Advancing Cyberbullying Detection: A Hybrid Machine Learning and Deep Learning Framework for Social Media Analysis

Bishal Shyam Purkayastha[1], Md. Musfiqur Rahman[2], Md. Towhidul Islam Talukdar[3] and Maryam Shahpasand[1]

[1]*Computer Science (Cyber Security), University of Staffordshire London, London, U.K.*

[2]*Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong, Bangladesh*

[3]*Department of Mechanical Engineering, Chittagong University of Engineering and Technology, Chittagong, Bangladesh*

Keywords: Automated Cyberbullying Detection, Transformer-Based Models, Social Media Text Analysis.

Abstract: Social media platforms have led to the prevalence of cyberbullying, seriously challenging the mental health of individuals. This research is on how effectively different machine learning and deep learning techniques can detect cyberbullying in online communications. Using two different tweet datasets obtained from Mandalay and Kaggle, we developed a balanced framework for binary classification. This research emphasizes comprehensive data preprocessing: text normalization and class balancing by random oversampling to increase the dataset's quality. Models used include several traditional machine learning classifiers: Random Forest, Extra Trees, AdaBoost, MLP, and XGBoost, and advanced deep learning architectures such as Bidirectional LSTM, BiGRU, and BERT. These results confirm that deep learning models, especially BERT, yield outstanding performance with an accuracy rate of 92%, hence showing the models' capability in effectively detecting and preventing cyberbullying through automated detection.

1 INTRODUCTION

Bullying is a deliberate act of aggression where individuals exploit their social or physical dominance to harm others, often targeting those who are less powerful. It manifests in various forms—verbal, physical, or social—and inflicts significant emotional and psychological suffering on victims. Cyberbullying, an extension of this behavior into digital environments, has emerged as a pressing social concern with the widespread use of the internet and social media platforms. Studies reveal that approximately 37% of young individuals in India have experienced cyberbullying, with 14% enduring chronic instances (Arif, 2021). This form of digital hostility manifests as harassment, intimidation, or public humiliation through social media and other online channels, profoundly affecting victims' mental health, academic performance.

The evolution of cyberbullying has closely followed advancements in technology depicted in Figure 1. In the 1990s, it began in internet forums and chat rooms, where anonymity enabled verbal harassment and rumor-spreading. The 2000s saw the rise of social networks like Myspace and Friendster, amplifying the impact of bullying through public humiliation and impersonation. By the 2010s, anonymous messaging apps such as Sarahah and Ask.fm further complicated the issue, allowing bullies to attack without fear of accountability. The late 2010s introduced deepfake manipulation, enabling bullies to create and share false, humiliating content. More recently, the 2020s have seen the emergence of AI-powered harassment, automating and intensifying bullying attacks, making them harder to monitor and counteract.

As technology advances, so do the tactics employed by cyberbullies, underscoring the urgent need for effective intervention mechanisms. Machine learning (ML) and deep learning (DL) methods have emerged as crucial tools in the fight against cyberbullying, offering the ability to analyze large volumes of text and multimedia data to identify patterns of abusive behavior (Bruwaene et al., 2020).

In response to these challenges, this study conducts a comprehensive evaluation of ML- and DL-based approaches for detecting and classifying cy-

Purkayastha, B. S., Rahman, M. M., Talukdar, M. T. I. and Shahpasand, M.
Advancing Cyberbullying Detection: A Hybrid Machine Learning and Deep Learning Framework for Social Media Analysis.
DOI: 10.5220/0013436200003929

In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 2*, pages 348-355
ISBN: 978-989-758-749-8; ISSN: 2184-4992

