An Ensemble Learning Framework for Robust Cyberbullying Detection on Social Media

Mohammed Hisham Saeed, Shakaib Ahmed Mohammed, Taufeeq Noamaan, Rania Mehreen Farooq, Mohammed Abdul Raheem

***Abstract:** Social networking platforms on the Internet are now an essential feature of daily life worldwide, as these networks have made bridging the gap and sharing content an effortless task. Twitter stands out as a leading platform with a gigantic user base and is used extensively for communication between people and spreading information. Besides the many advantages these websites offer, such as promoting worldwide communication and dialogue, they may also pose unintended side effects that can be destructive to humanitarian and social life. One of the negative impacts of social networking sites is cyberbullying. Cyberbullying can be defined as "willful and repeated harm inflicted through the medium of electronic text" [1]. The support of harmful actions, such as harassment, threats, and humiliation, by individuals in online environments has brought about significant emotional and psychological effects for targeted individuals. The anonymity associated with social media platforms has the effect of increasing the occurrence of such detrimental activities, as there is less fear of the consequences of their actions, thus escalating the negative impact of cyberbullying. The Cyberbullying Detection Algorithm, a unique research approach, is used to combat the increasing problem of cyberbullying through ensemble-based learning algorithms, achieving a set of features for the Twitter dataset using machine learning techniques. This algorithm will look down on user-generated tweets in real time and discover patterns that may indicate cyberbullying behaviour. The role of the framework is to make the cyberbullying detection model on Internet platforms such as Twitter more accountable and effective through a mix of Machine Learning algorithms such as Random Forest, BERT, LSTM, and Ensemble. Our findings from an evaluative study of the critical features extracted from the Twitter dataset showed their relevance in cyberbullying detection. The performance evaluation based on key metrics such as F1 Score, Accuracy, AUC, and Precision depicts how the detection of cyberbullying can be made more effective and efficient by utilising machine learning algorithms that can detect online harassment and create a secure digital space for everyone.*

Manuscript received on 30 September 2024 | First Revised Manuscript received on 16 October 2024 | Second Revised Manuscript received on 19 December 2024 | Manuscript Accepted on 15 February 2025 | Manuscript published on 28 February 2025.

*Correspondence Author(s)

Mohammed Hisham Saeed, Department of Computer Science and Engineering, Osmania University, Hyderabad (Telangana), India. Email ID: hishamsaeed2003@gmail.com, ORCID ID: 0009-0000-7523-5195

Shakaib Ahmed Mohammed, Department of Computer Science and Engineering, Osmania University, Hyderabad (Telangana), India. Email ID: shakaibahmed203@gmail.com, ORCID ID: 0009-0008-1099-4965

Taufeeq Noamaan*, Department of Computer Science and Engineering, Osmania University, Hyderabad (Telangana), India. Email ID: taufeeqnoamaan@outlook.com, ORCID ID: 0009-0007-5189-5261

Rania Mehreen Farooq, Department of Computer Science and Engineering, Osmania University, Hyderabad (Telangana), India. Email ID: raniamehreenfarooq@gmail.com, ORCID ID: 0009-0007-4826-7707

Mohammed Abdul Raheem, Department of Computer Science and Engineering, Osmania University, Hyderabad (Telangana), India. Email ID: maraheem@mjcollege.ac.in, ORCID ID: 0000-0002-1652-6109

***Keywords:** Cyberbullying Detection, Ensemble-Based Learning Algorithms, Random Forest, BERT, LSTM*

I. INTRODUCTION

Cyberbullying is one of the cruellest problems that have come out in different forms of modernity, and nowadays, it is just a minor topic to discuss. It is a modern type of bullying similar to simple bullying in which the bullies use the internet to frighten, harass, or intimidate the victims. Cyberbullying victims can experience it in different ways: through offensive text messages, the publishing of private information, or the posting of unpleasant and offensive content online. Cyberbullying is genuinely frightening, with 30 per cent of teens reporting that they had been victims of online harassment at least once. The social and computational stances on cyberbullying are vital to understanding the reason for it and the extent to which it occurs.

Cyberbullying is frequently seen as a network aggression that involves someone intentionally causing damage or fear to another through the use of technology. This can cause severe emotional harm, leading cyberbullying victims to experience depression, stress, and other mental issues. In addition to that, cyberbullying has significant effects on the school environment, leading to lower academic performance, reduced engagement, and disruptions in the school climate. Therefore, it is essential to implement effective strategies for preventing and addressing cyberbullying in educational settings. Among these strategies, developing innovative solutions is vital to stopping cyberbullying. Ensemble learning methods, based on multiple machine learning tools, have been shown to enhance the accuracy and robustness of cyberbullying detection algorithms, which range among the approaches offering promising results. However, many existing systems are limited because cyberbullying is often misrepresented in the literature, leading to inaccurate models that lack real-world applicability [2]. This method could, therefore, increase our capability to watch cyber-bullying, especially in real time. This paper presents a novel ensemble learning technique that is appropriate for detecting cyberbullying, using a combination of several machine learning models. We will discuss the design of our approach, including the data used for the evaluation. The preparation is based on a dataset of unlabelled cyberbullying cases. The proposed method efficiently locates cyberbullying cases with high accuracy and low false positive rates. Consequently, this technique can provide a solid base for identifying and preventing digital violence.

II. MOTIVATION

Twitter, a widely used social media platform, has become the primary source of information for researchers, journalists, and the general public. Twitter has 440 million active customers, which is well-suited for real-time communication. However, the high number of tweets by the hour can also serve the purpose of deterring cyberbullying. In the course of an investigation, it turned out that Twitter is a safe place for cyberbullying as people can harm and intimidate others with no identity and no charge very quickly. The research finds the fact that cyberbullying may trigger the development of various mental health disorders, such as depression and anxiety, as well as suicidal thoughts. The early diagnosis of cyberbullying is the first step to reducing negative consequences on the affected parties and providing the affected individuals with the necessary help.

Furthermore, the manual examination of tweets is hectic and based on human mistakes. This effort aims to solve the issue using machine learning algorithms which afford the implementation of systems that intelligently recognise cyberbullying instances. The idea of ensemble learning was born due to the need to combine several machine learning algorithms to reach an overall better functioning system. Using the algorithms of various kinds of machine learning, we plan to get a more precise and, therefore, trustworthy model for detecting cyberbullying on Twitter. The motivation for this research is driven by a more effective method for detecting cyberbullying. The enormous number of tweets on Twitter daily makes it challenging to analyse tweets for signs of cyberbullying manually. Moreover, the anonymity of Twitter users always prevents the identification of both offenders and victims. Through our system of cyberbullying detection, our research contributes to a more effective solution for identifying online bullying and harassment, ultimately contributing to a safer and more supportive online community.

III. PROBLEM STATEMENT

We aim to build an ensemble learning approach that connects mainstream models, like Random Forest, LSTM BERT, and other ensemble methods, for cyberbullying detection and classification on Twitter. The framework will contain features extracted from the tweets, such as linguistics, sentiment, and social networks. They can be identified by analysing the textual patterns of cyberbullying tweets with these features. These features hold an essential aspect, capturing the fine distinction of language usage, tone of emotions and social interactions, symbolising the case of online bullying and harassment. The metrics used to assess the performance of our ensemble learning model include F1 score, accuracy score, AUC, precision and recall, and help us determine how efficient the model is in cyberbullying detection. This will be achieved by using different functionalities together with the complex ensemble learning techniques that aim to produce a model that will be precise and very helpful for proper cyberbullying detection on Twitter.

