN, Multinomial Naive Bayes, Logistic Regression, SVM, XGBoost, and Random Forest. By evaluating the performance of these models on our pre-processed dataset, we aim to identify the most accurate and effective model for detecting cyberbullying.

The BERT model is a state-of-the-art language model that is well-suited for natural language processing tasks, including text classification. RNN (Bi-LSTM) is a recurrent neural network that can handle sequence data, making it a good choice for text classification tasks. CNN is a convolutional neural network that can be used for text classification tasks. Random Forest, SVM, Multinomial Naive Bayes, XGBoost, and Logistic Regression, are classical machine learning algorithms that have been used in text classification tasks and have shown good performance.

We perform a comparative analysis on various models in our pre-processed dataset to evaluate their F1 score, precision, recall, and accuracy, recall, precision, and F1 score. The best-performing model is then employed to create a cyberbullying detection system that can accurately identify cyberbullying in social media data. This technique has the potential to prevent cyberbullying before it occurs, thereby contributing to a safer online environment for everyone.

IV. Experimental Setup

A. Dataset

The dataset used in this study on cyberbullying detection is sourced from Kaggle and comprises a total of 47,692 tweets. The dataset consists of two primary columns, namely tweet_text and cyberbullying_type. The "tweet_text" column contains the actual text of the tweet, while the "cyberbullying_type" column indicates the type of cyberbullying present in the tweet.

The size of the dataset is quite substantial, which makes it ideal for training machine learning models for detecting cyberbullying. With a large dataset like this, we can ensure that the model can generalize well to new, unseen data. With the dataset being quite large, we have the ability to conduct a thorough analysis of the trends and patterns of cyberbullying on social media.

B. Metrics

We employed four common evaluation measures to assess how well the machine learning models performed in identifying cyberbullying: accuracy score, F1 score, Precision, and Recall. These metrics offer a thorough assessment of the algorithms' ability to distinguish tweets that are not cyberbullying from those that are.

TABLE II
CONFUSION MATRIX

Actual Class	**Predicted Class**	
	Positive	**Negative**
Positive	TP	FP
Negative	FN	TN

A confusion matrix, which can be found in Table II, is a table that summarizes the performance of a classification model by showing the predicted and actual classifications for each data point. The diagonal elements of the matrix contain the correctly predicted results, while the rest represent inaccurate classifications. The entries in the fields represent the following: True Positive (TP) refers to the number of tweets that were correctly identified as cyberbullying, while True Negative (TN) refers to the number of tweets that were correctly identified as not cyberbullying. False Positive (FP) refers to the number of tweets that were incorrectly identified as cyberbullying, and False Negative (FN) refers to the number of tweets that were incorrectly identified as not cyberbullying.

1) Accuracy Score: The percentage of tweets in the dataset that were successfully categorized makes up the accuracy score. This is how the accuracy score is determined:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \dots (1)$$

2) Precision: Precision refers to the proportion of correct positive predictions out of all the positive predictions made. The formula used to calculate precision is as follows:

$$Precision = \frac{TP}{TP+FP} \quad \dots (2)$$

3) Recall: Recall is the proportion of genuine positive incidents that are correctly identified as positive. The formula is as follows:

$$Recall = \frac{TP}{TP+FN} \quad \dots (3)$$

4) F1 score: F1 score is a metric that is calculated by taking the harmonic mean of recall and precision. The formula for the F1 score is as follows:

$$F1\ Score = \frac{2*precision*recall}{precision+recall} \quad \dots (4)$$

C. Experimental Analysis

In the experimental analysis of cyberbullying detection, a diverse set of models was employed. Each model had its own unique advantages and a proven track record in natural language processing tasks. These models included BERT, RNN (Bi-LSTM), CNN, Logistic Regression, SVM, Multinomial Naive Bayes, XgBoost, and Random Forest.

1) BERT: BERT, a bidirectional encoder model from transformers, is a game-changer in natural language processing. It leverages a bidirectional transformer architecture to capture contextual information effectively. The study conducted experiments involving fine-tuning BERT for cyberbullying detection, where important parameters such as epochs, batch size, and learning rate were adjusted.

2) RNN (Bi-LSTM): RNNs with bidirectional long short-term memory (Bi-LSTM) architectures are advantageous for processing sequential data because they can capture long-range dependencies. This model effectively captures contextual information and long-term dependencies, making it suitable for analysing text data to identify instances of cyberbullying.