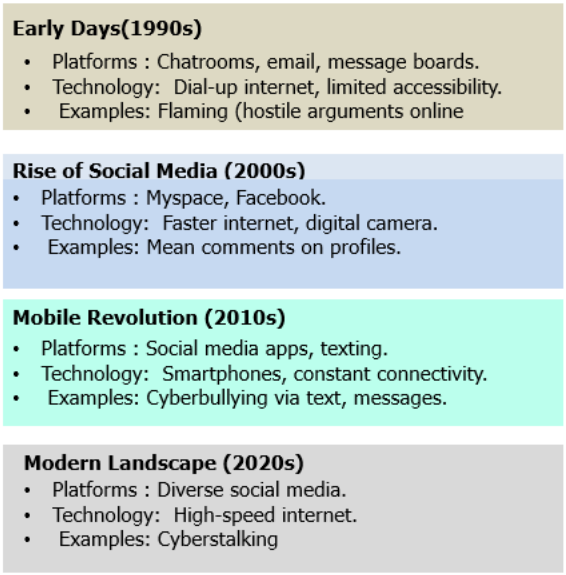

Figure 1: Evaluation of cyberbullying.

berbullying incidents. By leveraging two authentic datasets, the research develops and validates a binary classification framework that identifies cyberbullying occurrences within social media content. Rigorous data preprocessing, feature engineering, and advanced transformer-based models are employed to assess the strengths and limitations of various detection methods, demonstrating their potential to mitigate the growing issue of cyberbullying.

The main contributions of our work are:

1. Integrated two real-world datasets from Mandalay and Kaggle, addressing class imbalances through random oversampling to create a balanced binary classification framework.
2. Developed an effective data preprocessing and feature engineering workflow, enhancing the quality and reliability of input data for model training and evaluation.
3. Conducted an extensive evaluation of machine learning and deep learning models, highlighting BERT's superior performance with an accuracy of 92%.

2 RELATED WORK

Cyberbullying detection has garnered significant attention in recent years, with diverse approaches leveraging machine learning (ML) and deep learning (DL) methods to identify harmful behaviors in online communications. In (Balakrishnan et al., 2020), proposed an automatic detection method that utilizes the psychological characteristics of Twitter users for feature extraction and classification. The authors tested Random Forest (RF) and J48 classifiers using a dataset of 5,453 tweets, achieving promising results. Similarly, (Yadav et al., 2020) employed the BERT model, a transformer-based architecture, which effectively generated contextual embeddings and achieved reliable outcomes in detecting cyberbullying.

In paper (Dalvi et al., 2020), explored traditional approaches using TF-IDF vectorization combined with Naive Bayes (NB) and Support Vector Machines (SVM) for tweet classification, where SVM outperformed NB with an accuracy of 71.25%. In addition (Hani et al., 2019), implemented a supervised learning method incorporating n-gram language models and sentiment analysis, emphasizing feature extraction techniques to improve performance. In (Al-Ajlan and Ykhlef, 2018), introduced Optimized Twitter Cyberbullying Detection (OCDD), leveraging convolutional neural networks (CNNs) and Glove embeddings, further enhanced by meta-heuristic optimization methods.

Recent works emphasize integrated frameworks and ensemble methods. For example, (Unnava and Parasana, 2024) compared various ML techniques such as NB, k-Nearest Neighbors (kNN), Decision Tree (DT), RF, and SVM, demonstrating notable performance improvements with feature engineering. In (Atoum, 2020), highlighted that utilizing n-grams in NB outperformed SVM for Twitter-based datasets. Moreover (Mehendale et al., 2022), explored offensive language detection in multilingual datasets using Natural Language Processing (NLP) and ML techniques. Similarly, (Yuvaraj et al., 2021) proposed an integrated model incorporating user context, psychometric properties, and CB classification. Emerging methods focus on addressing the challenges posed by cyberbullying detection, including data imbalance, context analysis, and multimodal approaches. In (Roy and Mali, 2022), proposed a deep transfer learning model for image-based cyberbullying detection, achieving an accuracy of 89%, though with limited attention to textual data.

Despite advancements, critical gaps remain, including the development of scalable, real-time solutions and methods that generalize across modalities and languages.

3 CLASSIFICATION OF CYBERBULLYING

Cyberbullying encompasses various forms of harassment, each employing distinct tactics and method-

ologies depicts in Figure 2. Direct cyberbullying includes abusive messages, threats, and dissemination of false information. Cyberstalking involves the intimidation of victims through persistent surveillance and excessive communication. Flaming refers to heated and offensive exchanges in public forums. Banning pertains to the exclusion of individuals from online groups, often coupled with impersonation through the creation of fake profiles. Outing and fraud involve the unauthorized disclosure of personal information, while proxy cyberbullying leverages third parties to harass victims. Lastly, catfishing describes the emotional manipulation and exploitation of individuals using fabricated identities.

