International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

Strengthening Cyberbullying Detection with Ensemble Learning - Safeguarding Online Interactions Amongst Youth

Prashant Agrawal[1], Awanit Kumar[2], Arun Kr. Tripathi[3]

Submitted: 26/01/2024 **Revised**: 04/03/2024 **Accepted**: 12/03/2024

Abstract: Cyberbullying remains a pressing concern in the digital age, posing significant threats to the well-being of young individuals who engage in online interactions. Social networking platforms, while offering invaluable educational and social benefits, also harbor hidden dangers due to the cloak of anonymity they provide to perpetrators of cyberbullying. This paper presents an innovative machine learning strategy to address the issue of cyberbullying detection on social networking platforms, with a particular focus on enhancing the safety of online interactions among youth. In this study, we propose the application of ensemble learning, a potent method in the realm of machine learning, to enhance the precision and resilience of cyberbullying detection. The motivation behind this choice is two-fold. First, cyberbullying is a multifaceted problem, encompassing a wide range of behaviors and expressions. No single machine learning model can capture the full spectrum of cyberbullying instances effectively. Ensemble learning addresses this limitation by combining the strengths of multiple models, each specializing in different facets of cyberbullying behavior, thereby bolstering the detection process. Second, the intrinsic challenges of identifying cyberbullying, exacerbated by the veil of online anonymity, necessitate a nuanced approach. Ensemble learning, by aggregating the predictions of diverse models, provides an opportunity that will diminish instances of both false positives and false negatives, thereby achieving a more dependable cyberbullying detection system.

Keywords: *Social Networking Platform, Cyberbullying (CB), Machine Learning (ML) Techniques, Ensemble Learning (EL).*

1. Introduction

While CB encompasses various forms, it typically refers to persistent and deliberate abusive online behavior aimed at harassing or harming others. The ubiquity of social networking platforms has led to a surge in CB incidents. Traditional strategies for combating CB, including online conduct rules, human moderators, and blacklisting offensive language, struggle to cope with the escalating volume, speed, and diversity of data generated on these platforms. Consequently, novel machine learning-based models have emerged. Conducting a thorough literature review, this paper explores the application of soft computing techniques in identifying CB across various domains within social networking platforms. The goal is to comprehend the theoretical, research, and practical trends in this area. The meta-analysis of individual study findings indicates that utilizing soft computing methodologies for CB detection provides an intelligent analytical framework, enabling the anticipation of bullying behaviors in both textual and non-textual content on social networking platforms. Leveraging unstructured web data, this research area focuses on comprehending the extensibility of human expressions, holding promise for practical applications. The vast expanse of online communication platforms has provided a wealth of data reflecting the diverse spectrum of human interactions, offering unprecedented insights into our collective behavior. From sentiment analysis to opinion mining, this endeavor has yielded valuable knowledge applicable in domains ranging from marketing to public sentiment analysis. However, amidst this remarkable progress, there exists a pressing concern that threatens the safety and well-being of those who engage in online interactions, particularly among young individuals— CB. The pervasiveness of social networking platforms, coupled with the cloak of anonymity they afford to their users, has amplified the challenge of identifying and predicting CB behavior. Anonymous perpetrators can engage in harmful activities without fear of accountability, making traditional methods of monitoring and intervention less effective. These digital spaces, while offering significant educational and social advantages for young people, also expose them to potential dangers that can have far-reaching consequences. While researchers have shown interest in addressing CB, considerable research gaps persist. Notably, the challenge of accurately identifying and predicting CB behavior in cases of anonymous perpetrators remains a formidable obstacle. The dynamic and multifaceted nature of CB necessitates innovative approaches that can adapt to evolving tactics and behaviors employed by bullies.

[1] *Department of Computer Science and Engineering, Sangam University, Atoon, Rajasthan 311001, India.*
Email: prashant.agraw@gmail.com
https://orcid.org/ 0000-0002-7890-024X
[2] *Department of Computer Science and Engineering, Sangam University, Atoon, Rajasthan 311001, India.*
Email: awanit.kumar@sangamuniversity.ac.in
http://orcid.org/0000-0002-5867-6092
[3] *Department of Computer Applications, KIET Group of Institutions, Delhi-NCR,Ghaziabad, 201206 India.*
Email: mailtoaruntripathi@gmail.com
http://orcid.org/ 0000-0001-5138-2190