3) CNN: Convolutional Neural Network (CNN) is widely recognized for its success in image analysis, but its applicability extends to text classification as well. Through the utilization of filters and pooling operations, CNNs excel in extracting valuable features from textual input. This study delved into the exploration of CNNs for the purpose of cyberbullying detection, with specific emphasis on their proficiency in recognizing patterns and contextual details.

4) Logistic Regression: Logistic Regression, a conventional statistical model widely applied in binary classification tasks, estimates the likelihood of a sample belonging to a specific class. In this study, Logistic Regression was utilized as a reference model to gauge its performance against more intricate deep-learning models.

5) SVM: Support Vector Machine (SVM) is a widely employed supervised machine learning algorithm utilized for text classification. By constructing hyperplanes to distinguish data points into distinct classes, SVM proves useful in various domains. In this study, an SVM model is constructed utilizing extracted features from the text data.

6) Multinomial Naïve Bayes: Multinomial Naive Bayes is a probabilistic ML model that is based on Bayes' theorem. It assumes that the features of a text are conditionally independent given the class label, which makes it suitable for text classification tasks involving discrete features.

7) XgBoost: XgBoost, a gradient-boosting algorithm, constructs a robust ensemble model by combining multiple weak classifiers. It has garnered recognition for its efficiency and efficacy across diverse machine-learning applications. The evaluation conducted aimed to gauge XgBoost's performance specifically in the realm of cyberbullying detection.

8) Random Forest: Random forest is an ensemble learning method that creates a multitude of decision trees to make more accurate predictions. By reducing overfitting and yielding robust classification outcomes, Random Forest was evaluated as an additional ensemble model for detecting instances of cyberbullying.

V. RESULT ANALYSIS

The performance of binary machine learning models for binary classification using stemming is shown in Fig. 2 and Table III. The BERT model achieved the highest accuracy of 88.4%, followed by SVM (86.1%), Multinomial Naïve Bayes (85.6%), RNN (Bi-LSTM) and CNN (both 84%), Logistic Regression (83.8%), and XgBoost (82.9%). Overall, the BERT model achieved the highest accuracy for binary classification using stemming, and SVM scored considerably better than the rest except BERT.

TABLE III
ANALYSIS OF THE PERFORMANCE OF BINARY (ML MODELS) FOR BINARY CLASSIFICATION USING STEMMING

ANALYSIS OF THE PERFORMANCE OF BINARY (ML MODELS) FOR BINARY CLASSIFICATION USING STEMMING	
Model	**Accuracy**
BERT	88.4%
RNN (Bi-LSTM)	83.8%
CNN	84%
Logistic Regression	83.8%
SVM	86.1%
Multinomial Naïve Bayes	85.6%
XgBoost	82.9%
Random Forest	84%

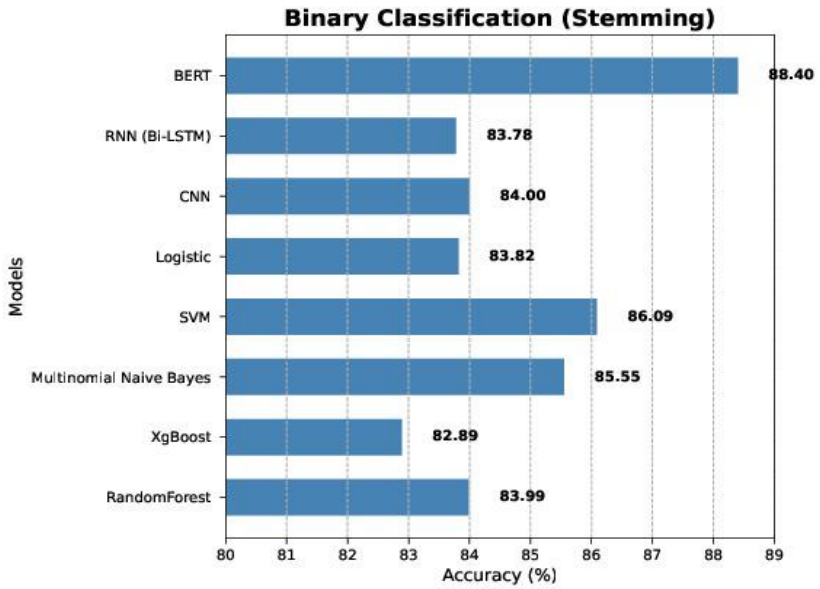

Fig. 2 Graphical representation of the Analysis of the performance of Binary (ML models) for Binary classification using Stemming