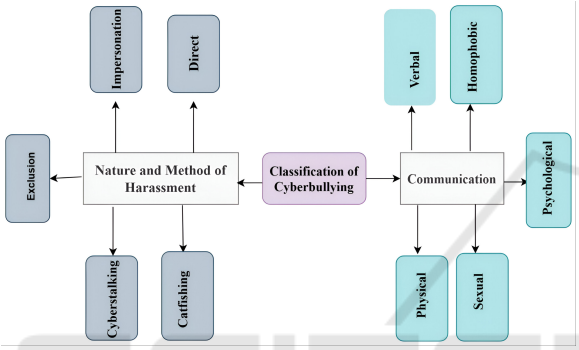

Figure 2: Classification of cyberbullying based on the nature and method of harassment.

Cyberbullying can be further categorized by the methods and media employed. Verbal cyberbullying uses harmful language, such as insults or threats via messages or comments. Physical cyberbullying involves unauthorized access to online accounts. Social cyberbullying occurs on platforms where false information is spread, or individuals are excluded. Psychological cyberbullying employs manipulation to inflict emotional distress. Sexual cyberbullying entails the dissemination of explicit content without consent, and homophobic, racist, or religious cyberbullying targets individuals based on identity, race, or beliefs, often leveraging hateful language or discrimination.

4 METHODOLOGY

This research employs a systematic approach, beginning with the collection of two open-source datasets from Mandalay and Kaggle, followed by preprocessing to enhance data quality. The processed dataset is then subjected to machine learning (ML) algorithms for cyberbullying detection. The workflow is depicted in Figure 3.

4.1 Dataset Collection

Two open-source datasets were utilized: a multiclass-labeled dataset from Mandalay - Cyberbullying Tweets, (Mehendale et al., 2022) and a binary-labeled dataset from Kaggle - Cyberbullying Classification (Kaggle, 2024). To create a binary-class hybrid dataset, relevant classes were relabeled such that all instances of cyberbullying were labeled as 1 and non-cyberbullying instances as 0. Due to class imbalance, random oversampling was employed, resulting in a balanced dataset comprising 90,276 records. This dataset was subsequently used to train ML models for cyberbullying detection.

4.2 Dataset Preprocessing

Preprocessing steps included the removal of hyperlinks, punctuation, extra spaces, and stop words to improve data quality. All text was converted to lowercase to address case inconsistencies, and non-English words were translated to English. Random oversampling was applied to balance the classes, ensuring equal representation of cyberbullying and non-cyberbullying instances. These steps enhanced the dataset's suitability for machine learning analysis.

5 BULLYING DETECTION MODEL

The proposed framework for cyberbullying detection integrates Natural Language Processing (NLP) and Machine Learning (ML). The methodology is categorized into two major components: neural network-based approaches and classical machine learning algorithms.

5.1 Neural Network Approaches

5.1.1 Bidirectional Long Short-Term Memory (BiLSTM)

The BiLSTM model processes sequences in both forward and backward directions, enabling a comprehensive understanding of contextual relationships in text data. For a given input sequence $X = [x_1, x_2, \ldots, x_T]$:

$$\overrightarrow{h}_t = \sigma(W_x x_t + W_h \overrightarrow{h}_{t-1} + b_h) \quad (1)$$

$$\overleftarrow{h}_t = \sigma(W_x x_t + W_h \overleftarrow{h}_{t+1} + b_h) \quad (2)$$

Here, σ is the activation function, W_x and W_h are weight matrices, and b_h is the bias term. The final