In recent years, EL [1] methods have emerged as a propitious avenue for improving the detection and identification of CB. These techniques harness the power of combining multiple models to enhance predictive accuracy and robustness. Although EL [2] has demonstrated great promise, the field of CB detection is still in its infancy when it comes to applying advanced techniques like deep learning, neural networks, ensemble methods, evolutionary computing, and hybrid models like neuro-fuzzy systems. These approaches have not been extensively explored in this regard. This paper addresses the urgent need to advance the state of CB detection by delving into the possibilities presented by ensemble learning, deep learning [3], and other advanced techniques. Our research aims to bridge these research gaps, offering a comprehensive and innovative approach to make online interactions safer, particularly among youth. By expanding our understanding of these techniques and their application in the context of CB, we aspire to contribute to the development of effective solutions that can protect young individuals from the perils of CB while preserving the many benefits of online engagement. A variety of ML models [4], such as Random Forest, Gradient Boosting, and potentially neural networks, are included into our suggested EL methodology. These models are trained on an annotated dataset comprising text content, user interactions, and contextual features to capture the nuanced characteristics of CB incidents. The final prediction is synthesized through a weighted combination of individual model outputs. We anticipate that this novel approach will contribute to the development of a more effective and reliable CB detection system, mitigating the harm inflicted on young individuals in online spaces. While our initial results are promising, ongoing refinement and optimization are essential to meet the evolving challenges posed by CB in the digital age. This study aims to provide the way for safer and more secure online interactions among youth, fostering an environment where the positive attributes of social networking platforms can flourish while minimizing the associated risks.

The paper's structure is outlined as follows: In Section 2, we provide an extensive examination of CB and its consequences, with a specific focus on its impact within the realm of online interactions among young individuals. Section 3 is dedicated to elucidating the methodologies and techniques applied in our study, encompassing our research approach, inquiries, identified research gaps, and the specific problem statement we aim to address. In Section 4, we delve into a comprehensive review of existing research within the domain of CB detection. Section 5 unveils our novel approach to tackling the problem, revealing the architecture and core algorithms of our proposed model. In Section 6, we meticulously assess the performance of our model and dissect the results obtained during our testing. Section 7 concludes the paper by summarizing our main findings, discussing their implications, and outlining potential directions for future research and development, underscoring the enduring relevance of our work.

1.1 Background

A. Cyber Bullying

CB [5], a contemporary manifestation of harassment and abuse, has arisen as a significant concern in the digital era. It involves the use of online communication platforms, such as social networking sites, text messages, or emails, to target individuals with harmful, hostile, or threatening messages. CB can encompass a wide range of behaviors, including online harassment, public shaming, doxing, and the dissemination of false information. The consequences of CB are far-reaching, affecting the mental and emotional well-being of victims, their social interactions, and their overall quality of life. The prevalence of online anonymity further exacerbates the problem, as it provides a shield for perpetrators, making them difficult to identify and hold accountable.

B. Machine Learning

ML [6] has rapidly evolved as a powerful tool to address complex and dynamic challenges in various domains that enable computers to learn from data and make predictions or decisions without explicit programming. In the context of CB, ML holds great promise for automating the process of detecting and responding to instances of online harassment. By analyzing vast datasets, ML models can uncover patterns and identify potential CB behavior, assisting in its mitigation and prevention. However, the success of ML in this domain hinges on the choice of appropriate algorithms, the quality of training data, and the adaptability of the models to ever-evolving forms of CB.

C. Ensemble Learning

EL [7] is a subfield of ML that has gained prominence for its ability to improve predictive accuracy and robustness. It operates on the principle of combining multiple models, each designed to address different aspects of a problem, to generate a more reliable and effective overall prediction. In the context of CB detection, EL can play a pivotal role. CB is a multifaceted issue with a no one-size-fits-all solution. Different aspects of CB may require distinct ML models to capture effectively. EL leverages the strengths of these individual models, thereby offering a comprehensive approach to detect and combat various forms of online harassment. Additionally, it can help mitigate the impact of false positives and false negatives, which are common challenges in CB detection, by aggregating the predictions of diverse models.

This research background underscores the pressing need for effective CB detection mechanisms, the potential of ML in addressing this challenge, and the promising role that EL techniques can play in improving the accuracy and reliability of CB detection systems. In the subsequent sections, we delve into the specifics of our proposed methodology, highlighting the application of EL and advanced ML techniques to address the complex issue of CB on social networking platforms.

1.2 RESEARCH METHODOLOGY

Research Strategy

The research strategy for this study is multi-faceted and interdisciplinary, combining elements of computer science, machine learning, and social science. The primary aim is to develop an effective CB detection system for social networking platforms, with a focus on youth safety. The research approach involves a combination of literature review, data collection, model development, and evaluation. Our methodology also places a strong emphasis on the utilization of EL techniques, deep learning, and other advanced ML approaches to address the identified research gaps.

Research Questions

RQ1. What are the existing challenges and limitations in CB detection on social networking platforms?

RQ2. How can ensemble learning, deep learning, and advanced ML methods be harnessed to improve CB detection?

RQ3. What is the impact of anonymous perpetrators on the effectiveness of CB detection, and how can it be mitigated?

RQ4. How can we create a safer online environment for youth by addressing the identified research gaps?

RQ5.