The performance of binary machine learning models for multiclass classification using stemming is shown in Fig. 3 and Table IV. The BERT model achieved the highest accuracy of 85.3%, followed by XgBoost (84.7%), SVM (83.1%), Random Forest (82.%), CNN (81.3%), RNN Bi-LSTM (79.2%), and Multinomial Naïve Bayes (76.3%). Overall, the BERT model achieved the highest accuracy for multiclass classification using stemming, and XgBoost scored considerably better than the rest except BERT.

TABLE IV
ANALYSIS OF THE PERFORMANCE OF BINARY (ML MODELS) FOR MULTICLASS CLASSIFICATION USING STEMMING

ANALYSIS OF THE PERFORMANCE OF BINARY (ML MODELS) FOR MULTICLASS CLASSIFICATION USING STEMMING	
Model	**Accuracy**
BERT	85.3%
RNN (Bi-LSTM)	79.2%
CNN	81.3%
SVM	83.1%
Multinomial Naïve Bayes	76.3%
XgBoost	84.7%
Random Forest	82.7%

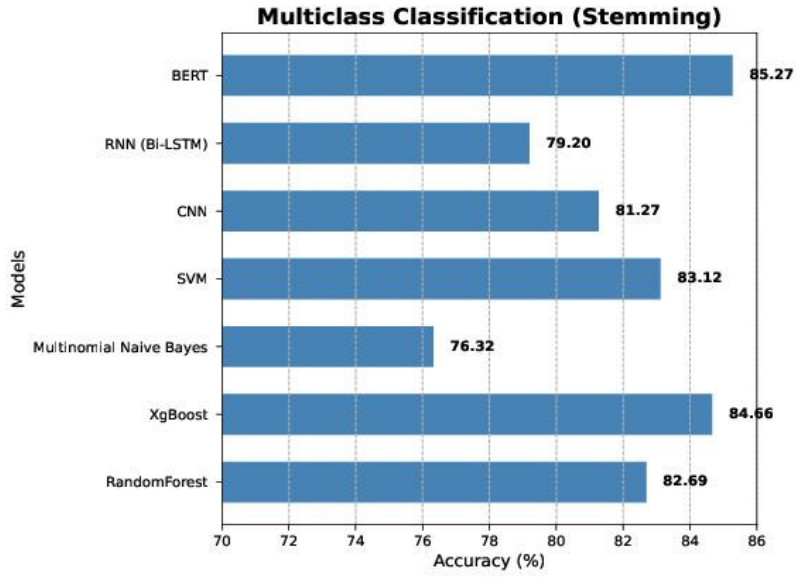

Fig. 3 Graphical representation of the Analysis of the performance of Binary (ML models) for Multiclass classification using Stemming

The performance of binary machine learning models for binary classification using lemmatization is shown in Fig. 4 and Table V. The SVM model achieved the highest accuracy of 86.2%, followed by CNN (85.9%), Multinomial Naïve Bayes (85.6%), RNN Bi-LSTM (84.8%), Logistic Regression (84%), Random Forest (83.9%), and XgBoost (82.8%). Overall, the SVM model achieved the highest accuracy for binary classification using lemmatization, and CNN scored considerably better than the rest except SVM.

TABLE V
ANALYSIS OF THE PERFORMANCE OF BINARY (ML MODELS) FOR BINARY CLASSIFICATION USING LEMMATIZATION

ANALYSIS OF THE PERFORMANCE OF BINARY (ML MODELS) FOR BINARY CLASSIFICATION USING LEMMATIZATION	
Model	**Accuracy**
RNN (Bi-LSTM)	84.8%
CNN	85.9%
Logistic Regression	84%
SVM	86.2%
Multinomial Naïve Bayes	85.6%
XgBoost	82.8%
Random Forest	83.9%