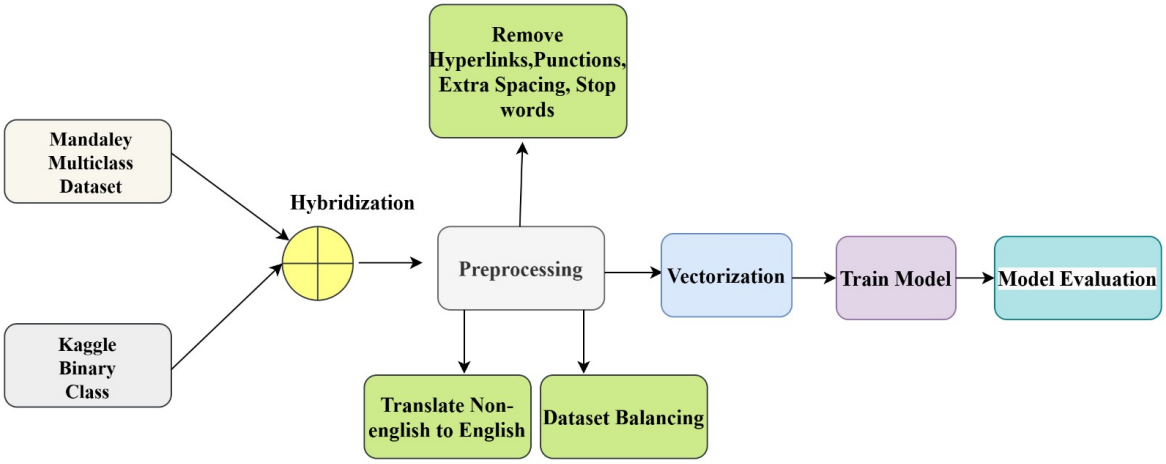

Figure 3: Classification of cyberbullying based on nature and method of harassment.

hidden state is the concatenation of forward and backward states:

$$\mathbf{h}_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \tag{3}$$

This bidirectional approach ensures that the model captures dependencies from both past and future contexts, making it particularly effective for sentiment analysis and text classification tasks (Gada et al., 2021).

5.1.2 Gated Recurrent Unit (GRU)

The GRU simplifies Long Short-Term Memory (LSTM) models by using fewer gates while maintaining performance. The update and reset gates are computed as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{4}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{5}$$

The hidden state is updated as:

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{6}$$

where $\tilde{h}_t = \tanh(W_h x_t + r_t \odot (U_h h_{t-1}))$. GRUs are computationally efficient and work well for sequential data tasks like language modeling and named entity recognition (Fang et al., 2021).

5.1.3 Bidirectional Encoder Representations from Transformers (BERT)

BERT utilizes self-attention mechanisms to generate context-aware embeddings by processing text bidirectionally. For an input sequence X:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{7}$$

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

Here, Q, K, and V represent the query, key, and value matrices, and d_k is the dimensionality of the keys. BERT employs masked language modeling (MLM) and next sentence prediction (NSP) for pre-training, enabling it to understand sentence relationships and word context efficiently (Paul and Saha, 2022). It is particularly powerful for tasks like sentiment analysis and query answering.

5.2 Classical Machine Learning Algorithms

5.2.1 Naive Bayes Classifier

The Naive Bayes classifier is a probabilistic approach based on Bayes' theorem. Given a feature vector $X = [x_1, x_2, \ldots, x_n]$, the probability of class C is computed as:

$$P(C|X) = \frac{P(C)\prod_{i=1}^{n} P(x_i|C)}{P(X)} \tag{9}$$

The decision rule is:

$$C^* = \arg\max_C P(C)\prod_{i=1}^{n} P(x_i|C) \tag{10}$$

This classifier assumes independence among features, making it simple yet effective for text classification tasks, especially with well-engineered features (Paul and Saha, 2022).

5.2.2 Random Forest

The Random Forest algorithm operates as an ensemble of decision trees. For a dataset D, the classification decision is obtained by aggregating predictions from T trees:

$$F(x) = \text{majority\_vote}(T_1(x), T_2(x), \ldots, T_T(x)) \tag{11}$$

Each tree is trained on a bootstrap sample of the dataset, and feature selection during tree construction introduces diversity. The majority vote determines the final class label (Raj, 2021).