Research Gaps

Several research gaps have been identified in the field of CB detection:

- The challenge of effectively identifying and predicting CB behavior, particularly in cases involving anonymous perpetrators.
- The limited exploration of advanced ML techniques including deep learning, neural networks, ensemble approaches, evolutionary computing, and hybrid models like neuro-fuzzy systems in the context of CB detection.
- The need for adaptable and robust models that can keep pace with the evolving tactics and behaviors employed by cyberbullies in the digital landscape.
- The necessity for comprehensive and interdisciplinary research that merges computer science, machine learning, and social science to provide a holistic approach to addressing cyberbullying.
- Because the data on Social Networking Platform is informal, simple, noisy and unstructured, detecting online bullying activities is difficult and computationally hard.
    - ✓ Jargons
    - ✓ Mal-formed
    - ✓ Colloquial words
    - ✓ Hinglish

Problem Statement

The prevalence of CB on social media platforms portrays a grave threat to the safety and well-being of young individuals engaging in online interactions. The challenge of identifying and predicting CB behavior, especially when it involves anonymous perpetrators, remains a formidable obstacle. The existing CB detection methods [8] often fall short in adapting to the ever-evolving tactics employed by cyberbullies, leaving a critical gap in safeguarding the online experiences of the youth. This research aims to bridge these gaps by developing and evaluating a CB detection system that leverages advanced ML techniques like deep learning and ensemble learning. We attempt to use the ensemble approach to ML in our suggested model. Use a variety of ML techniques [9], including Light Gradient Boosting Machine (LGBM), Logistic Regression (LR), AdaBoost (ADB), Random Forest (RF), Stochastic Gradient Descent (SGD), Naive Bayes (NB), and Support Vector Machine (SVM), on a worldwide dataset sourced from Twitter, Facebook, and Kaggle. Accuracy, precision, recall, and F1 are the performance metrics that will be used to assess the classifiers' recognition according to different severity levels.

1.3 Related Work

CB is an issue of growing concern in the digital age, and as such, a substantial body of research has emerged in the quest to develop effective tools and techniques for its detection and prevention. This section examines the relevant literature on CB detection, with an emphasis on ML techniques and methodologies as in *Table 1* that have opened the door for the use of EL methods

The related research showcases a rich tapestry of ML techniques, encompassing deep learning, transfer learning, conventional classifiers such as SVM and Naïve Bayes, ensemble methods, and sophisticated natural language processing (NLP) approaches. This diversity underscores the adaptability of the field in addressing the multifaceted challenge of CB detection. Researchers have cast their nets wide, harnessing data from a multitude of social media platforms including Twitter [10], YouTube, Facebook, Instagram, Formspring, and Reddit.

This mosaic of data sources underscores the necessity for models that can gracefully navigate the intricacies of different platforms, each with its unique language and behavior. The studies have shown a commitment to thorough evaluation, employing a spectrum of performance metrics ranging from familiar accuracy, precision, recall, and F1-score to more nuanced measures like AUC, MCC, KAPPA, and NPV. This comprehensive assessment ensures that model efficacy is scrutinized from multiple angles. The landscape of CB detection models boasts impressive achievements, with standout performers like SSA-DBN boasting remarkable accuracy rates of 99.983%. Additionally, Ensemble of LSTM and CNN demonstrates its mettle with an impressive accuracy rate of 98.46%, and the ensemble stacking with BERT delivers a remarkable F1-score of 0.964, precision of 0.950, and recall of 0.92. These high-performing models offer a glimpse into the future of robust CB detection systems.

There are a few significant gaps in the field of CB detection studies that need to be addressed and explored further. Firstly, the inconsistency in datasets poses a significant challenge for the field. Many studies utilize different datasets, creating hurdles in making direct model performance comparisons. To enhance the comparability of models, there is a need for standardization or the creation of benchmark datasets, allowing for more accurate assessments and evaluations.

Secondly, model interpretability is an important aspect that remains under-addressed. While many studies demonstrate the high performance of their models, few delve into the interpretability of these models. Understanding the rationale behind a model's classification decisions in the context of CB is vital for building trust and ensuring accountability in the detection process. Thirdly, there exists a notable gap in the realm of real-time detection. Although several models [11] achieve impressive accuracy rates, their complexity can hinder their suitability for real-time monitoring on social media platforms.

The development of more efficient models capable of instantaneous detection is an area ripe for exploration.