IV. LITERATURE SURVEY

A study by Alam et al. (2024) [3] examines automation in detecting cyberbullying in social media and focuses on various classifiers for detecting abusive language. Based on the result, the Logistic Regression and Bagging ensemble models have shown promising results individually. Still, the SLE and DLE proposed voting classifiers have shown a higher accuracy of 96% using the feature extraction method as TF IDF (Unigram) with the validation sampling techniques incorporated via K-folds cross-validation. This proved that ensemble methods, especially SLE and DLE, are extremely powerful techniques for increasing the detection accuracy of cyberbullying. As a result of the research, the requirements for the feature selection algorithms for MAP and the automation of the detection processes of offensive language were identified as contributing to safer internet use.

Abdulkarim Faraj Alqahtani and Mohammad Ilyas (2024) [4] emphasised utilising a stacking ensemble model to determine whether the tweets associated with cyberbullying are aggressive or non-aggressive. They used four feature extraction methods: the bags of words model, term frequency-inverse document frequency, Word2Vec, and GloVe models. Among these, the Word2Vec model was the most effective in improving the F1 score. Their stacking ensemble model of Decision Trees, Random Forest, Logistic Regression, K-nearest neighbours and Linear Support Vector Classification was even more accurate at 94% than general machine learning models. Further, recall and precision measures were the parameters incorporated to qualify the classifiers. Some limitations were associated with comparing soft voting ensemble methods and evaluating toxic text type, which indicates that the presented topics should be researched further.

In another study, Pankaj Shah and Shivali Chopra (2024) [5] sought to address the detection of cyberbullying in mixed languages based on the use of several learning machines. From the study, it was established that among all the algorithms, SVM has high accuracy in classifying cases of cyberbullying, especially with those that are multilingual in content. Specifically, the study focused on the aspect of feature engineering to enhance the current ML models with a focus on detecting cyberbullying in mixed language texts, which is even more complex with the current increase in online activity due to the COVID-19 pandemic.

Jinan Redha Mutar (2024) [6] proposed a machine learning and deep learning model to detect cyberbullying using TF-IDF for traditional algorithms and embedding text for deep learning. The study employed fifteen classifiers in the current research, and the voting classifier obtained the highest accuracy level of 96.5%. Cross-validation and Recursive Feature Elimination (RFECV) were utilised to enhance the model's performance, and implementing a voting classifier was the most effective approach. This shows that voting classifiers could be applied on social media. However, the study has some limitations, such as the absence of real-time data from social media platforms and biases in the data labelling of the set.

Raj et al. (2024) [7] examined the feasibility of integrating ensemble methods

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

such as Random Forest and gradient boosting in threat identification. The work also demonstrated that ensemble models were more accurate than linear models and neural networks. For that matter, Random Forest was correct in distinguishing between the malicious content and the rest and is less prone to overfitting. In the study, the authors noted that ensemble methods perform significantly better than the basic approach, providing more scalable and accurate solutions to threats. However, the work presented did not have real-time data.

Jalda et al [8]. stated that the following machine learning algorithms were investigated for detecting cyberbullying on social media: SVM, Random Forest, Multinomial Naive Bayes and Logistic Regression. The highest accuracy was recorded by SVM at 96.5%, making it the best algorithm of the four. Random Forest and Multinomial Naive Bayes produced decent results with overall accuracy rates of 94.2% and 92.8%, respectively. The study explored how SVM can be used in cyberbullying and noted that there is a high need to adopt early intervention measures to prevent online bullying.

Bhagyashree Kadam (2023) [9] used a Decision Tree, Random Forest, Support Vector Machine (SVM) and Multilayer Perceptron for cyberbullying with the help of a hate speech dataset. The Decision Tree received the highest accuracy in detecting cyberbullying among all the algorithms used. The study adopted text mining on user messages and identifying bullying with an invention that this combination of machine learning techniques could provide a solution to early cyberbullying intervention. Some of the gaps included data unavailability, privacy issues, and lack of feedback and demographic details in case of detection.

Muneer et al. (2023) [10] proposed a stacking ensemble using BERT to detect cyberbullying on social media platforms like Twitter and Facebook. The proposed model yields an accuracy of 97.4% on the Twitter dataset compared to the baseline and the proposed modification using BERT. The stacking model again rose to the occasion, and its performance was impressive on the new Twitter and Facebook integrated combined data set with an accuracy of 90.97%. In particular, ensemble stacking learning and feature extraction processes were identified as providing a more advanced approach to cyberbullying identification.

Aaminah Ali & Adeel M. Syed (2022) [11] used different algorithms such as Supervised Vector Machine, Naïve Bayes, Random Forest, and an ensemble method to develop an efficient understanding of cyberbullying. SVM was identified as the best among the models, as it produced high accuracy on different datasets and, on average, obtained an accuracy of 79.3%. The authors found that SVM and Logistic Regression yielded better results than Decision Tree and Random Forest when analysing Twitter text data. This supports the effectiveness of the proposed SVM and CNN solutions for identifying cyberbullying on social networks.

Both the studies by Alam et al. (2024) and Alqahtani & Ilyas (2024) are focused on the use of ensemble methods in the case of cyberbullying detection. Alam et al. emphasise the idea of a voting classifier and achieving an accuracy of 96%, while Alqahtani & Ilyas rely on a stacking ensemble model, with an accuracy of 94%. Alam et al. highlighted the importance of automating the detection of offensive language, whereas Alqahtani & Ilyas showcased the effectiveness of the Word2Vec feature extraction method. The two studies noted limitations when comparing these soft voting methods and calling for diverse datasets. On the other hand, Pankaj Shah and Shivali Chopra (2024) discussed the detection of cyberbullying in mixed languages where outcomes showed SVM yield high accuracy in multilingual environments; this means that there is a need for an improved feature engineering strategy. Jinan Redha Mutar (2024) proposed a hybrid combination of machine learning and deep learning models, which attained 96.5% accuracy with voting classifiers. Raj et al. (2024) also showed that ensemble learning approaches gave a better threat detection performance than linear models, but again, they lacked real-time data, too. Jalda et al. (2024) noted SVM as the effective algorithm with an accuracy of 96.5%, emphasising the need for early intervention measures, while Bhagyashree Kadam (2023) used Decision Tree and Random Forest, reporting Decision Tree as the most effective but pointing to issues like data unavailability and privacy. Muneer et al. (2023) used a stacking ensemble model implementing the BERT method. They demonstrated that it can achieve 97.4% accuracy on Twitter data, highlighting that more advanced techniques are required in recognising cyberbullying. Aaminah Ali & Adeel M. Syed (2022) found that SVM is the best-performing algorithm across various datasets, emphasising the model's applicability in recognising cyberbullying issues. Collectively, these studies underline the efficacy of ensemble and hybrid approaches while revealing common limitations, such as the lack of real-time data and the need for diverse and robust datasets.

V. METHODOLOGY

The cyberbullying detection model encompassing various stages is shown in Figure 1. The overall model has four main stages: Twitter data collection, data preprocessing, feature extraction (i.e., classifying as either Normal or Hate speech or offensive language) and cyberbullying detection and classification (i.e., classifying as cyberbullying or not). The present study examined cyberbullying on Twitter by utilising advanced technologies and techniques such as Natural Language Processing (NLP), machine learning (Random Forest) and deep learning models (BERT and LSTM).

A. Dataset Description

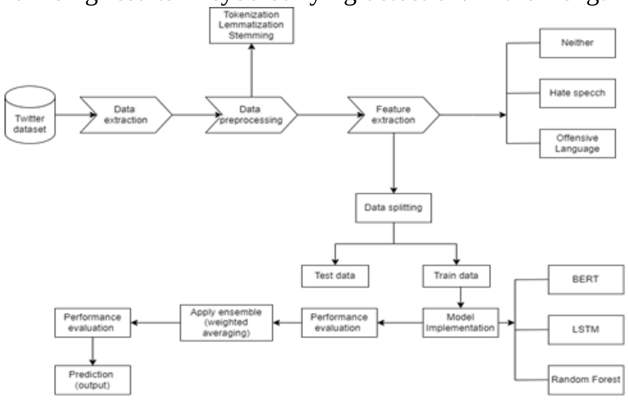
The tweet dataset used in this research, sourced from online repositories, consists of 26,954 unlabelled tweets.