5.2.3 XGBoost

XGBoost (Extreme Gradient Boosting) optimizes gradient boosting by minimizing a regularized loss function. Given the prediction $\hat{y}_i$, the objective is:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{12}$$

where $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$, T is the number of leaves, and w are the leaf weights. This algorithm is particularly effective for structured data (Raj, 2021).

5.2.4 AdaBoost

AdaBoost combines weak classifiers to form a strong classifier. For each iteration t:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - e_t}{e_t} \right) \tag{13}$$

where e_t is the weighted error rate. The updated weights for misclassified samples emphasize their importance in subsequent iterations (Raza et al., 2020).

5.3 Feature Engineering

Feature engineering is crucial for distinguishing between cyberbullying and neutral content. Measurable attributes such as lexical and syntactic features, demographic details, and sentiment scores are incorporated. The sentiment score for a document is computed as:

$$\text{Sentiment Score} = \sum_{i=1}^{n} w_i \cdot s_i \tag{14}$$

where w_i is the weight assigned to term i, and s_i is the sentiment score of term i. Effective feature selection and generation enhance classification performance by ensuring the model focuses on relevant aspects of the data.

5.4 Explainable AI (XAI)

Explainable AI (XAI) techniques, such as SHAP (Shapley Additive Explanation) and LIME (Locally Interpretable Model-Agnostic Explanation), are employed to interpret model decisions. The SHAP value for a feature i is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \tag{15}$$

where S is a subset of features, N is the set of all features, and $v(S)$ is the model prediction for feature subset S.

6 RESULTS AND DISCUSSION

This study provides a detailed comparative analysis of machine learning (ML) and deep learning (DL) classifiers for cyberbullying (CB) detection. The experiments utilized hybrid datasets sourced from Mandalay and Kaggle to address the challenges associated with detecting diverse categories of cyberbullying. Data preprocessing, feature engineering, and advanced classification algorithms were applied to ensure comprehensive analysis.

6.1 Data Preparation and Evaluation Metrics

The datasets underwent rigorous preprocessing steps to enhance model performance. These steps included:

- Removal of punctuation, stop words, numbers, and emojis.
- Spelling correction and language translation.
- Text normalization through compression management and lowercase conversion.

Post preprocessing, the data was split into 70% training and 30% testing sets. Feature engineering methods such as term frequency-inverse document frequency (TFIDF) and sentence embeddings were employed.

Evaluation metrics used for the models include:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{18}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{19}$$

6.2 Performance of ML and DL Models

Table 1 illustrates the performance metrics for six ML classifiers. Among these, the Extra Trees classifier emerged as the best performer with a 90.24% accuracy and 90.21% F1 score. This model was optimized using hyperparameters such as *n_estimators*: 109, *learning_rate*: 0.1, *max_depth*: 25, *min_samples_split*: 9, and *min_samples_leaf*: 1.

Table 1: Performance Metrics of ML Classifiers.

Model	F1 Score	Accuracy	Recall	Precision
Random Forest	0.8831	0.8836	0.8836	0.8955
Extra Trees	**0.9021**	**0.9024**	**0.9024**	**0.9112**
AdaBoost	0.8024	0.8039	0.8039	0.8188
MLP	0.8873	0.8876	0.8876	0.8959
XGBoost	0.8397	0.8416	0.8416	0.8654
Gradient Boost	0.8036	0.8064	0.8064	0.8332

Table 2 highlights the performance of three DL models. BERT achieved the highest accuracy of 91.36% and an F1 score of 91.24%, outperforming other DL models. The results underscore the effectiveness of BERT in capturing contextual relationships in textual data.

Table 2: Performance Metrics of DL Models.

Model	F1 Score	Accuracy	Recall	Precision
BiLSTM	0.9084	0.9105	0.8605	0.9619
BiGRU	0.9053	0.9053	0.8779	0.9345
BERT	**0.9124**	**0.9136**	**0.8723**	**0.9564**

The confusion matrix for the Extra Trees classifier in Figure 4 demonstrates its ability to accurately classify cyberbullying and non-cyberbullying events. True positives (TP) and true negatives (TN) significantly outweighed false positives (FP) and false negatives (FN), affirming the model's robustness.