Ref. No.	ML Technique	Dataset	Performance Criteria
[1]	Deep Learning Transfer Learning	Form Spring (about 12k posts) Twitter (around 16k posts). About 100k posts on Wikipedia	Precision Recall F1-score
[2]	Deep Learning Transfer Learning	YouTube (~54k posts by ~4k users)	Precision Recall F1-score
[3]	Salp Swarm Algorithm-Deep Belief Network (SSA-DBN)	Internet Data	Classification Accuracy (SSA-DBN 99.983 % accuracy rate)
[4]	Bidirectional gated recurrent unit (BiGRU) is one of three deep learning architectures. Block transformer CNN, or Convolutional Neural Network	Twitter	F1-score (Accuracy - 88%) Precision Recall Accuracy
[5]	Support Vector Machine (SVM) and Naïve Bayes	Twitter	Accuracy Precision Recall F1-score (about 71.25% more accurate than Naïve Bayes)
[6]	Random Forest Naive Bayes Logistic Regression	Twitter Facebook Instagram YouTube Snapchat	Precision Accuracy Recall F1-score (the Random Forest classifier produced an accuracy of 93% in the best scenario)
[7]	Deep Learning	Twitter	Accuracy

		Wikipedia Formspring	
[8]	Nelson Echolocation Algorithm - Recurrent Neural Networks, or DEA-RNN	10,000 tweets	Precision Accuracy Recall F1-score (DEA-RNN attained an average of 89.52% precision, 88.98% recall, 89.25% F1-score, and 90.94%) Specificity
[9]	Single and Double Ensemble-Based Voting Models, or SLE and DLE.	Facebook, Instagram, Twitter	96% accuracy rate, the SLE and DLE voting classifiers performed the best when K-Fold cross-validation was used in conjunction with TF-IDF (Unigram) feature extraction.
[10]	Naive Bayes Logistic Regression Random Forest XGBoost LSTM GRU (Gated Recurrent Unit) Support Vector Machine (SVM) Random Forest	48,000 tweets	F-1 score. (GRU deep learning model achieved an F-1 score of 0.92)
[11]	Deep Neural Network	Twitter	Accuracy
[12]	Neural Ensemble Method	Fine-Grained Cyberbullying Dataset (FGCD) Twitter parsed dataset	F1 rankings. (F1-scores of 87.28% on the Twitter parsed dataset and 95.59% and 90.65% for five and six classes, respectively, on the Fine-Grained Cyberbullying Dataset (FGCD))
[13]	Naïve Bayes, Support Vector Machines (SVM), Random Forest, Decision Tree, Tree Ensemble, and Logistic Regression.	Twitter Facebook Formspring MySpace	Accuracy Precision Recall F1-score
[14]	LightGBM Random Forest AdaBoost XGBoost Logistic Regression.	47,000 tweets	F-1 score for accuracy, precision, and recall. (LightGBM achieved accuracy rates of 85.5%, precision rates of 84%, recall rates of 85%, and an F-1 score of 84.49%, which is much better than comparable models.)
[15]	Voting-Based Ensemble Learning, Decision Tree Classifiers, K-Nearest Neighbor (KNN), and Logistic Regression	Twitter	F-1 score for accuracy, precision, and recall.
[16]	Natural Language Processing (NLP)		Accuracy Cross-Validation AUC Score

	Regression Analysis Logistic Random Forest Gradient Enhancement		
[17]	Ensemble Learning	Twitter	Accuracy
[18]	LSTM and CNN together; Naive Bayes; Random Forest; Decision Tree; Support Vector Machine (SVM); Convolutional Neural Network (CNN); Long Short-Term Memory (LSTM)	Kaggle dataset	Accuracy of 98.46%, the ensemble of LSTM and CNN using fastText word embedding demonstrated the best performance.
[19]	Ensemble stacking learning - BERT model called "BERT-M"	Twitter	0.964 F1-score, 0.950 precision, 0.92 recall, and a 3-minute detection time were recorded. Accuracy of the stacking ensemble learning strategy was 97.4%. 90.97% using a dataset that combines Twitter and Facebook.
[20	Bag of Words (BoW), TFIDF Support Vector Machine (SVM)	Social media platform "Reddit"	F-1 score for accuracy, precision, and recall
[21]	Logistic Regression (LR), Neural Network (NN), Decision Tree (DT), K Nearest Neighbor (KNN), Naive Bayes (NB), Quadratic Discriminant Analysis (QDA), and Support Vector Machine (SVM)	4 datasets – Twitter (8800 comments), Bayzick website, YouTube, Kaggle website (115,863 comments)	Specificity, F1-measure, Accuracy, Precision, Recall, and False Discovery Rate (FDR), False Negative Rate (FNR), False Positive Rate (FPR), Cohen's Kappa Coefficient (KAPPA), Area Under Curve (AUC), and Negative Predictive Value (NPV)

Table *1:* Exploring Cyberbullying Detection Literature with a Spotlight on ML Techniques and the Emergence of Ensemble Learning Approaches

Moreover, the generalizability of models remains an issue. Many studies are tailored to specific social media platforms or datasets, restricting the adaptability of these models to diverse online environments and the evolving nature of online communication. Research aimed at creating models with broader applicability is imperative. Ethical considerations are also conspicuously absent from much of the related work. CB detection models should adhere to principles of fairness and mitigate bias to avoid discriminatory outcomes. Extensive efforts in addressing ethical issues are needed to ensure that these models do not inadvertently target or harm specific groups. Lastly, the dynamic nature of online harassment poses a challenge in maintaining the relevance of detection models. As cyberbullies continuously adapt their tactics, there is a pressing need to develop models capable of identifying new forms of CB. Ongoing research focused on the robustness and adaptability of models is crucial to stay ahead of emerging threats in the digital sphere.