B. Model Description

BERT, LSTM and Random Forest (RF) models have emerged as superior choices for cyberbullying detection, among other traditional machine learning models such as SVM, KNN, logistic regression and Naive Bayes. This superiority is because such advanced models can more effectively handle the complicated language contexts and patterns found in social media data.

BERT, a transformer-based pre-trained language model, has been successful in doing much better than other models in detecting cyberbullying content [21]. BERT achieved the highest accuracy of 80.165% in the Bengali language text cyberbullying detection [22]. For cyberbullying detection in the Chinese language, a hybrid model based on XLNet, one of the variants of BERT, combined with bi-LSTM, performed better than all the baseline models [23]. In hate speech detection, transformer models such as BERT and DistilBERT performed much better when compared with other deep learning and machine learning algorithms [27].

LSTM, a type of recurrent neural network, has also shown promising results in cyberbullying detection. In the Bengali text study, LSTM-based approaches achieved the highest accuracy among deep learning models, at 77.804% [21]. The combination of BERT and Bi-LSTM for Indonesian cyberbullying detection yielded a high accuracy rate of 90% [28].



[Fig.1: Proposed System] [27]

Although the Random Forest algorithm has not outperformed deep learning models, it has performed commendably in some studies. When compared for their efficacy in detecting cyberbullying in social networks, RF was among the top-performing models alongside SVM and Logistic Regression [23].

It is worth noting that several investigations showed that traditional ML techniques like SVM, Naive Bayes, and others are effective in specific situations. For example, in the investigation of Bengali text, Multinomial Naive Bayes performed best (78.816%) among the classical machine learning methods, as reported in [21]. Although the general trend is that in detecting cyberbully behaviour, deep learning and transformer-based models, BERT and its modifications, in particular, perform much better than the machine learning approaches [22].

In conclusion, the cyberbullying detection performance of BERT, LSTM, and RF is higher than that of other algorithms as these approaches can effectively learn the intricate language contexts, semantics, and patterns within the text [24]. These model's capacity to process and learn from large amounts of data makes them particularly well-suited for detecting cyberbullying in various languages and contexts [25].

BERT - Bidirectional Encoder Representations from Transformers (BERT), developed by Google, is a powerful language representation model designed for natural language processing (NLP) tasks [26]. This neural network model employs a transformer architecture that develops deep bidirectional representations from unlabelled text [29]. It achieves this by concurrently considering the context from prior and following layers during the pre-training process [17]. By doing so, BERT obtains a contextual comprehension of words within sentences and is exceptionally suited for various NLP jobs. BERT's main technical breakthrough is the application of bidirectional training to Transformer, a vital attention model in language modelling [20]. BERT was trained on two distinct tasks to develop powerful language-understanding capabilities. The first task is the masked language model (MLM), in which certain words in a sentence are masked, and the model is trained to predict these masked words [15]. Training BERT from scratch requires masking a certain percentage of the words in the corpus, with a recommended masking rate of 15%. This percentage is crucial for balancing the model's training efficiency and effectiveness because too little masking makes the training process prohibitively expensive. In contrast, too much masking can obscure the context necessary for the model to learn effectively. The second task involves predicting the following sentence. In this task, BERT is given pairs of sentences and aims to understand their relationship, predicting whether two sentences will follow each other in a coherent narrative [16]. For instance, given Sentence A ("The dog barked loudly") and Sentence B ("It was warning the intruder"), BERT is tasked with determining if Sentence B is a plausible continuation of Sentence A. This binary classification task helps BERT to assess the relationship and coherence between the two sentences. BERT is pre-trained on a large corpus of text, which allows it to generalise well to various text-based tasks with minimal fine-tuning [17]. This pre-training advantage is absent in models such as traditional n-grams or simple feed-forward networks.

LSTM - RNNs Recurrent Neural Networks (RNNs) are a type of neural network designed to process sequential data. LSTM networks, a type of recurrent neural network (RNN), are distinguished by their ability to address the challenge of learning long-term dependencies in sequential data, a task that traditional RNNs struggle with because of issues such as the vanishing gradient problem and was first introduced by [18]. LSTM - Long Short-Term Memory, the term implies that the network has a short-term memory of the immediate past events for decision making; however, at the same time, the network also has a long-term memory for decision making. Besides the RNN's hidden state, an LSTM block also has memory cells for storing prior knowledge and adds a sequence of gates known as input, output, and forget gates. The gates prevent errors from vanishing or exploding by controlling the information entering and leaving the memory cell. The Input Gate controls new information added to the cell state from the current input. The model uses a sigmoid activation function, which takes the current input and the previous hidden state as inputs and converts information between 0 and 1. The Forget Gate determines the information to discard from the memory cell. A sigmoid activation function is used to perform this process. The sigmoid activation function outputs values between 0 and 1 from the weighted input and previous hidden state. The Output Gate controls the information revealed from the updated cell state to the output.

It outputs a value between 0 and 1. These gates collectively enable the network to selectively retain or discard information, contributing to the

proficiency of LSTM in handling long-term dependencies and sequential data, such as speech recognition and sentiment analysis. Thus, LSTM models are particularly effective for learning and predicting sequence data because they can retain memory over long sequences and capture dependencies and patterns over time. The gating mechanisms in LSTM
that control the flow of information mitigate the vanishing gradient problem that plagues standard RNNs. Thus, learning is better over long sequences compared to simpler models.

Random Forest - is an ensemble learning algorithm with base classifiers as decision trees generated from random subsets of features, each capturing a distinct pattern in the data [19]. It is, by far, the most popular technique used by several cyberbullying detection studies for its performance in dealing with multiple features that may be correlated. Random Forest is used for classification and regression tasks, and in this study, we only consider random forest as a classifier. Random Forest is a supervised machine learning algorithm that generates a "forest" of decision trees. It combines the predictions from these decision trees to improve the accuracy and performance, thereby addressing the limitations of individual decision trees. Each tree is trained on a random subset of the data, allowing the model to incorporate a diverse range of perspectives. When making predictions, each tree in the forest provides a classification (classification tasks) or output (regression tasks), and the final decision is determined by majority voting or averaging the results across all trees. The fundamental principle behind Random Forest is the "bagging" method, where a combination of learning models produces generally more reliable results than single models. Random Forest is resilient to noisy data owing to its ensemble nature, where multiple trees vote on the final classification. This robustness makes it superior to single classifiers, such as decision trees, which may be more affected by noise.

A. ENSEMBLE IMPLEMENTATION

A. Preprocessing

It entails pre-processing the tweets and extracting critical information. Real-world tweets contain unnecessary characters and words such as articles, punctuations, prepositions, and common verbs. Before applying the machine learning algorithms, the tweets were cleaned and prepared for the detection phase to enable the models to understand and classify each text. Data preprocessing involved the removal of symbols, special characters, numbers, and stop words from each tweet in the dataset. Words like “um”, “like,” “a,” “an,” “the,” “but” and “or” are filler words that carry little information. When developing natural language processing (NLP) models, algorithms must programmed not to include these words. These uninformative words that do not add context are called stop words. Before feature extraction, we performed tokenization, lemmatization, and stemming as a part of the preprocessing phase. NLTK, Python’s API library, was utilised to conduct the following tasks on textual data: tokenization, lemmatization and stemming. Tokenization involves breaking text into smaller units called tokens. These tokens can be words, sub-words, or characters, depending on the type of tokenization used. Tokenization is critical because most NLP models, especially those under the deep learning framework, require text data representation in a structured form for further analysis. For instance, the phrase “Cyberbullying detection” is tokenized into “Cyberbullying” and “detection”. Stemming strips away suffixes, prefixes, and other elements of a given word until only its root remains. It often reduces words to their base or root form without considering their context. Example: "running", "runs", and "runner" are reduced to “run”. “Better” is reduced to “bet”. The stemmer reduces the word ‘better’ to ‘bet’, which differs from what better means. Lemmatization reduces individual tokens from a sentence to their base form. The base is called the lemma. This process considers the context and uses part-of-speech tagging (POS). For example, “running” becomes “run”, and “better” becomes “good”; hence, it has more semantic accuracy than stemming.