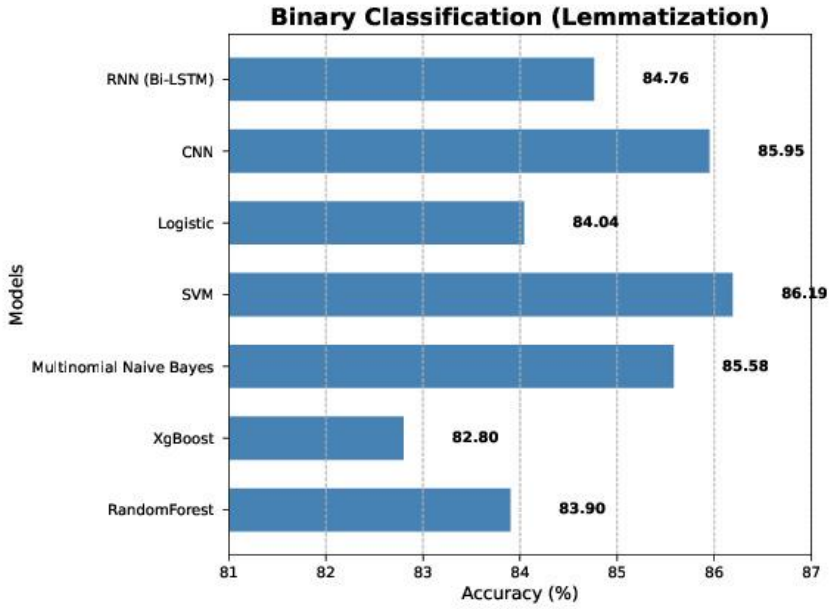

Fig. 4 Graphical representation of the Analysis of the performance of Binary (ML models) for Binary classification using Lemmatization

The performance of binary machine learning models for multiclass classification using lemmatization is shown in Fig. 5 and Table VI. The XgBoost model achieved the highest accuracy of 84.6%, followed by SVM (83.2%), Random Forest (82.9%), RNN Bi-LSTM (81.1%), CNN (81%), and Multinomial Naïve Bayes (76.6%). Overall, the XgBoost model achieved the highest accuracy for multiclass classification using lemmatization, and SVM scored considerably better than the rest except XgBoost.

TABLE VI
ANALYSIS OF THE PERFORMANCE OF BINARY (ML MODELS) FOR MULTICLASS CLASSIFICATION USING LEMMATIZATION

ANALYSIS OF THE PERFORMANCE OF BINARY (ML MODELS) FOR MULTICLASS CLASSIFICATION USING LEMMATIZATION	
Model	**Accuracy**
RNN (Bi-LSTM)	81.1%
CNN	81%
SVM	83.2%
Multinomial Naïve Bayes	76.6%
XgBoost	84.6%
Random Forest	82.9%

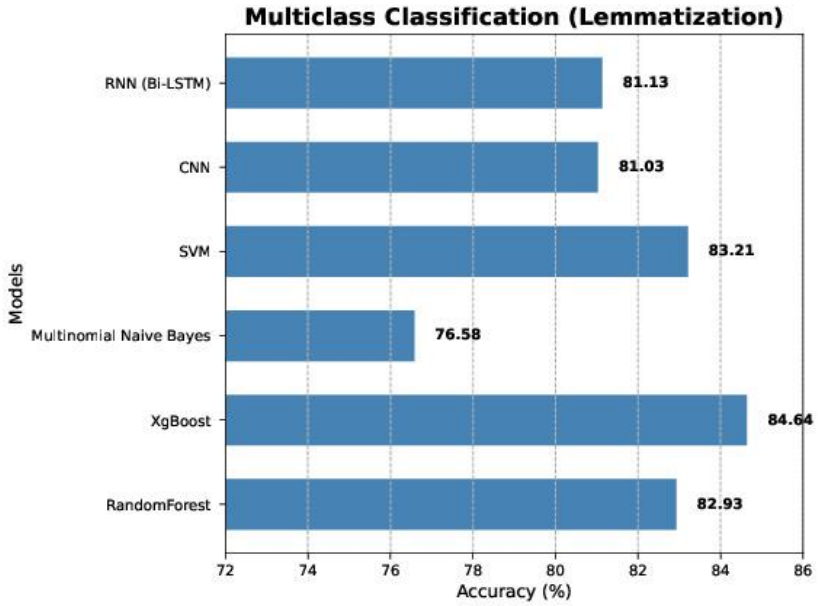

Fig. 5 Graphical representation of the Analysis of the performance of Binary (ML models) for Multiclass classification using Lemmatization