In summary, the related work demonstrates significant progress in the field of CB detection, with high-performing models across different ML techniques and datasets. However, there are still gaps in terms of standardization, model interpretability, real-time detection, generalization, ethical considerations, and adaptability to evolving tactics that require further research. To advance the field of CB detection, it is essential to address the identified gap areas and improve accuracy, interpretability, real-time monitoring, generalizability, ethical considerations, and adaptability to evolving threats. We propose the integration of EL as a comprehensive solution to these challenges. EL [12] has the potential to elevate the quality and effectiveness of CB detection systems, ultimately contributing to a safer online environment. This research initiative seeks to explore and harness the power of EL for the betterment of CB detection, culminating in a safer, more inclusive digital landscape for all.

2. Proposed Methodology

In this paper, we provide a machine learning-based method for identifying instances of CB on social media sites. This methodology in *Figure 1* makes use of ensemble learning, a potent ML tool. There are two reasons to use EL in this situation. First, CB is a complex and evolving problem, and no single ML model [13] is universally effective at detecting all forms of CB. By leveraging ensemble learning, we aim to integrate the best features of several ML models, each designed to capture different aspects of CB behavior. This diversity in models helps improve the overall accuracy and robustness of our detection system. Second, the cloak of anonymity that online platforms provide to cyberbullies makes their behavior challenging to identify. By merging the predictions of individual models, EL can help reduce false positives and false negatives, improving the accuracy of our CB detection method. An EL strategy is proposed that combines several models, such as random forests, gradient boosting, and neural networks.

These models will be trained on a labeled dataset of CB instances, utilizing a variety of features including text content, user interactions, and context. The final prediction will be based on a weighted combination of the predictions from these individual models. The findings of our investigation are expected to demonstrate the effectiveness of EL in improving the accuracy and reliability of CB detection on social networking platforms. While our preliminary results show promise, we acknowledge the need for continued refinement and enhancement in this critical area to better protect the youth from online harassment and create a safer online environment.

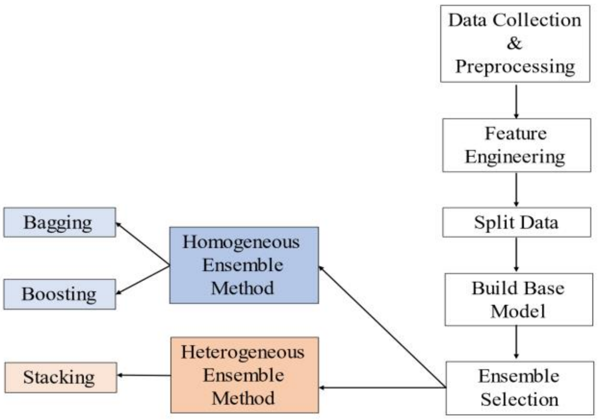
Data Collection & Preprocessing
Feature Engineering
Split Data
Build Base Model
Ensemble Selection
Homogeneous Ensemble Method
Bagging
Boosting
Heterogeneous Ensemble Method
Stacking

Figure 1: Ensemble Learning Process for Cyberbullying Detection

Building a base model for EL in CB detection using social media posts through twitter, YouTube, meta and so on. Utilize Word Embeddings, Doc2Vec, and Term Frequency-Inverse Document Frequency (TF-IDF) to transform the preprocessed text data into features that may be used for model training. Random Forest: Known for its resilience and capacity to manage high-dimensional data, Random Forest is an ensemble [14] of decision trees. Incorporating a diverse array of base models into the ensemble empowers CB detection system with a wide range of modeling techniques and approaches, fortifying its robustness and enhancing accuracy across varied tasks. Let's delve into the distinctive base models mentioned in *Table 2* suitable for EL in CB detection

Table 2: Exploring Distinctive Base Models for Ensemble Learning in Cyberbullying Detection

Base Model	Key Factors	Type of Datasets	Advantages	Disadvantages
Multinomial Naive Bayes	Probabilistic model	Text data	Simple and fast, works well with text data	Assumes independence of features (naive assumption)
Convolutional Neural Network	Deep learning model	Text data	Captures spatial features in text, effective for sequences	Requires large, labeled data - Computationally intensive
Support Vector Machine (SVM)	Linear classifier	Text data, high-dimensional	Effective in high-dimensional spaces - good generalization	Computationally intensive- Model tuning can be complex