B. Feature Extraction

To achieve the objectives of this study, we extracted the following features from the dataset - Hate Speech, Offensive and Normal.

i. **Hate Speech**: Any speech that denigrates or discriminates against a person or group on account of their race, ethnicity, gender, or religion could be considered hate speech and may pose a danger to societal harmony [12]. It can take different forms, including verbal abuse, written messages and visual images such as memes. Examples include racial insults, insulting remarks about an individual’s belief system and incitements to violent acts against specific communities.

ii. **Offensive**: It refers to tweets or comments that are vulgar or obscene, which breach commonly accepted standards of decency but do not necessarily incite violence or hatred. We assumed that all hate speech was also offensive.

iii. **Normal**: It encompasses tweets or comments that are neither hate speech nor offensive. These include neutral or positive remarks, general discussions, and content that does not contain harmful language or intent. Examples include conversations between friends, greetings, casual talks, and informative posts.

C. TF-IDF for Feature Extraction

We used Term Frequency-Inverse Document Frequency (TF-IDF) as a feature extraction technique for the ensemble's traditional machine learning model (Random Forest). It is a statistical measure used to evaluate the importance of a word in a document relative to that in a collection of documents (corpus). This model avoided the bag-of-words (BoW) approach. The BoW model is often avoided because of its limitations in capturing the significance of the terms within a document. While BoW treats all words equally, TF-IDF weighs terms based on frequency across documents, enhancing the representation of rare but informative words.

The purpose of TF-IDF is to measure a term's significance in textual data based on how frequently it appears in the document and how seldom it occurs across documents in the corpus. TF-IDF converts textual data into numerical vectors, assigning each word a weight based on its frequency in a document (tweet) and inverse

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

frequency across the entire corpus. Under this system, common expressions such as ‘the’ and ‘is’ are minimised in importance, while rare but significant words, including ‘offensive slurs’ or ‘targeted harassment terms’ are prioritised.

i. **Term Frequency (TF)**: This is the ratio of the number of times a word occurs in a document to the total number of words in that document. The aim is to emphasise words that are frequent in the document.

ii. **Inverse Document Frequency (IDF)**: IDF of a term reflects the proportion of documents in the corpus that contain the term. The IDF calculates the rarity of a term in the corpus. It is the logarithm of the ratio of the total number of documents to the number of documents containing the word. The main aim is to quantify the importance of a term within a document relative to a corpus of documents.

Multiplying a term's TF and IDF scores gives the TF-IDF score. Thus, TF-IDF is particularly valuable in detecting offensive language and cyberbullying because it highlights discriminative terms that may not be frequent across all tweets but indicate harmful content in a particular tweet. For instance, in a given dataset, the words hate, kill or stupid are likely to receive higher TF-IDF scores in messages with offensive language as compared to their presence in everyday conversation. Usual conversational posts may contain words like “movie”, “great”, or “fun”, which often come with low TF-IDF scores. By applying such a statistical method, the model can learn which terms are more critical in terms of being offensive than neutral conversations. TF-IDF features are fed into the Random Forest model, which can effectively use these weighted word representations to split data based on the presence or absence of harmful language. This approach provides a more interpretable and structured approach to text classification, making it a valuable addition to the deep learning models in the ensemble. Unlike TF-IDF, which gives a brief and statistical overview of the text, the ensemble of deep learning models (BERT and LSTM), on the other hand, provides deeper, contextualised, and sequential interpretations of tweets.

D. BERT for Feature Extraction

BERT (Bidirectional Encoder Representations from Transformers) extracts rich, contextual features from text data. The pre-trained language model BERT generates embeddings representing different meanings of words in their contexts, consequently providing an understanding of subtle language differences. Every word or subword token within a tweet is described using an embedding in a high-dimensional space that defines its meaning relative to the other tokens [13]. This representation ensures that specific phrases indicative of harmful intent are detected, thereby enhancing the detection of sarcasm or indirect insults that simpler models may have overlooked. For instance, in the statement, “You did a great job on that project; your hard work paid off,” the word " great job" is encouraging and positive and shows support.

On the other hand, the use of ‘great job’ in the statement “Oh yeah, ‘great job’ on that project, maybe next time you’ll actually try” is sarcastic. The speaker believes the work done was substandard and is mocking the person. BERT's contextual understanding allows it to classify the first statement as positive and supportive while interpreting the second statement as negative and potentially harmful, highlighting its effectiveness in cyberbullying detection.

E. LSTM for Feature Extraction

LSTM models are effective for capturing sequential dependencies in text. Each tweet is represented as a sequence of word embeddings, and the LSTM processes them stepwise, keeping track of previous words [14]. This is important, especially for tweets where the meaning of a word changes according to preceding or succeeding text. For instance, the phrase: “I really enjoy football, but my coach really annoys me,” can still be analysed positively primarily because of its initial part despite the latter part. The LSTM would recognise that the broader context remains predominantly positive, which aids the differentiation from something offensive like, “I hate my coach; he’s the most terrible person ever!” Here, the structure and sequence in which phrases are used signal the LSTM to classify the latter as potentially offensive because of the strong negative adjectives and overall directive intention. The LSTM model improves the understanding of time-based or sequential relationships in a language. It can pick up on repeated patterns of harassment in different tweets or identify shifts in sentiments that cause cyberbullying. LSTM manages these sequences superior to static feature extraction methods such as TF-IDF.

F. Ensemble Integration

Using an ensemble of TF-IDF, BERT, and LSTM produces a solid framework for feature extraction. TF-IDF provides statistical significance, BERT captures complex and contextual meanings, whereas LSTM retains sequential information, which helps to identify patterns in language flow across a single tweet or several tweets. The ensemble incorporates each model’s feature extraction through weighted averaging. Each model contributes according to its unique feature representation. Through this multifaceted approach, we can perform a more refined classification that enhances the model efficacy at differentiating between normal, offensive, and hate speech.

G. Cyberbullying Detection and Classification

The ensemble classifies the input tweets as cyberbullying or non-cyberbullying by integrating the three models, BERT, LSTM, and Random Forests, through weighted averaging.

Weighted averaging method: Weighted Averaging is a technique used to combine predictions from multiple models, where each model's contribution is proportional to its performance. This method assigns weights to each model based on its evaluation metrics, such as accuracy, precision, or F1-score, calculated on a validation or holdout dataset.

The proposed ensemble model combines the outputs of the three models: BERT, LSTM, and Random Forest. Each model classifies an input into one of the three classes: ‘offensive’, ‘hate’, or ‘normal’. Each model’s output is assigned a weight based on its performance. These individual outputs are then combined using the weighted averaging technique. Thus, the final classification for a given input is determined by aggregating the predictions of all three models. After aggregating the results for the three classes, we focus specifically on the percentage of the "normal" class prediction. If the aggregated output exceeds a certain threshold for the "normal" class, the input is classified as "non-cyberbullying". Conversely, if the "normal" percentage is below this threshold, the input is flagged as "cyberbullying". This method leverages the strengths of each model to improve overall accuracy in detecting harmful content.

