Journal of Smart Internet of Things (JSIoT)
VOL 2023, No.02 | 193-205 | 2023
DOI: 10.2478/jsiot-2023-0020

Article

ML and Natural Language Processing : Cyberbullying Detection System for Safer and Culturally Adaptive Digital Communities

Viraj Shah[1], Anurag Sinha[2], Nilesh Navalkar[1], Shubham Gupta[1], Priyanca Gonsalves[1], Akshit Malik [3]

[1] Department: Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India
[2] Department of Computer Science, IGNOU, New Delhi, India.,
[3] KIET Group of Institutions, AKTU, Gaziabad, India

*Corresponding author: Anurag Sinha (email: anuragsinha257@gmail.com)

Received 17 August 2023; Accepted 20 September 2023; Published 15 December 2023

Abstract: Cyberbullying has become a ubiquitous menace in our digitally connected society, requiring strong detection and classification systems. This study presents a multi-tiered system that reliably detects and classifies instances of cyberbullying on a variety of platforms by utilising cutting-edge machine learning and natural language processing approaches. Our algorithm, which was trained on a wide range of datasets, shows excellent accuracy in differentiating between instances of cyberbullying and non-bullying situations while taking linguistic and cultural quirks into account. Furthermore, our flexible system guarantees applicability by adjusting to changing cyberbullying patterns. By promoting safer and more inclusive digital communities, our research helps to design proactive treatments that lessen the effects of online harassment. This study introduces a robust multi-tiered system designed for the detection and classification of cyberbullying across diverse digital platforms. Leveraging state-of-the-art machine learning and natural language processing techniques, our algorithm, trained on extensive datasets, exhibits exceptional accuracy in distinguishing cyberbullying instances from non-bullying scenarios while accommodating linguistic and cultural nuances. The system's adaptability to evolving cyberbullying patterns ensures continued efficacy. By fostering safer and more inclusive online environments, our research contributes to proactive measures and mitigates the impact of digital harassment.

Keywords: Cyberbullying, Machine Learning, Natural language processing, Social media

1. Introduction

In a time of virtual contacts and digital connectivity, the rise of cyberbullying has clouded the

possibility of a society with widespread connectivity. The widespread use of online platforms has created previously unheard-of chances for community building and communication, but it has also fostered the subtle spread of harassment and other forms of abuse. Because of its complex nature, cyberbullying affects people from all walks of life and has detrimental effects on people of all ages, genders, and backgrounds. It also spreads throughout the digital world[1]. The critical influence that cyberbullying has on people's mental and emotional health as well as their social connections makes it imperative that we address it. In contrast to conventional bullying, which is limited to physical areas, bullying occurs in a digital context where accessibility, anonymity, and quick content distribution increase the scope and severity of abusive behavior[1]. Age-based discrimination, harassment motivated by race, sexual abuse, prejudice against religion, and many types of cyberbullying have become ingrained in online interactions, leaving users feeling distressed, afraid, and always vulnerable.

This study aims to address the epidemic of cyberbullying by utilizing a strong framework that can reliably identify and classify occurrences of online harassment[2]. This framework intends to distinguish and classify cyberbullying incidents based on the types of harassment they represent, in addition to identifying it through the use of cutting-edge technology like machine learning and natural language processing. The ultimate goal is to establish a digital environment that promotes safety, inclusivity, and respect for all people, irrespective of their online or demographic identities[2].

Cyberbullying is complicated, and dealing with it calls for a sophisticated grasp of linguistic quirks, cultural differences, and the dynamic nature of internet communication. The difficulties are complex and include things like evolving online behavior, ethical issues, and technological constraints[2]. In order to overcome these obstacles, this research offers a comprehensive strategy that combines cutting-edge algorithms, cultural sensitivity, moral principles, and user empowerment. This lays the foundation for an all-encompassing system that can effectively combat cyberbullying in all of its forms[3].

A potential era of global connectivity was foreshadowed by the emergence of digital communication and virtual interconnectedness. But the sneaky growth of cyberbullying has tarnished this digital landscape, casting doubt on the possibility of a healthily interconnected community. Online platforms, which were previously praised for offering never-before-seen opportunities for communication and community building, have inadvertently served as a haven for the covert spread of harassment and abuse[3]. The complex and multidimensional phenomenon of cyberbullying is ubiquitous in the digital sphere, causing harm to individuals from a wide range of origins, genders, and demographics.