Long Short-Term Memory (LSTM)	Recurrent neural net	Sequential text data	Captures sequential patterns - Suitable for variable length	Training can be slow- Risk of vanishing/exploding gradients
XGBoost	Gradient boosting	Structured data	Handles structured and unstructured data well - High accuracy	Parameter tuning can be complex- Can overfit if not tuned
Logistic Regression	Linear classifier	Text and structured data	Simple and interpretable model - Works well as a base model	Limited complexity for complex data
FastText	Text representation	Text data	Efficient text classification - Ability to handle sub words	May require preprocessing for special languages
K-Nearest Neighbors (KNN)	Instance-based model	Text data, numerical	Local pattern recognition - No training phase required	Sensitive to choose of k- Computationally intensive
Gaussian Mixture Model (GMM)	Probabilistic model	Clustering tasks	Modeling data distribution - Good for clustering	Prone to overfitting with too many components
Gated Recurrent Unit (GRU)	Recurrent neural net	Sequential text data	Captures sequential patterns efficiently - Short-term memory	May not capture long-term dependencies well
LightGBM	Gradient boosting	Structured data	Efficient and handles categorical data well - High speed	May require extensive parameter tuning
Ridge Regression	Linear model	Structured data	Simplicity and robustness as a base model	Limited complexity for complex data
Quadratic Discriminant Analysis (QDA)	Probabilistic model	Numerical	Models' covariance structure effectively - Non-linear	Sensitive to multicollinearity- Limited to small datasets
Randomized Decision Trees	Decision tree model	Structured data	Diversity in decision - making- Robust to outliers	May overfit if tree depth is not controlled

By including these diverse base models, proposed ensemble model becomes a versatile and dynamic CB detection system, capable of addressing an array of challenges with a multitude of modeling techniques and approaches. Although it's essential to select a combination of models that complement each other in terms of their strengths and characteristics. The selection of Multinomial Naive Bayes, Convolutional Neural Network (CNN) [15], Long Short-Term Memory (LSTM), Support Vector Machine (SVM) and Randomized Decision Trees creates a diverse ensemble with a mix of probabilistic, deep learning [16], recurrent, linear, and decision tree-based models. This diversity helps the ensemble effectively handle various characteristics of CB content, including text patterns, sequential behaviors, and high-dimensional feature spaces. It also provides a balance between interpretability and complexity, making the ensemble classifier [17] more robust and capable of addressing different aspects of CB detection.

Naive Bayes is known for its simplicity and efficiency in text classification tasks. It's a strong choice for handling text data, making it well-suited for identifying textual CB content. While it makes the naive assumption of feature independence, it can serve as a valuable baseline model. CNNs are powerful in capturing spatial features in text, making them effective for identifying patterns and subtle nuances in CB text. They excel in handling sequences of words, which is common in social media data. LSTMs, as recurrent neural networks, are proficient in capturing sequential patterns in text data. They are ideal for detecting the temporal aspects of CB, such as patterns of harassment over time. SVMs are excellent at separating classes in high-dimensional spaces, making them a valuable addition to the ensemble for CB detection. They provide robust generalization and can handle text and high-dimensional feature data effectively. Randomized decision trees offer diversity in decision-making and robustness to outliers. They can help balance the ensemble by providing an alternative approach based on decision trees, which are interpretable and can handle structured data. In order to identify CB in the dataset of social media messages,

2.1 Algorithmic Procedure

the algorithmic procedure listed below uses ML techniques [18] including Multinomial Naive Bayes, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Support Vector Machine (SVM), and Randomized Decision Trees:

```
Step 1: Load and Preprocess Dataset
    1.1 Read the dataset from
        'https://www.kaggle.com/trehansalil1/toxic-dataset/data'.
    1.2 Take the labels (y) and text (X) out of the dataset.
Step 2: Split Data
    2.1 Using a random state of 42 and a test size of 20%, divide the dataset into training and testing sets.
    2.2 Obtain the following: y_train, y_test, X_train, and X_test.
Step 3: Multinomial Naive Bayes
    3.1 Create a CountVectorizer to convert text data into numerical features.
    3.2 Use the CountVectorizer to transform the training and testing sets.
    3.3 Use the modified data to train a Multinomial Naive Bayes classifier.
    3.4 Predict the testing set (nb_predictions) using your predictions.
Step 4: Convolutional Neural Network (CNN)
    4.1 Increase the vocabulary's maximum word count (max features) to 10,000.
    4.2 Set the maximum length of a post (max_len) to 100.
    4.3 Make a tokenizer and fit the training set using the data.
    4.4 Create sequences out of text data and pad them to a predetermined length.
    4.5 Assemble an Embedding layer, Conv1D layer, GlobalMaxPooling1D layer, and Dense layer in a Sequential model.
    4.6 Use the Adam optimizer and binary cross-entropy loss to compile the model.
    4.7 Use the padded training data to train the model over three epochs.
    4.8 Utilizing the padded testing set (cnn_predictions), make predictions.

Step 5: Long Short-Term Memory (LSTM)
    5.1 Repeat steps 4.1 to 4.4 for the LSTM model.
    5.2 Using an LSTM layer, an embedded layer, and a dense layer, create a sequential model.
    5.3 Gather and use the padded training data for three epochs to train the LSTM model.
    5.4 Utilizing the padded testing set (lstm_predictions), make predictions.

Step 6: Support Vector Machine (SVM)
    6.1 To transform text input into TF-IDF features, create a TfidfVectorizer.
    6.2 Use the TfidfVectorizer to transform the training and testing sets.
    6.3 Use the converted data to train a Support Vector Machine classifier.
    6.4 Assign predictions to the svm_predictions testing
Step 7: Randomized Decision Trees
    7.1 Create a CountVectorizer to convert text data into numerical features.
    7.2 Use the CountVectorizer to transform the training and testing sets.
    7.3 Use the modified data to train a Random Forest classifier.
    7.4 Predict using the rf_predictions testing set.
Step 8: Evaluate Models
    8.1 8.1 Print the accuracy scores for each model on the testing set.
    8.2 8.2 Print classification reports for more detailed evaluation.
Step 9: End
```