Thus, in this approach, each model—whether BERT, LSTM, or Random Forest—predicts the probabilities or percentages of the input falling into three categories: “hate speech”, “offensive speech”, and “normal speech”. These predicted percentages are then combined using the weighted averaging method, where higher-performing models influence the final output more. A certain threshold is applied to the combined “normal speech” percentage to classify a tweet as cyberbullying or not. If this percentage exceeds the threshold, the tweet is classified as non-cyberbullying; otherwise, it is flagged as cyberbullying. This method leverages the strengths of each model while using weighted averaging to improve classification accuracy.

VI. RESULTS AND DISCUSSION

We adopted a two-step process to classify tweets as either cyberbullying or non-cyberbullying. In the **first step**, we classified the tweets or comments as Hate Speech, Offensive, or normal using the Random Forest, BERT and LSTM models individually. In the **second step**, we classified the tweets into Cyberbullying and not cyberbullying using the ensemble approach and weighted averaging method.

- **Cyberbullying**: All tweets classified as hate speech were labelled as cyberbullying tweets.
- **Not Cyberbullying**: All tweets classified as Normal or Offensive were labelled ‘not cyberbullying.’

A. Evaluation Metrics

Understanding the performance of a model is an essential part of any classification algorithm. The performance of a classifier is measured using several methods, such as recall, precision, accuracy, F1, and AUC.

True Positive: The number of times the predicted value is positive and the actual value is positive.

False Positive: The number of times the predicted value is positive and the actual value is negative.

True Negative: The number of times the predicted value is negative and the actual value is negative.

False Negative: The number of times the predicted value is negative and the actual value is positive.

The metrics evaluated in this study are:

1. **Accuracy**: Accuracy is the proportion of correctly classified instances in a dataset. It measures the overall effectiveness of a classifier. It is the sum of all true values divided by the total values.

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

2. **Precision**: This is the ratio of accurately predicted positive observations to the total predicted positive observations. It indicates how many of the predicted positive instances are actually positive. Precision is crucial in contexts where the costs of false positives are significant, as it measures the proportion of true-positive predictions among all positive predictions.

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \text{----- (2)}$$

3. **Recall (Sensitivity)**: The true positives are divided by the total number of actual positive values.

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

4. **F1-Score**: This is the weighted average of the precision and recall. It considers the cognizance of both false positives and false negatives. F1 is usually more beneficial than accuracy. This is useful when seeking a balance between Precision and Recall.

$$\text{f1 score} = \frac{2{*}\text{precison}{*}\text{recall}}{\text{precison+recall}} \quad \text{----- (4)}$$

5. **AUC** stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1). AUC ranges in value from 0 to 1. A model with predictions that are 100% wrong has an AUC of 0.0; one with 100% correct has an AUC of 1.0. Thus, the Area Under the Curve (AUC) is a powerful indicator of the model's ability to discriminate between positive and negative outcomes across various classification thresholds.

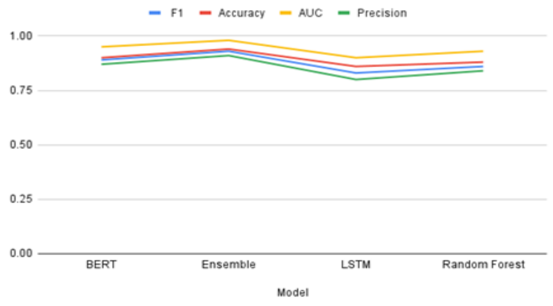
B. Performance Evaluation

We evaluated random forest, Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), and the ensemble approach to assess the various models' effectiveness in detecting cyberbullying. The proposed ensemble model combines three models (Bidirectional Encoder Representations from Transformers (BERT), LSTM (long short-term memory (LSTM), and Random Forest (RF), using the weighted averaging method. The evaluation metrics include F1-score, accuracy, AUC, and precision.

All classifiers were trained on the same dataset to determine their performance using the relevant metrics. Table 1 presents the results of the analysis of the classifier metrics.

Table I: Results

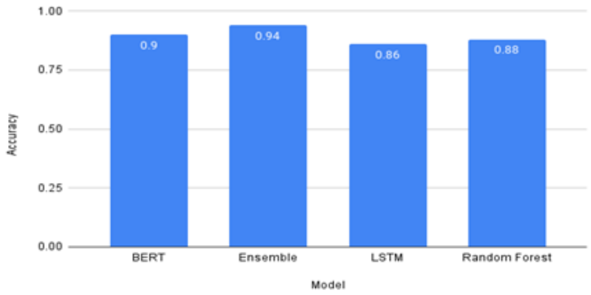
Model	F1	Accuracy	AUC	Precision
Random Forest	0.86	0.88	0.93	0.84
LSTM	0.83	0.86	0.9	0.8
BERT	0.89	0.9	0.95	0.87
Ensemble	0.93	0.94	0.98	0.91



[Fig.2: Comparison of Metrics Used]

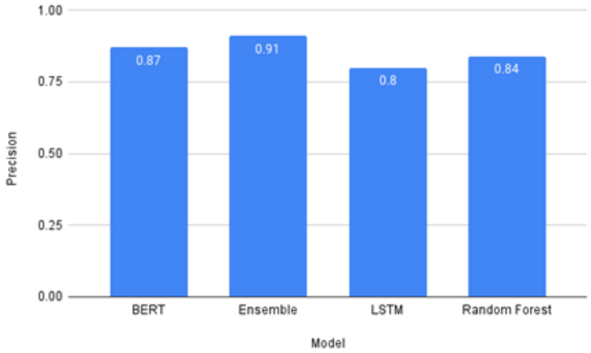
Figure 2 shows the combined line graph of the results obtained by evaluating the metrics for all the classifiers.

The line graph shows that the classifiers performed best when measuring the AUC and performed least when measuring precision. LSTM and Random Forest generally perform worse than other classifiers concerning all the metrics used. Figures 3-6 show bar graphs visualising the individual metrics against the classifiers.



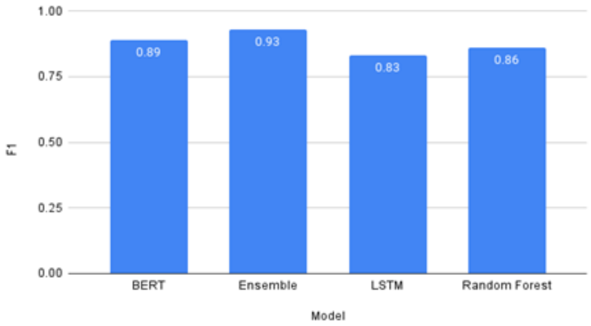
[Fig.3: Accuracy Score Analysis]

When evaluated by Accuracy, the Ensemble model led with a value of 0.94, suggesting that it correctly classified 94% of the cases in the dataset. BERT and Random Forest performed well with accuracies of 0.90 and 0.88, respectively, while LSTM had a slightly lower accuracy of 0.86. These accuracy scores reflect the models' ability to distinguish between classes, although the accuracy can be misleading in imbalanced datasets.



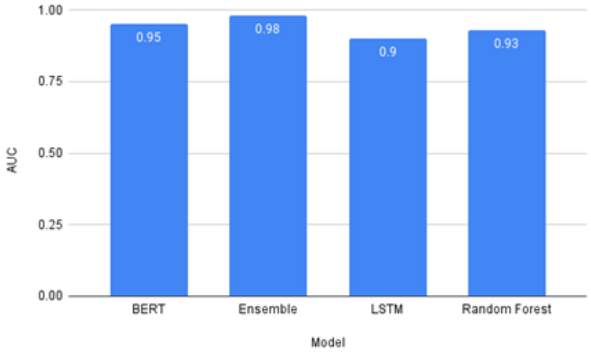
[Fig.4: Precision Score Analysis]

The Ensemble model attained a precision score of 0.91, followed by BERT with 0.87, indicating effective identification of actual positive cases. Random Forest achieved a precision of 0.84, whereas LSTM scored lower at 0.80.