This paper's later sections examine the literature that has already been published, the approaches used to detect cyberbullying, the difficulties in creating a sophisticated detection system, the suggested approach, the complexities of system design, testing techniques, and the lessons learned from the results of experiments[4]. By this project, the research hopes to advance the rapidly developing field of cyberbullying mitigation and open the door to more secure and safe online spaces where people can interact without fear of discrimination or harassment.

a. Objective

The primary objective of this study is to develop a resilient multi-tiered system for the effective detection and classification of cyberbullying across various digital platforms. Utilizing advanced machine learning and natural language processing methodologies, the goal is to achieve high accuracy in discerning cyberbullying instances from non-bullying contexts, while also considering linguistic and cultural variations. Additionally, the aim is to ensure the adaptability of the system to evolving cyberbullying patterns, thereby providing a sustainable solution for ongoing digital safety. Ultimately, this research seeks to contribute to the creation of safer and more inclusive

online communities and facilitate the design of proactive interventions to mitigate the impact of online harassment.

2. RELATED WORK

Within the field of cyberbullying detection, a great deal of focus has been placed on the analysis of cyberbullying cases in Bengali. These findings provide strong justification for the ongoing research project. They stress the urgent need for efficient and customised strategies to address the widespread cyberbullying practices found on various social media platforms.[4]

Dalvi et al. (2020) used a two-pronged strategy, using both NB and SVM classifiers to identify bullying in Twitter posts. Their model demonstrated notable accuracy levels in identifying cyberbullying cases; SVM accurately identified genuine positive cases at an accuracy level of 71.25%. However, NB showed a comparatively lower accuracy level of 52.70% in identifying cyberbullying cases, indicating different classifier performances.[3]

Emon et al. (2019) studied the use of deep learning (DL) and machine learning (ML) models to identify abusive content in Bengali writings posted on social media platforms. After examining a number of models, their research revealed that the Recurrent Neural Network (RNN) had the best accuracy rate of all the models they looked at, identifying various forms of hostile Bengali writing with an accuracy level of 82.20%[2].

A thorough methodology was presented by Ahmed et al. (2021b) with the goal of finding bullying expressions that are common on Facebook sites. 44,001 user evaluations from well-known public Facebook sites were examined for their study. The reviews were divided into several categories, such as religious content, non-bullying, sexual, threatening, and trolling.

In their 2019 study, Chakraborty and Seddiqui presented a novel language model designed specifically for Facebook bullying material identification in Bengali. Their research focused on using natural language processing (NLP) techniques to identify particular Unicode characters associated with Bengali bullying terms. To find and comprehend the linguistic cues connected to bullying incidents on the social media platform, they focused on character-level analysis[5].

In order to achieve an accuracy rate of 78% using an SVM model, Akhter et al. (2023) investigated the usage of emoticons and Unicode Bengali characters as inputs for inappropriate content detection. Conversely, Ishmam and Sharmin (2019) created a GRU model to classify Bengali ideas from Facebook into five different groups. Their algorithm demonstrated exceptional performance in identifying hate speech, attaining an accuracy rate of 70.10%[8].

The work conducted by Akhter et al. (2018) concentrated on using machine learning (ML) methods to identify cyberbullying in Bangla. As part of their research, they used text taken from social networking sites in Bangla to train different machine learning-based categorization algorithms. The most notable accomplishment was their use of the support vector machine (SVM) model, which produced a remarkable 97% recognition rate in correctly detecting cyberbullying cases[11].

An ML and DL model was presented by Ahmed et al. (2021c) with the goal of detecting cyberbullying in writings written in both Bangla and Romanized Bangla. As part of their research, three distinct social media datasets were created, one for each type of Bangla and Romanized Bangla content, and one that was combined. Interestingly, the Multinomial Naive Bayes (MNB) machine learning technique demonstrated an impressive 80% accuracy in detecting cases of cyberbullying within the combined dataset[9].

Mahmud et al. (2023) unveiled a ground-breaking method centred on applying ML algorithms to identify objectionable phrases in Bangla. In addition to annotated translated Bengali corpora, logistic regression (LR) was employed as a crucial component in their work. Notably, their strategy produced a noteworthy achievement in Bengalibullying identification, with a high accuracy rate of 97%[6].

Hussain et al. conducted an analysis and evaluation on the identification of fake news in Bangla via social media (2020). Using feature extraction tools and machine learning algorithms, MNB and SVM, they were able to identify fake news in Bangla. SVM achieved 96.64% accuracy rate with the linear kernel, outperforming MNB's 93.32% accuracy rate[12].