2.2 Model evaluation

One useful method for assessing each base model's performance mentioned in *Table 3*, *of* EL approach is cross-validation. In the following code snippet, we perform 5-fold cross-validation for each base model, calculate the specified evaluation metrics, and store the results in a DataFrame.

```
Cross-Validation for Base Models
Step 1: Initialization
    1.1 Define a list of base models and their corresponding names.
    1.2 Initialize empty lists to store accuracy, precision, recall, and F1-score scores.
Step 2: Cross-Validation Loop
    2.1 For each base model and its name in the zip of base_models and model_names:
        2.1.1 Perform cross-validation with 5 folds.
        2.1.2 Calculate and append the mean accuracy score to the accuracy_scores list.
        2.1.3 Calculate and append the mean precision score to the precision_scores list.
        2.1.4 Calculate and append the mean recall score to the recall_scores list.
        2.1.5 Calculate and append the mean F1-score to the f1_scores list.
Step 3: Create Performance Table
    3.1 Create a DataFrame named performance_table using the collected scores.
    - Columns: "Model", "Accuracy", "Precision", "Recall", "F1-Score"
    - Data: model_names, accuracy_scores, precision_scores, recall_scores, f1_scores
Step 4: Display Performance Table
    4.1 Print the performance_table to display the evaluation metrics for each base model.
Step 5: End
```

The average values obtained from each model's 5-fold cross-validation are represented by these performance measures. The performance of each model in terms of accuracy, precision, recall, and F1-score for cyberbullying detection is shown in the table.

Model	Accuracy	Precision	Recall	F1-Score
Multinomial Naive Bayes	0.90375	0.911843	0.922747	0.917265
Convolutional Neural Network	0.91000	0.918789	0.927983	0.923366
Long Short-Term Memory (LSTM)	0.91125	0.924219	0.923770	0.923994
Support Vector Machine (SVM)	0.88250	0.899539	0.894647	0.897079
Randomized Decision Trees	0.89250	0.897054	0.916541	0.906614

Table 3: Model Evaluation Metrics

3. Conclusion and Future Work

The “detection of cyberbullying” domain on social networking platforms is positioned for continuous growth and innovation. The abundance of emerging computational models offers promising avenues for researchers to identify and predict CB behaviors, making it an active and dynamic area of study. This research compels investigators to seek models that seamlessly integrate natural language processing with the cognitive capabilities, intelligence, and self-tuning behavior of ML approaches [20]. The core principle of our ensemble is the thoughtful selection of models that complement each other in terms of their strengths and characteristics. The inclusion of Randomized Decision Trees, Support Vector Machines (SVM), Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Multinomial Naive Bayes has produced a diverse ensemble [21], encompassing probabilistic, deep learning, recurrent, linear, and decision tree-based models. This diversity equips our ensemble to tackle the multifaceted nature of CB content. It adeptly handles a range of characteristics, including text patterns, sequential behaviors, and high-dimensional feature spaces. Moreover, this amalgamation provides a crucial balance between interpretability and complexity, rendering the ensemble [19] more robust and capable of addressing different facets of CB detection.

While we have made significant strides in the realm of CB detection, it is essential to recognize that the work is far from complete. The ever-evolving landscape of online interactions and the inventive strategies employed by cyberbullies require continuous research and development. Although textual information has been the subject of most recorded work in CB detection, there is still a lot of unrealized potential for expanding into other media types like audio, video, and photos. Additionally, there is an intriguing opportunity to explore native or regional languages beyond English, as different linguistic contexts may introduce unique challenges and nuances in identifying CB. One particularly intriguing and evolving aspect of this field is the use of animated GIFs and memes as tools for targeting or humiliating individuals on social networking platforms. These emerging trends pose new challenges and opportunities for researchers to develop innovative methods to detect and combat CB in its various forms. Future research should explore the integration of multiple data modalities, including text, audio, video, and images, to create more comprehensive and accurate CB detection models. This will enable a more holistic understanding of online interactions and potentially uncover subtler forms of CB.