[Fig.5: F1 Score Analysis]

These models can be ranked based on their F1 Scores. The Ensemble model achieved the highest F1 Score of 0.93, indicating a superior performance in identifying positive cases while minimising false positives and false negatives. The BERT model closely follows with an F1 Score of 0.89. Random Forest and LSTM had lower F1 Scores of 0.86 and 0.83, respectively, but still demonstrated competent performance.

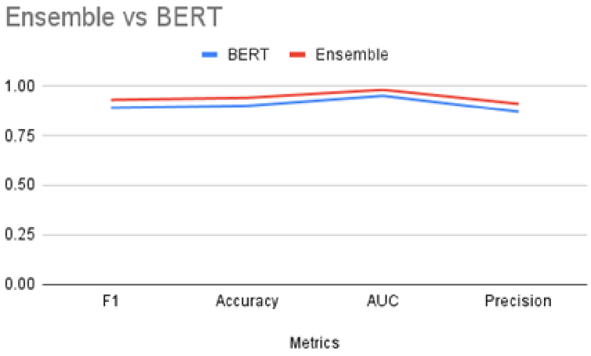


[Fig.6: AUC Score Analysis]

The Ensemble model also excelled here, achieving the highest AUC of 0.98, indicating exceptional overall performance. BERT followed with an AUC of 0.95, with Random Forest and LSTM achieving 0.93 and 0.90, respectively.

C. Comparison of Ensemble with BERT, LSTM and RF

i. Ensemble vs BERT



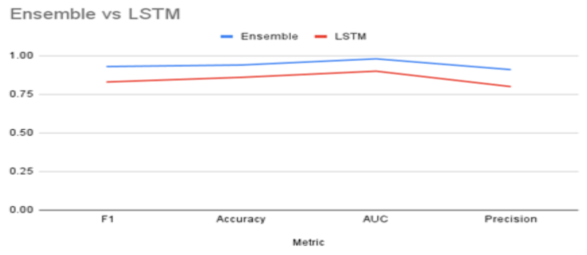
[Fig.7: Comparison of Ensemble and BERT Metrics]

When examining accuracy, the ensemble model takes the lead with an accuracy of 0.94 compared to BERT's accuracy of 0.90. This demonstrates that the ensemble model provides a higher correct classification rate than BERT in the given dataset. The ensemble model achieves a precision score of 0.91, while BERT's precision score is 0.87, indicating that the ensemble is more effective in minimising false positives. In terms of the F1 Score, the ensemble model outperforms BERT. The ensemble achieves an F1 Score of 0.93, indicating its strong balance between precision and recall. In contrast, BERT's F1 Score is slightly lower at 0.89, reflecting its effectiveness but not quite matching the ensemble's performance. The ensemble model excels with an AUC of 0.98, suggesting its superior ability to distinguish between positive and negative classes. BERT follows with a high AUC of 0.95. Both models exhibit strong performance; however, the ensemble demonstrates a slight advantage. BERT excels in linguistic context and has shown remarkable capabilities in various NLP tasks, benefiting from its architecture designed to understand language intricacies. However, it may

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

fall short in overall performance compared to ensembles that leverage multiple models to achieve more robust predictions across different metrics.

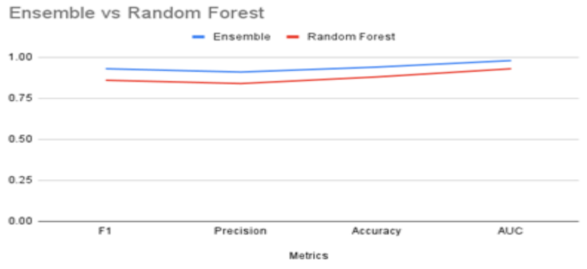
ii. Ensemble vs LSTM



[Fig.8: Comparison of Ensemble and LSTM Metrics]

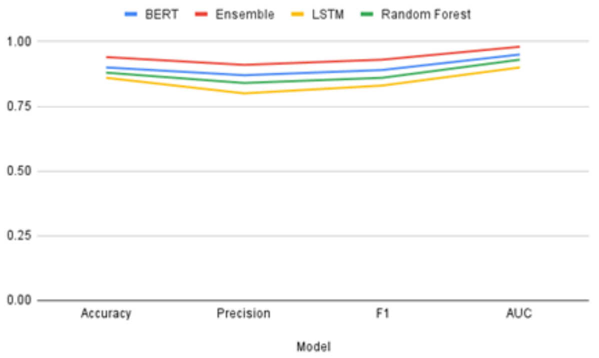
There is a significant difference between all the metrics used. In terms of accuracy, the Ensemble model leads with an accuracy of 0.94, while LSTM reports an accuracy of 0.86. This suggests that the Ensemble model is generally more reliable in correctly classifying instances than the LSTM model. The Ensemble model scores 0.91 for precision, indicating that it is more effective in identifying true positive instances than LSTM, which has a precision of 0.80. The Ensemble model outperforms LSTM in terms of the F1 Score, achieving a score of 0.93 compared to LSTM's score of 0.83. The Ensemble model has a higher AUC of 0.98, which indicates exceptional performance, while LSTM has an AUC of 0.90. This shows that the Ensemble model is superior in distinguishing between classes. The Ensemble model outperforms LSTM across multiple metrics, including F1 Score, accuracy, AUC, and precision.

iii. Ensemble vs Random Forest



[Fig.9: Comparison of Ensemble and RF Metrics]

When evaluating accuracy, the Ensemble model outperforms Random Forest with an accuracy of 0.94 versus 0.88 for Random Forest. This suggests that the Ensemble model is generally more reliable in correctly classifying instances within the dataset. The Ensemble model scored 0.91 for precision, notably higher than Random Forest's precision score of 0.84. This reveals that the Ensemble model is more effective in minimising false positives than the Random Forest model. The Ensemble model demonstrates a higher F1 Score of 0.93 than the Random Forest model, which achieved an F1 Score of 0.86. This indicates that the Ensemble model maintains a better balance between precision and recall, making it more effective in handling classification tasks. The Area Under the Curve (AUC) metric also favours the Ensemble model, which has an AUC of 0.98, indicating a superior ability to discriminate between positive and negative classes. In contrast, the Random Forest model has a lower AUC of 0.93, showing that the Ensemble model is more adept at distinguishing between the two classes. The Ensemble model outperforms Random Forest across multiple metrics, including F1 Score, accuracy, AUC, and precision.



[Fig.10: Comparison of Models]

Figure 10 indicates that leveraging an ensemble approach using the weighted averaging method for cyberbullying detection yields the best outcomes, making it the recommended choice for future implementations in this field. The ensemble model demonstrated the best performance across all metrics, followed by BERT, Random Forest, and LSTM. Using a Weighted Average Ensemble Model to detect cyberbullying offers several advantages over standalone machine learning models. Here are the key reasons why this approach is considered superior, as concluded from the results of this study:

1. Improved Accuracy: The ensemble model combines the predictions of multiple algorithms (Random Forest, BERT, and LSTM), which helps to mitigate the weaknesses of individual models. This results in a higher overall accuracy, as evidenced by the improved accuracy of 0.94, the highest among all the models.
2. Reduction of Overfitting: Standalone models, especially Random Forest, can easily overfit the training data, capturing noise rather than the underlying pattern. The ensemble approach helps to balance this by averaging the predictions, leading to a better generalisation of unseen data.
3. Diverse Perspectives: Each algorithm in the ensemble has its strengths and weaknesses. For example, LSTM is effective at understanding sequential data and long-term dependencies, while BERT is effective at understanding contextual relationships in language. By combining these models, the ensemble captures a broader range of patterns in the data, leading to more reliable predictions.
4. Error Compensation: When using multiple models, the errors made by one model can be compensated by the correct predictions of others. This collaborative approach reduces the likelihood of poor predictions occurring within a single model, enhancing the robustness of the predictions.
5. Flexibility and Adaptability: The Weighted Average Ensemble Model can be easily adjusted by changing the weights assigned to each algorithm based on their performance. This adaptability allows for fine-tuning the model to achieve optimal results for specific datasets or conditions.