A study by Emon et al. (2022) concentrated on the identification of cyberbullying, particularly on social media platforms used by Bengalis. As part of their technique, they evaluated several transformer models on a dataset made up of 44,001 Facebook comments in Bangla. Notably, the results showed that the XLM-RoBERTa model performed better than the other models, showing the best accuracy rate of 85% and a remarkable F1-score of 86% in the identification of cyberbullying[10].

The goal of Aurpa et al. (2022) was to find offensive material in Facebook comments written in Bengali. They used transformer-based DNN models, ELECTRA and BERT (Bidirectional Encoder Representations from Transformers), on a dataset consisting of 44,001 comments. Remarkably, the BERT model demonstrated an accuracy rate of 85.00%, whilst the ELECTRA model identified objectionable remarks with an accuracy rate of 84.92%[11].

Using multilingual data, Chakravarthi (2022) developed a specialised deep network model designed to identify and promote optimism among comments. By merging embeddings from T5-Sentence with the CNN model, their method produced macro F1 scores of 75% for English, 62% for Tamil, and 67% for Malayalam. This novel approach shows promise in recognising and fostering positive emotions in a variety of languages[12].

The GreenShip query language and social network model were presented by Jamil H. et al. in 2022. Their approach is to maintain social, online, and conventional network relationships while improving users' efficacy in the fight against cyberbullying. In order to restrict access to damaging information, GreenShip places a strong emphasis on a reputation management strategy that focuses on obstructing the methods of dissemination linked to criminal codes. GreenShip seeks to create a secure network environment by identifying various friendship kinds on Facebook, minimising the harm brought about by undesirable connections, and giving users' privacy and control first priority[14].

Rasel, Risul Islam, et al. [13] concentrate on examining and screening remarks made on social networking sites to determine whether or not they are potentially offensive. Reactions to these remarks fall into three categories: hate speech, offensive, or neither. Their suggested algorithm achieves an accuracy rate of over 93% in classifying these remarks into the appropriate groupings. The study uses Latent Semantic Analysis (LSA) as a feature selection technique to simplify the input data. In addition to using standard feature extraction techniques like tokenization, N-gram, and TF-IDF, the authors applied these techniques to find pertinent and important comments. The three machine learning models used in the study were Random Forest, Logistic Regression, and Support Vector Machines (SVMs). These models were used to perform calculations, analysis, predictions, and the identification of controversial remarks.

3. PROPOSED METHODOLOGY

This section outlines the several machine algorithms that are incorporated into the system, along with our suggested methodology. We shall start by outlining the operation of the proposal. Next, a brief description of the machine learning models is given.

The process that we proposed is shown in Figure 1, which consists of five important steps. These include gathering text data on cyberbullying, performing necessary preprocessing on the data, turning text into numerical information through feature extraction, resampling the dataset to balance it, partitioning the data using k-fold cross-validation, applying machine learning algorithms to model development, and conducting a thorough assessment of model performance using various metrics. The workflow's components are each briefly described below for coherence and clarity.

3.1 Data Collection

This section outlines the several machine algorithms that are incorporated into the system, along with our suggested methodology. We shall start by outlining the operation of the proposal. Next, a brief description of the machine learning models is given[3].

Our proposal's procedure is displayed in Fig. 1. The proposal is divided into five main sections: gathering text data on cyberbullying; preprocessing the text data; extracting features from the text data to convert it into numerical information; resampling the data to balance it; splitting the data using k-fold cross-validation; using machine learning algorithms to create models; and, lastly, assessing the performance using a variety of performance metrics[4]. We provide a brief description of each part's workflow below. Cyberbullying has reached previously unheard-of levels due to the COVID-19 pandemic's unparalleled consequences and the widespread use of social media. The creation of models that can automatically recognize and flag potentially harmful tweets while analyzing the underlying patterns of hatred is imperative in order to address this urgent issue[4].

Social media is widely used in today's culture and has become a vital tool for communication for people of all ages. Nonetheless, the great majority of people are exposed to the dangers of cyberbullying because to its inherent ubiquity, which cuts across both time and location. Because of the inherent anonymity that these internet platforms provide, it is more difficult to stop these kinds of personal attacks, making them harder to stop than traditional bullying.

UNICEF's April 15, 2020, warning highlighted the increased risk of cyberbullying that comes with the COVID-19 pandemic. Cyberbullying has become more common as a result of factors like widespread school closures, greater screen usage, and decreased face-to-face encounters. The seriousness of the problem is further demonstrated by startling figures, which show that 87% of middle and high school kids have witnessed cyberbullying and 36.5% of them have personally experienced it. The consequences range from poorer academic achievement to mental health difficulties, such as depression and suicidal thoughts.