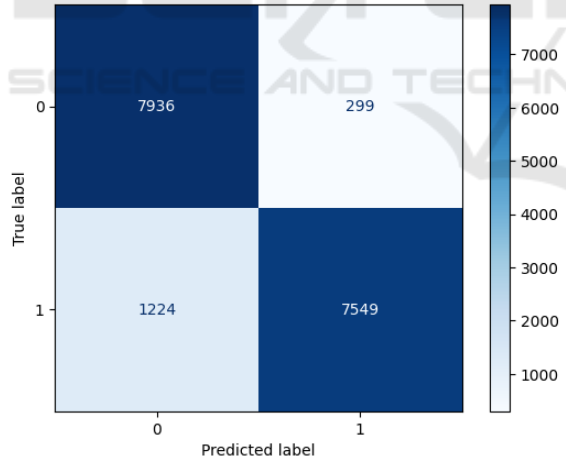

Figure 4: Confusion Matrix for Extra Trees Classifier.

Similarly, the ROC curve in Figure 5 highlights the model's efficiency, achieving an AUC score of 0.96, indicative of its high sensitivity and specificity in distinguishing cyberbullying instances.

6.2.1 Interpretability and XAI Analysis

Model interpretability was enhanced using explainable AI (XAI) techniques like SHAP and LIME. Figure 6 illustrates the LIME-based interpretation for the

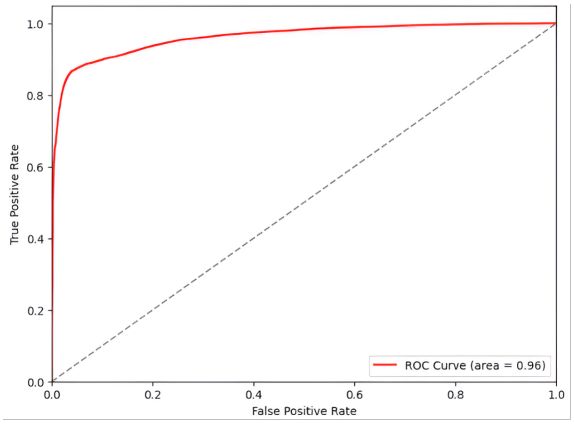

Figure 5: ROC Curve for Extra Trees Classifier.

Extra Trees classifier, highlighting feature contributions in predicting cyberbullying.

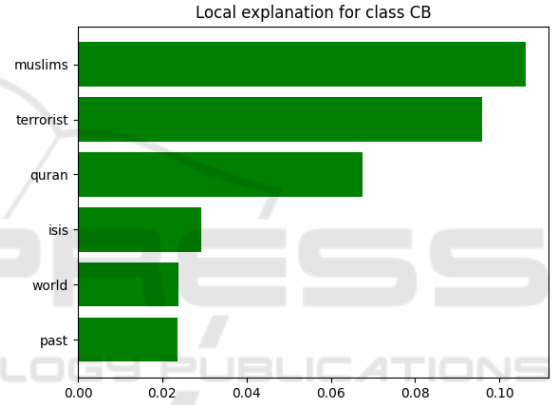

Figure 6: Local Interpretability with LIME: Contribution of Words to the CB Class.

LIME identified words such as *Muslim*, *terrorist*, and *Qur'an* as key indicators for the CB class. These insights enhance transparency and foster trust in automated cyberbullying detection systems. Figure 7 further exemplifies the interpretability of the classifier by highlighting specific words, such as *ignorant*, *reason*, and *ghetto*, that influenced predictions. Each panel demonstrates the contribution of these words toward classifying instances as bullying or not bullying. The visualizations emphasize the importance of these features in understanding the classifier's decision-making process.