In summary, the Weighted Average Ensemble Model provides a more accurate, robust, and flexible approach to cyberbullying detection than standalone models. By leveraging the strengths of multiple algorithms, it effectively addresses the limitations of individual models, leading to superior performance in prediction tasks.

VII. FUTURE SCOPE AND LIMITATIONS

The potential for the future scope of this research in detecting cyberbullying through enhanced machine-learning procedures is broad and encouraging. Based on the current study, several directions for subsequent development may improve the effectiveness of the proposed models and their relevance for practical use. One significant gap within this domain is the lack of proper knowledge of distillation and model-pruning applications. These strategies can be useful to reduce the computational burden related to large pre-trained models and make them smaller and faster without damaging their performance. More specifically, ensemble methods can be made faster to be used for real-time applications in settings with limited computing resources.

Another possible area of opportunity is the incorporation of active learning modes. This approach facilitates the labelling methodology by only incorporating unidentified samples that could provide the most significant annotations. In addition, it contributes to the improvement of ensemble performance. It increases the effectiveness of the active sample-learning method by optimising the identification of the most representative cases of cyberbullying. Moreover, further experimental studies involving modern weighting scenarios, including attention-based weighted or adaptive weightings, would be profound, leading to substantially enhancing the model. Through the dynamic setting of weights proportional to the measured performance or content specificity, decision-making can be improved, and the stability of predictions increases. The use of multimodal fusion techniques has significant potential for increasing the classification efficiency of cyberbullying cases. When using text data with user metadata or network structure information, it will be easier to predict context and ballot, making it easier to identify cyberbullying behaviour. Moreover, feature extraction using large language models (LLMs) or fine-tuning the models for sentiment analysis can boost detection accuracy. Because LLMs can be used to understand language subtlety compared to traditional models, it could be better to comprehend contextual and sentiment-related features.

Based on the analysis of the results, model transparency will remain an important factor in future work, for which explainability techniques should also be used, such as saliency maps or feature importance analysis. It is always helpful to know why a model has made a particular decision, and as more details about the model are provided, stakeholder trust increases.

Despite these promising directions, several limitations must be acknowledged and addressed in future studies. The challenge of accommodating long texts within social media constraints remains significant, necessitating innovative text segmentation or summarization approaches. However, several limitations that are worth mentioning and that should be circumvented in subsequent research should be outlined. The issue of handling long texts, given the character limits of social media, has yet to be fully resolved, although the issue of text segmentation or summarisation is in progress. Moreover, resolving a disproportionate representation of hostile and constructive comments would similarly and likely require a careful approach to dataset balancing and potentially synthetic data creation. There is also the problem of understanding natural language because text interpretation can be ambiguous; therefore, using updated methods in natural language processing while considering context and sentiment is crucial. Additionally, the high level of informality, which significantly characterises Internet communication, makes its identification more challenging. Future research should focus on developing robust models capable of understanding informal language variations commonly observed in social media discourse. In conclusion, the effectiveness and flexibility of existing and advanced cyberbullying detection systems can be improved by exploring these directions in future research to contribute to safer online environments.

VIII. CONCLUSION

The current state of research involves the application of ensemble learning in cyberbullying detection and proposes a framework that can be considered a potential solution to this still-growing problem. This study harnessed three distinct models, BERT, LSTM, and Random Forest, to categorise textual data into three primary classes: normal, hate, and offensive. The performance of the models was established by applying several metrics to understand their usefulness in supporting the final classification.

Moving from a three-class classification to a binary classification necessitated the identification of the normal percentage from the obtained confidence scores.

This step was very important, especially when the output of the confidence scores was low, but the hate and offensiveness scores were high, suggesting cases of cyberbullying.

To overcome this challenge, this study proposed the concept of weighted averages of the models' outputs, which provided the basis for setting a threshold for categorising interactions as normal as opposed to those that possess features of cyberbullying. This ensemble approach led to a vast improvement in accuracy and ensured that misclassification rates were greatly reduced, contributing to a safer online space. In summary, the proposed model and the results of this research contribute to recognising the potential of advancing ensemble learning methods in solving real-life issues, such as cyberbullying detection. Furthermore, the findings give the reader an understanding of research areas in this evolving field and the fact that strategies for handling cyberbullying are dynamic.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been sponsored or funded by any organization or agency. The independence of this

research is a crucial factor in affirming its impartiality, as it has been conducted without any external sway.