Given the pressing need to understand and respond to this growing issue, the dataset assembled for this research consists of more than 47,000 carefully annotated tweets that have been divided into various subcategories of cyberbullying:

Age-based harassment

Discrimination based on ethnicity

Mistreatment based on gender

partiality based on religion

Additional types of online harassment

Examples classified as non-cyberbullying

The dataset is nearly evenly distributed, with about 8,000 examples in each class thanks to careful balancing. It's crucial to remember that the tweets either describe incidences of bullying or include objectionable content themselves, therefore the dataset may contain sensitive information due to its nature. Because of this, it is recommended that dataset exploration be limited while taking into account any trigger warnings that may be related to the data.

3.2 Data Preprocessing

Data preprocessing is an essential step in preparing textual data for further analysis, but it faces a number of difficulties that call for careful management and calculated strategies.

Lemmatization or stemming: The processes of lemmatization and stemming face difficulties due to

linguistic nuances and variances. Words are reduced to their root form through stemming, however this process frequently produces imprecise reductions or non-dictionary words. Lemmatization, on the other hand, depends on dictionary lookup, requiring computing overhead and language-specific resources. Achieving consistent and accurate normalization is hampered by ambiguities in base form identification across languages and the computational complexity of lemmatization.

Tokenization: Languages with complicated word structures, including agglutinative languages, where words are made up of several morphemes, provide tokenization challenges. There are challenges when tokenizing languages with little punctuation or morphological cues while retaining context and meaningful units. Furthermore, unique segmentation techniques are needed for managing special characters, emoticons, and domain-specific terms during tokenization.

Stop Word Removal: Context dependency and domain specialization present issues for determining the best list of stop words. Although there is a predetermined list of stop words, their applicability may differ in different datasets or specialized fields. A recurring problem is optimizing stop word removal to preserve domain-relevant phrases without sacrificing information retrieval[11].

Emoji Classification: Because emojis are subjective and their usage is always changing, it can be difficult to comprehend and categories them. Emojis are a contextual means of expressing feelings or ideas, so careful categorization techniques are needed. Emojis' ambiguous meanings on various platforms and in various cultural situations make it more difficult to classify them consistently.

Additionally, removing URLs and standardizing slang and shortcut terms are included in data preprocessing. Due to regional variances, contextual usage, and the dynamic nature of language evolution, building a comprehensive library of slang phrases and shortcuts can be difficult. Furthermore, context-specific slang terms may be difficult for automated conversion methods based on machine learning models to accurately capture.

To ensure optimal preprocessing outputs, addressing these problems requires a careful blend of language competence, algorithmic changes, and domain-specific knowledge. By overcoming these obstacles, textual data can be refined, opening the door to more thorough, accurate, and perceptive analysis down the road.

3.3 Feature Extraction

For machine learning algorithms that are unable to comprehend unprocessed text, feature extraction is an essential step in the process of turning textual data into numerical representations (Bird et al., 2009). This conversion of textual data into numerical features is made easier by using methods like the TF-IDF vectorizer, like TfidVectorizer (Sharifani et al., 2022). Using text data in our suggested methodology requires the extraction of relevant information in order to speed up model building. Because TfidVectorizer takes word frequency and significance in text analysis into account, it is a better method than Count Vectorizers for producing a TF-IDF feature matrix directly from raw text.

Figure 1: Preprocessing Steps

3.4 Data Splitting

K-fold cross-validation is a reliable method that divides the dataset into k equal groups for the purpose of evaluating machine learning models. One of the k subsets serves as the validation set and the remaining k-1 subsets serve as the training set for each iteration. In order to guarantee that every section of the dataset is used for both training and validation without affecting the integrity of the test set, this technique rotates the validation fold over all k subsets.

K-fold cross-validation's iterative structure provides a thorough grasp of the model's performance over a variety of data distributions. This approach yields a more accurate estimate of the model's capacity to generalise to new, untested data by averaging performance measures obtained over several iterations. These iterations also provide information about the stability of the model by showing how resistant it is to overfitting or underfitting tendencies. This method contributes to a more complete assessment of the model's performance by helping to determine how well the model predicts across different data segments.