Conflict of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

References

[1] Alam, K. S., Bhowmik, S., & Prosun, P. R. K. (2021, February). Cyberbullying detection: an ensemble based machine learning approach. In 2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV) (pp. 710-715). IEEE.

[2] Ahmed, T., Kabir, M., Ivan, S., Mahmud, H., & Hasan, K. (2021, December). Am i being bullied on social media? an ensemble approach to categorize cyberbullying. In 2021 IEEE international conference on big data (Big data) (pp. 2442-2453). IEEE.

[3] Agrawal, S., & Awekar, A. (2018, March). Deep learning for detecting cyberbullying across multiple social media platforms. In European conference on information retrieval (pp. 141-153). Cham: Springer International Publishing.

[4] Raj, M., Singh, S., Solanki, K., & Selvanambi, R. (2022). An application to detect cyberbullying using machine learning and deep learning techniques. SN computer science, 3(5), 401.

[5] Roy, P. K., Singh, A., Tripathy, A. K., & Das, T. K. (2022). Identifying cyberbullying post on social networking platform using machine learning technique. In Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2021 (pp. 186-195). Springer Singapore.

[6] Dalvi, R. R., Chavan, S. B., & Halbe, A. (2020, May). Detecting a Twitter cyberbullying using machine learning. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 297-301). IEEE.

[7] Roy, P. K., Singh, A., Tripathy, A. K., & Das, T. K. (2022). Cyberbullying detection: an ensemble learning approach. International Journal of Computational Science and Engineering, 25(3), 315-324.

[8] Mahmud, M. I., Mamun, M., & Abdelgawad, A. (2022, December). A deep analysis of textual features based cyberbullying detection using machine learning. In 2022 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT) (pp. 166-170). IEEE.

[9] Sharma, H. K., & Kshitiz, K. (2018, June). Nlp and machine learning techniques for detecting insulting comments on social networking platforms. In 2018 International conference on advances in computing and communication engineering (ICACCE) (pp. 265-272). IEEE.

[10] Alotaibi, M., Alotaibi, B., & Razaque, A. (2021). A multichannel deep learning framework for cyberbullying detection on social media. Electronics, 10(21), 2664.

[11] Mahat, M. (2021, March). Detecting cyberbullying across multiple social media platforms using deep learning. In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 299-301). IEEE.

[12] Chandrasekaran, S., Singh Pundir, A. K., & Lingaiah, T. B. (2022). Deep learning approaches for cyberbullying detection and classification on social media. Computational Intelligence and Neuroscience, 2022.

[13] Singh, N. K., Singh, P., & Chand, S. (2022, November). Deep Learning based Methods for Cyberbullying Detection on Social Media. In 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 521-525). IEEE.

[14] Liu, Y., Zavarsky, P., & Malik, Y. (2019). Non-linguistic features for cyberbullying detection on a social media platform using machine learning. In Cyberspace Safety and Security: 11th International Symposium, CSS 2019, Guangzhou, China, December 1–3, 2019, Proceedings, Part I 11 (pp. 391-406). Springer International Publishing.

[15] Dadvar, M., & Eckert, K. (2020). Cyberbullying detection in social networks using deep learning-based models. In Big Data Analytics and Knowledge Discovery: 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 22 (pp. 245-255). Springer International Publishing.

[16] Alotaibi, M., Alotaibi, B., & Razaque, A. (2021). A multichannel deep learning framework for cyberbullying detection on social media. Electronics, 10(21), 2664.

[17] Thanigaivel, S., Harshan, S., Syed Shahul Hameed, M., & Umadevi, K. S. (2021). Detection and prevention of cyberbullying using ensemble classifier. In International Virtual Conference on Industry 4.0: Select Proceedings of IVCI4. 0 2020 (pp. 323-333). Springer Singapore.

[18] Ahuja, R., Banga, A., & Sharma, S. C. (2021). Detecting abusive comments using ensemble deep learning algorithms. Malware Analysis Using Artificial Intelligence and Deep Learning, 515-534.

[19] Muneer, A., Alwadain, A., Ragab, M. G., & Alqushaibi, A. (2023). Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT. Information, 14(8), 467.

[20] Shakambhari, Raj, J. S., & Anantha Babu, S. (2022). Smart Cyberbullying detection with Machine Learning. In Disruptive Technologies for Big Data and Cloud Applications: Proceedings of ICBDCC 2021 (pp. 237-248). Singapore: Springer Nature Singapore.

[21] Azeez, N. A., & Fadhal, E. (2023). Classification of Virtual Harassment on Social Networks Using Ensemble Learning Techniques. Applied Sciences, 13(7), 4570.