- **Ethical Approval and Consent to Participate:** The data provided in this article is exempt from the requirement for ethical approval or participant consent.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Authors Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Patchin, Justin & Hinduja, Sameer. (2006). Bullies Move Beyond the Schoolyard A Preliminary Look at Cyberbullying. Youth Violence and Juvenile Justice. 4. 148-169. DOI: https://doi.org/10.1177/1541204006286288
2. Rosa, Hugo & Salgado Pereira, Nádia & Ribeiro, Ricardo & Ferreira, Paula & Carvalho, Joao & Oliveira, Sofia & Coheur, Luisa & Paulino, Paula & Veiga Simão, Ana Margarida & Trancoso, Isabel. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior. 93. 333-345. DOI: https://doi.org/10.1016/j.chb.2018.12.021
3. K. S. Alam, S. Bhowmik and P. R. K. Prosun, "Cyberbullying Detection: An Ensemble Based Machine Learning Approach," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 710-715, Doi: https://doi.org/10.1109/ICICV50876.2021.9388499
4. Abdullah, Alqahtani., Mohammad, Ilyas. (2024). A Machine Learning Ensemble Model for the Detection of Cyberbullying. DOI: 10.48550/arxiv.2402.12538 https://doi.org/10.5121/ijaia.2024.15108
5. Pankaj, Shah., Shivali, Chopra. (2024). Mixed Language Text Classification Using Machine Learning: Cyberbullying Detection System. 514-518. DOI: https://doi.org/10.1201/9781003405580-83
6. Jinan, Redha, Mutar. (2024). Cyberbullying Messages Detection Using Machine Learning and Deep Learning. International journal of advances in scientific research and engineering, 10(03):19-29. DOI: https://doi.org/10.31695/IJASRE.2024.3.3
7. K. S. Raj, K. Tej, N. K. S, S. K. T and S. Vajipayajula, "Ensemble Techniques for Malicious Threat Detection," 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2024, pp. 1543-1545, DOI: https://doi.org/10.1109/ICICT60155.2024.10544694
8. Prasad, K. L., Anusha, P., Rao, M., & Rao, Dr. K. V. (2019). A Machine Learning based Preventing the Occurrence of Cyber Bullying Messages on OSN. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 3, pp. 1861–1865). DOI: https://doi.org/10.35940/ijrte.a9164.078219
9. Jalda, C.S., Polimetla, U.B., Nanda, A.K., Nanda, S. (2024). A Comparison Study of Cyberbullying Detection Using Various Machine Learning Algorithms. In: Sathees kumaran, S., Zhang, Y., Balas, V.E., Hong, Tp., Pelusi, D. (eds) Intelligent Computing for Sustainable Development. ICICSD 2023. Communications in Computer and Information Science, vol 2122. Springer, Cham. DOI: https://doi.org/10.1007/978-3-031-61298-5_4
10. Bhagyashree, Kadam. (2023). Cyberbullying Detection using Machine Learning Algorithms. International Journal For Science Technology And Engineering, 11(5):1326-1328. DOI: https://doi.org/10.22214/ijraset.2023.51749
11. Muneer A, Alwadain A, Ragab MG, Alqushaibi A. Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT. Information. 2023; 14(8):467. DOI: https://doi.org/10.3390/info14080467
12. Ali, A., & Syed, A. M. (2022). Cyberbullying Detection using Machine Learning. Pakistan Journal of Engineering and Technology, 3(2), 45–50. DOI: https://doi.org/10.51846/vol3iss2pp45-50
13. Patil, P., Raul, S., Raut, D., & Nagarhalli, T. (2023). Hate Speech Detection using Deep Learning and Text Analysis. 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), 322–330. DOI: https://doi.org/10.1109/iciccs56967.2023.10142895
14. Hondor, Saragih., Jonson, Manurung. (2024). 1. Leveraging the BERT Model for Enhanced Sentiment Analysis in Multicontextual Social Media Content. Jurnal Manajemen Informatika C.I.T. Medicom, DOI: https://doi.org/10.35335/cit.Vol16.2024.766.pp82-89
15. Amisha, Sharma., Diya, Khajuria., Ayushi., Ritu, Rani., Garima, Jaiswal., Mala, Saraswat. (2023). LSTM-Based Model for Classification of Tweets. 1-7. DOI: https://doi.org/10.1109/ASIANCON58793.2023.10270665
16. Yamaguchi, A., Margatina, K., Chrysostomou, G., & Αλέτρας, N. (2021). Frustratingly Simple Pretraining Alternatives to Masked Language Modeling. cornell university. DOI: https://doi.org/10.48550/arxiv.2109.01819
17. Sun, Y., Hao, C., Zheng, Y., & Qiu, H. (2021). NSP-BERT: A Prompt-based Few-Shot Learner Through an Original Pre-training Task--Next Sentence Prediction. cornell university. DOI: https://doi.org/10.48550/arxiv.2109.03564
18. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv (Cornell University). DOI: https://doi.org/10.48550/arxiv.1810.04805
19. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735
20. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Ensemble learning. In Elsevier eBooks (pp. 479–501). DOI: https://doi.org/10.1016/b978-0-12-804291-5.00012-x
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017b, June 12). Attention Is All You Need. arXiv.org. https://arxiv.org/abs/1706.03762
22. Hoque, M. N., & Seddiqui, M. H. (2024). Detecting cyberbullying text using the approaches with machine learning models for the low-resource Bengali language. IAES International Journal of Artificial Intelligence, 13(1), 358. DOI: https://doi.org/10.11591/ijai.v13.i1.pp358-367
23. Chen, S., He, K., & Wang, J. (2024). Chinese Cyberbullying Detection Using XLNet and Deep Bi-LSTM Hybrid Model. Information, 15(2), 93. DOI: https://doi.org/10.3390/info15020093
24. Shibly, F. H. A., Sharma, U., & Naleer, H. M. M. (2022). Performance Comparison of Machine Learning and Deep Learning Algorithms in Detecting Online Hate Speech (pp. 695–706). Springer Nature Singapore. DOI: https://doi.org/10.1007/978-981-19-2821-5_59
25. Farasalsabila, F., Utami, E., & Hanafi, H. (2024). Deteksi Cyberbullying Menggunakan BERT dan Bi-LSTM. Jurnal Teknologi, 17(1). DOI: https://doi.org/10.34151/jurtek.v17i1.4636
26. Sunitharam, Dr. C., Nandini, P. S., & K, R. (2023). Detection of Cyber-Bullying Through Sentimental Analysis. In International Journal of Soft Computing and Engineering (Vol. 13, Issue 1, pp. 16–20). DOI: https://doi.org/10.35940/ijsce.a3594.0313123
27. Angelis, J. D., & Perasso, G. (2020). Cyberbullying Detection Through Machine Learning: Can Technology Help to Prevent Internet Bullying? In International Journal of Management and Humanities (Vol. 4, Issue 11, pp. 57–69). DOI: https://doi.org/10.35940/ijmh.k1056.0741120
28. Sharma, P. (2023). Advancements in OCR: A Deep Learning Algorithm for Enhanced Text Recognition. In International Journal of Inventive Engineering and Sciences (Vol. 10, Issue 8, pp. 1–7). DOI: https://doi.org/10.35940/ijies.f4263.0810823
29. Prashar, S., & Bhakar, S. (2019). Real
Time Cyberbullying Detection. In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 2, pp. 5197–5201). DOI: https://doi.org/10.35940/ijeat.b4253.129219

AUTHOR'S PROFILE

Mohammed Hisham Saeed, is a dedicated computer science engineering student at Muffakham Jah College of Engineering and Technology, with a strong focus on software development and artificial intelligence. He has experience in Python, TensorFlow, Hugging Face, and the MERN stack. He has led teams on various projects, winning multiple hackathons and delivering innovative solutions. His strengths include analytical thinking, problem-solving, and effective communication, and he excels in working collaboratively to achieve results. Passionate about technology, he constantly seeks opportunities to learn and grow while contributing to impactful projects.

Shakaib Ahmed Mohammed, is a dedicated computer science engineering student at Muffakham Jah College of Engineering and Technology, with a strong focus on machine learning. As an experienced machine learning enthusiast with a deep interest in artificial intelligence and the MERN stack, he has won multiple competitions for his innovative solutions. His strengths include critical thinking, creative problem-solving, and working effectively in collaborative environments. With hands-on experience in machine learning, Shakaib is skilled at tackling complex challenges and developing impactful solutions. Passionate about AI and full-stack development, he continuously seeks opportunities to apply his knowledge to research and projects. His drive for continuous learning allows him to adapt to the fast-paced advancements in technology and contribute meaningfully to the field.

Taufeeq Noamaan, is an aspiring software developer with a passion for creating impactful solutions through technology. With hands-on experience in developing AI-driven products and building CI/CD pipelines, he thrives in collaborative environments where innovation is key. As a GitHub Campus Expert and Microsoft Learn Ambassador, he has guided peers in app development and showcased projects at various tech expos. Proficient in languages like Java and Python and frameworks such as Flutter and MERN, Taufeeq is always eager to tackle challenging projects. His research interests focus on AGI, Human-Computer Interaction, ethics in AI, and multi-agent systems.

Rania Mehreen Farooq, is pursuing a bachelor’s degree in computer science and engineering at Muffakham Jah College of Engineering and Technology, affiliated with Osmania University. She has developed a strong programming and software development foundation, having excelled in courses such as Data Structures and Algorithms, Software Engineering, and Database Management Systems. Rania's research interests include artificial intelligence, machine learning, and data mining. She has actively participated in several hackathons, demonstrating her ability to innovate and collaborate while developing solutions to real-world challenges in a competitive environment. Outside academics, Rania enjoys building personal projects and engaging with the tech community. She is passionate about using technology to solve real-world problems and aspires to make a meaningful impact in the industry.

Mohammed Abdul Raheem, is an Assistant Professor with over ten years of experience in the field of Computer Science and Engineering, specialising in teaching Artificial Intelligence and Machine Learning. He holds a B.Sc., MCA, and M.Tech (CSE), which have provided him with a robust academic and technical foundation. His areas of expertise also include Web Programming, Software Engineering, and Computer Architecture. Passionate about fostering growth in his students, he encourages them to explore and excel in AI and ML. His research interests focus on computer networks and wireless networks, and he continues to explore emerging innovations. Mohammed is always open to collaboration and engaging in forward-thinking research initiatives.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.