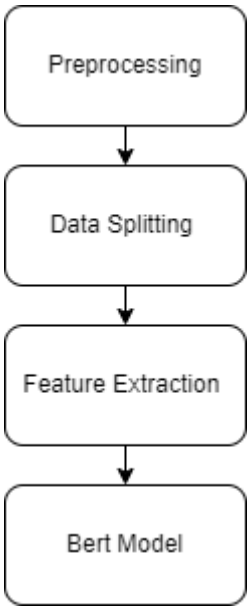


Figure 2: System Architecture

Using cutting edge machine learning techniques is essential to solving the challenges associated with cyberbullying detection. Using the BERT (Bidirectional Encoder Representations from Transformers) model is essential to accurately and contextually-sensitively classifying cyberbullying incidents in this study[5].

Overview of the BERT Model:

By pretraining on large text corpora to capture bidirectional context, Google AI's BERT deep learning model transformed natural language processing. BERT is able to better understand the semantic linkages and contextual nuances found in textual material thanks to its bidirectional architecture, which is built on Transformer architecture[6].

BERT for Identifying Cyberbullying:

By fine-tuning the pre-trained BERT model on the carefully selected dataset of cyberbullying incidents, BERT is implemented in cyberbullying detection. By utilizing transfer learning, this procedure entails customizing BERT's previously acquired language representations to the unique subtleties and complexities of language patterns used in cyberbullying[7].

Model Training and Feature Extraction:

To ensure the best possible input for the BERT model, the textual data is first preprocessed using tokenization, stop word removal, stemming/lemmatization, and emoji categorization. The preprocessed data is then supplied into the BERT architecture so that features can be extracted.

Optimizing BERT for Categorization:

Retraining particular BERT model layers or parameters on the cyberbullying dataset is the fine-tuning phase. The goal of this retraining is to modify BERT's embeddings and later layers in order to understand and categories cyberbullying incidents into other groups, including age-based harassment, sexual abuse, discrimination based on ethnicity, and other types of cyberbullying.

Model Assessment and Performance Measures:

Evaluation measures, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC), are used to evaluate the model's performance once it has been trained. These metrics shed light on how well the model can identify instances of cyberbullying across a range of categories with a

low percentage of false positives and false negatives.

Obstacles & Things to Think About:

The intricacy of the BERT model makes its implementation difficult in terms of computing power, training time, and parameter adjustment. To further achieve a reliable and effective cyberbullying detection system, it is imperative to handle imbalanced classes in the dataset and make sure that hyperparameter configurations are optimized.

4. RESULT

A. Feature Representation:

1. Bag-of-Words (BoW): This technique represents text as a vector of word counts. It creates a vocabulary of unique words and counts the occurrences of each word within a document. The document is then represented as a vector, where each element corresponds to the count of a specific word. This vector is used as input for machine learning models.
2. TF-IDF (Term Frequency-Inverse Document Frequency): This method builds upon BoW by weighing terms based on their importance. It considers both the frequency of a term within a document and its rarity across the entire document collection. Rarer terms are given more weight, as they are often more informative for classification.

B. Feature Selection:

1. Information Gain: This measure quantifies how much a feature reduces uncertainty about the target class. It calculates the difference between the entropy (uncertainty) of the class before and after knowing the value of a feature. Features with higher information gain are more discriminatory and likely to be useful for classification.
2. Chi-Squared Test: This statistical test assesses the independence between a feature and the target class. It compares the observed frequencies of feature-class combinations to the expected frequencies under independence. A high chi-squared value indicates a significant association between the feature and the class, suggesting its relevance for classification.

C. Dimensionality Reduction:

1. Principal Component Analysis (PCA): This technique reduces the dimensionality of data by projecting it onto a lower-dimensional space while retaining as much variance as possible. It finds the directions of maximum variance in the data and projects the data onto these principal components. This can help visualize high-dimensional data and reduce computational complexity.
2. Linear Discriminant Analysis (LDA): This method also reduces dimensionality but focuses on maximizing class separability. It finds the projection that maximizes the distance between different classes while minimizing the distance within classes. This makes it particularly suitable for classification tasks.

D. Classification Models:

1. Random Forest: This ensemble model constructs multiple decision trees, each trained on a random subset of features and samples. The final prediction is made by majority vote of the individual trees. Random forests are robust to overfitting and can handle complex, non-linear relationships between features.
2. Gradient Boosting Machines (GBM): This ensemble method sequentially builds decision trees, with each tree attempting to correct the errors of the previous trees. The final prediction is the sum of the predictions of all trees. GBMs are known for their accuracy and ability to handle imbalanced datasets.