6.3 Discussion

The advent of social networking platforms has revolutionized communication, offering unparalleled opportunities for global interaction. However, these platforms have also become a medium for cyberbullying, where individuals are harassed, intimidated, or

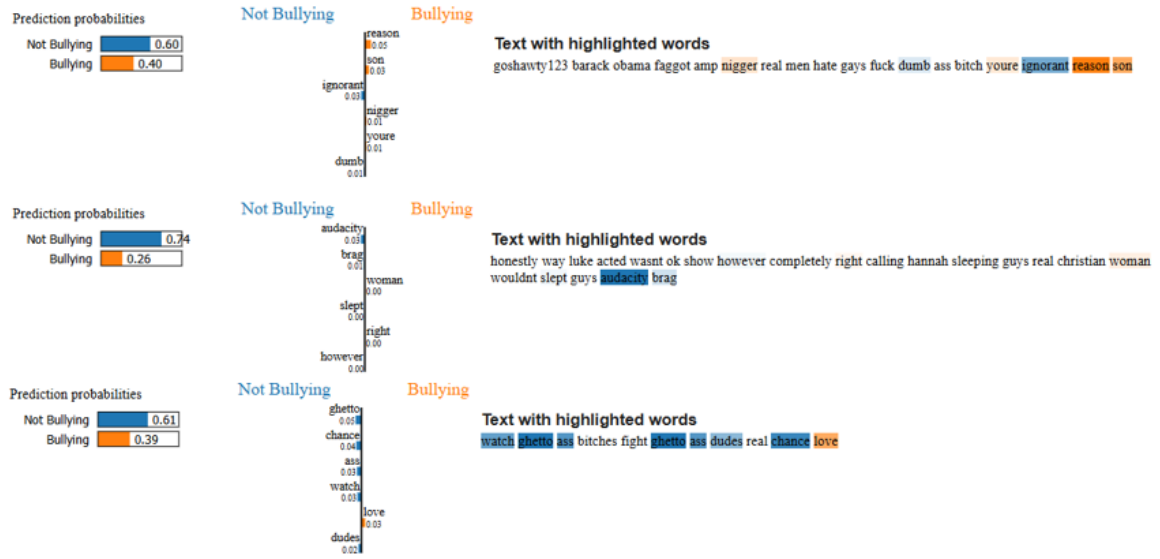

Figure 7: LIME-based Interpretability Analysis: Contribution of Words to Bullying and Not Bullying Classes.

harmed through electronic communication. This pervasive issue poses significant risks to mental health and social well-being, necessitating robust detection and prevention mechanisms. This study investigates the efficacy of machine learning (ML) and deep learning (DL) models in addressing this critical challenge, emphasizing their performance, limitations, and future implications.

6.3.1 Key Findings and Performance Analysis

The results underscore the effectiveness of ML and DL approaches in detecting cyberbullying on social media platforms. By employing algorithms such as Random Forest, Extra Trees, MLP, XGBoost, AdaBoost, Gradient Boost, BiLSTM, BiGRU, and BERT, a comprehensive evaluation was conducted using accuracy, F1 score, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC). Among the ML classifiers, Extra Trees emerged as the most effective, achieving an accuracy of 90.24% and an F1 score of 90.21%. For DL models, BERT outperformed others with 91.36% accuracy and an F1 score of 91.24%, due to its bidirectional contextual information processing.

6.3.2 Advantages of Deep Learning Models

Deep learning models, particularly BiLSTM, BiGRU, and BERT, showed superior performance compared to traditional ML algorithms. Their capacity to capture sequential dependencies and contextual nuances in textual data makes them particularly suitable for analyzing interactions on social media platforms. BERT's ability to leverage attention mechanisms helped it detect subtle linguistic features, enhancing its differentiation between cyberbullying and non-cyberbullying content. These attributes underscore the potential of DL models in handling the complexities of cyberbullying detection.

6.3.3 Challenges and Limitations

While the findings highlight significant progress, challenges remain. The dynamic nature of online communication presents difficulties in accurately identifying cyberbullying. Language nuances, cultural differences, and adversarial behaviors, such as coded language, often elude static models. Additionally, reliance on labeled datasets introduces potential biases, limiting the generalizability of models across diverse contexts.

The findings of this study, underscores the importance of robust evaluation mechanisms, such as confusion matrices and AUC-ROC analysis, in optimizing detection accuracy. Deep learning models, particularly BERT, showed exceptional performance, leveraging contextual understanding to address cyberbullying complexities. However, challenges related to language dynamics, adversarial behaviors, and ethical considerations remain. Addressing these issues through innovative approaches and interdisciplinary collaboration is key to creating a safer online environment. Future efforts should prioritize multilingual, multimodal detection systems and the integration of explainability techniques to ensure transparency and fairness, playing a crucial role in combating cyberbullying while upholding ethical standards.