VI. CONCLUSION

The pervasive nature of cyberbullying on social media platforms has become a pressing concern, as it can have severe emotional and psychological consequences for victims. This study addresses the need for effective cyberbullying detection by leveraging ML and DL techniques to enhance the efficiency and accuracy of cyberbullying detection, facilitating timely intervention and support for affected individuals. In this comprehensive analysis of existing systems, various ML and DL models were extensively tested for cyberbullying detection. The evaluated models include Random Forest, XgBoost, Naive Bayes, SVM, CNN, RNN, and BERT. Pre-processed datasets were utilized to train and evaluate the models. To evaluate the ability of each model to reliably identify cyberbullying in social media data, performance metrics such as F1 score, recall, precision, and accuracy were used. The findings of this study demonstrate the efficacy of different ML and DL models in monitoring cyberbullying in social media data. Among the models evaluated, the BERT model exhibited exceptional performance, achieving the highest accuracy rates of 88.8% for binary classification and 86.6% for multiclass classification. These results demonstrate the potential of advanced NLP techniques to address the widespread problem of cyberbullying. This implementation of pipeline and SMOTE techniques also helped to address the imbalanced nature of the dataset. In conclusion, this research provides a foundation for further research and development in the field of NLP and its applications in combating online harassment. The BERT model and the proposed approach can be used to create robust and accurate cyberbullying detection systems, making social media platforms safer for users.

ACKNOWLEDGEMENT

We are deeply grateful to all individuals who have contributed to the successful completion of this project on Cyberbullying Detection. We would like to extend our thanks to the authors, J. Wang, K. Fu, and C. T. Lu for generously providing us with the Cyberbullying Tweets dataset. The support and guidance offered by our mentors and colleagues throughout this project have been invaluable.

Furthermore, we would like to express our appreciation to VNRVJIET, Hyderabad, for their generous provision of the essential resources and infrastructure that enabled us to conduct this research. We are also grateful to the Department of Computer Science and Engineering - AIML and IOT faculty members for their invaluable feedback and unwavering support throughout the project.

REFERENCES

[1] Craig, W., Boniel-Nissim, M., King, N., Walsh, S. D., Boer, M., Donnelly, P. D., Harel-Fisch, Y., Malinowska-Cie´slik, M., Gaspar de Matos, M., Cosma, A., Van den Eijnden, R., Vieno, A., Elgar, F. J., Molcho, M., Bjereld, Y., Pickett, *Social Media Use and Cyber-Bullying: A Cross-National Analysis of Young People in 42 Countries*. In: Journal for Adolescent Health (2020).

[2] Ferrara, P., Ianniello, F., Villani, A., Corsello, G.: *Cyberbullying: A Modern Form of Bullying - Let's Talk About This Health and Social Problem*. In: Italian Journal of Pediatrics (2018).

[3] Peebles, E.: *Cyberbullying: Hiding Behind the Screen*. In: Paediatrics & Child Health (2014).

[4] Sathyanarayana Rao, T. S., Bansal, D., Chandran, S.: *Cyberbullying: A Virtual Offense with Real Consequences*. In: Indian Journal of Psychiatry (2018).

[5] Nixon, C.: *Current Perspectives: The Impact of Cyberbullying on Adolescent Health*. In: Adolescent Health, Medicine, and Therapeutics (2014).

[6] Wan Ali, W. N. H., Mohd, M., Fauzi, F.: *Cyberbullying Detection: An Overview*. In: 2018 Cyber Resilience Conference (CRC) (2018).

[7] Salawu, S., He, Y., Lumsden, J.: *Approaches to Automated Detection of Cyberbullying: A Survey*. In: IEEE Transactions on Affective Computing (2020).

[8] Habeeb Ur Rahman, M.: *Cyberbullying Detection Using Natural Language Processing*. In: International Journal for Research in Applied Science and Engineering Technology (2022).

[9] Al-Ajlan, M. A., Ykhlef, M.: *Optimized Twitter Cyberbullying Detection Based on Deep Learning*. In: 2018 21st Saudi Computer Society National Computer Conference (NCC) (2018).

[10] Mahlangu, T., Tu, C.: *Deep Learning Cyberbullying Detection Using Stacked Embbedings Approach*. In: 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI) (2019).