7 CONCLUSION

Cyberbullying has become a critical concern in the digital age, impacting individuals and society on multiple levels. This study highlights the importance of employing robust detection mechanisms to address this growing issue effectively. Among the machine learning classifiers, the Extra Trees algorithm outperformed traditional methods, achieving notable results with an accuracy of 90.24%, precision of 91.12%, recall of 90.24%, and an F1-score of 90.21%. While these results are significant, deep learning models demonstrated even greater efficacy. In particular, attention-based architectures and bidirectional neural networks emerged as the most effective approaches, with the BERT-based model achieving the highest metrics: 91.36% accuracy, 93.45% precision, 87.23% recall, and 91.24% F1-score. This underscores the advantage of using neural networks to capture the nuanced and context-dependent nature of cyberbullying-related text. Notably, our shallow neural network framework offers a resource-efficient alternative, reducing the need for complex deep neural networks while maintaining competitive performance.

Future research should explore hybrid and ensemble methods to further improve detection accuracy and resilience. By focusing on these areas, researchers and practitioners can develop more comprehensive and scalable solutions to combat cyberbullying, ensuring safer online environments for all users.

REFERENCES

Al-Ajlan, M. A. and Ykhlef, M. (2018). Optimized twitter cyberbullying detection based on deep learning. In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pages 1–5. IEEE.

Arif, M. (2021). A systematic review of machine learning algorithms in cyberbullying detection: Future directions and challenges. *Journal of Information Security and Cybercrimes Research*, 4(1):01–26.

Atoum, J. O. (2020). Cyberbullying detection through sentiment analysis. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 292–297. IEEE.

Balakrishnan, V., Khan, S., and Arabnia, H. R. (2020). Improving cyberbullying detection using twitter users' psychological features and machine learning. *Computers & Security*, 90:101710.

Bruwaene, L. V., Houtte, E. V., and Backer, M. D. (2020). Utilizing machine learning and deep learning methods for detecting and analyzing cyberbullying. *Computers in Human Behavior*, 108:106–116.

Dalvi, R. R., Chavan, S. B., and Halbe, A. (2020). Detecting a twitter cyberbullying using machine learning. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 297–301. IEEE.

Fang, Y. et al. (2021). Gated recurrent units in natural language processing. *AI Research Letters*.

Gada, X. et al. (2021). Bidirectional lstm for text processing. *NLP Journal*.

Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., and Mohammed, A. (2019). Social media cyberbullying detection using machine learning. *International Journal of Advanced Computer Science and Applications*, 10(5):703–707.

Kaggle (2024). Cyberbullying tweets. Retrieved September 30, 2024, from https://www.kaggle.com/datasets/soorajtomar/cyberbullying-tweets.

Mehendale, N., Shah, K., Phadtare, C., and Rajpara, K. (2022). Cyberbullying detection for hindi-english language using machine learning. Available at SSRN 4116143, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4116143.

Paul, D. and Saha, A. (2022). Bidirectional encoder representations from transformers (bert). *Machine Learning Applications*.

Raj, K. (2021). Xgboost: An improved gradient boosting algorithm. *Data Science Advances*.

Raza, S. et al. (2020). Adaboost for weak classifiers. *Ensemble Learning Journal*.

Roy, P. and Mali, P. (2022). A deep transfer learning model for image-based cyberbullying detection in social networks. *International Journal of Advanced Computer Science and Applications*, 13(5):45–51.

Unnava, S. and Parasana, S. R. (2024). A study of cyberbullying detection and classification techniques: A machine learning approach. *Engineering, Technology & Applied Science Research*, 14(4):15607–15613.

Yadav, J., Kumar, D., and Chauhan, D. (2020). Cyberbullying detection using pre-trained bert model. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1096–1100. IEEE.

Yuvaraj, N. et al. (2021). Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking. *Mathematical Problems in Engineering*, 2021:1–12.