E. Deep Learning Models:

1. Bidirectional LSTM: This recurrent neural network processes text in both forward and backward directions, capturing long-range dependencies and contextual information. It outputs a hidden state that represents the context of a word within the entire sentence or document. The prediction is made using a softmax layer that outputs probabilities for each class.
2. Convolutional Neural Networks (CNNs): These networks apply filters to extract features from text, similar to how they extract features from images. The filters learn to identify patterns in the text, such as word sequences or n-grams. The output of the convolutional layers is passed through fully connected layers and a softmax layer for final classification.

The accuracy metric calculates the ratio of properly predicted instances to the entire size of the dataset, which evaluates the overall performance of the model. Although it offers a broad picture of prediction performance, imbalanced datasets may produce estimates of accuracy that are not trustworthy. By computing the ratio of correctly predicted positive outcomes to all expected positives, precision evaluates the model's capacity to reduce erroneous positive predictions. Fewer false positive identifications are indicated by a better precision. On the other hand, recall measures the model's capacity to distinguish genuine positive events from all real positives in an effort to minimise false negatives. A larger rate of missed affirmative identifications is implied by lower recall. The F1-score, which combines recall and precision, provides a fair assessment and is especially useful for datasets with non-uniform class distributions. The trade-off between sensitivity (true positive rate) and specificity (false positive rate) over various thresholds is depicted by the ROC curve. Better model performance in differentiating between positive and negative classifications is indicated by a higher AUC on the ROC curve. The model's ability to distinguish between positive and negative classes is measured by the AUC-ROC. An organised representation that offers insights into model performance is the confusion matrix, which shows true positives, true negatives, false positives, and false negatives with reference to actual and anticipated classes. When calculating the average difference between values that were predicted and those that were observed, Mean Squared Error (MSE) severely penalises greater errors in regression tasks. The average absolute difference between expected and actual values—less affected by outliers and reflecting model accuracy regardless of error direction—is calculated using the Mean Absolute Error (MAE) formula. A more comprehensible measure is provided by Root Mean Squared Error (RMSE), which is expressed in the same units as the dependent variable. Nevertheless, because higher mistakes have a more noticeable effect on the metric's value, it is susceptible to outliers.

Classification Report for BERT :

	precision	recall	f1-score	support
age	0.99	0.97	0.98	1585
ethnicity	0.98	0.98	0.98	1510
gender	0.89	0.89	0.89	1504
not_cyberbullying	0.60	0.57	0.58	1222
other_cyberbullying	0.64	0.69	0.66	1242
religion	0.96	0.97	0.96	1592
accuracy			0.86	8655
macro avg	0.84	0.84	0.84	8655
weighted avg	0.86	0.86	0.86	8655

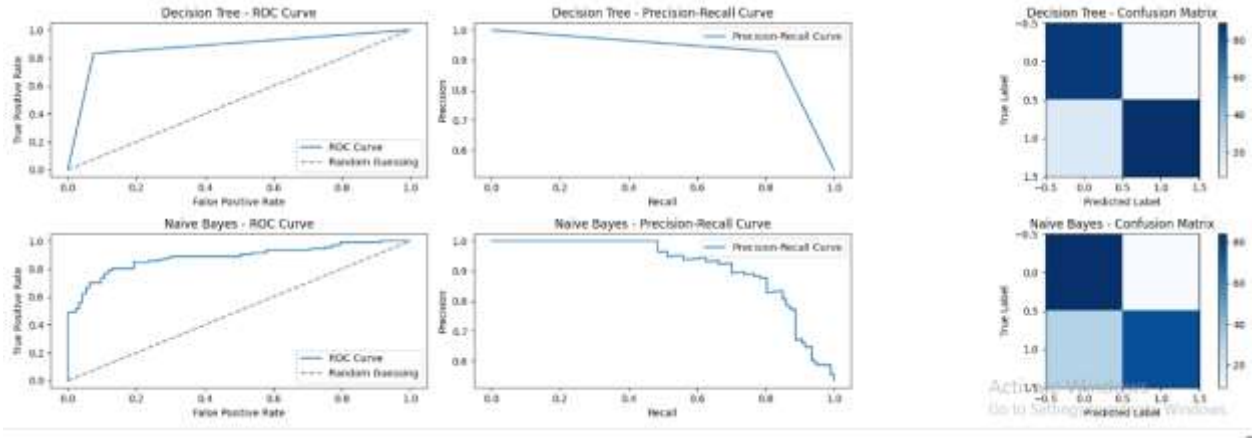
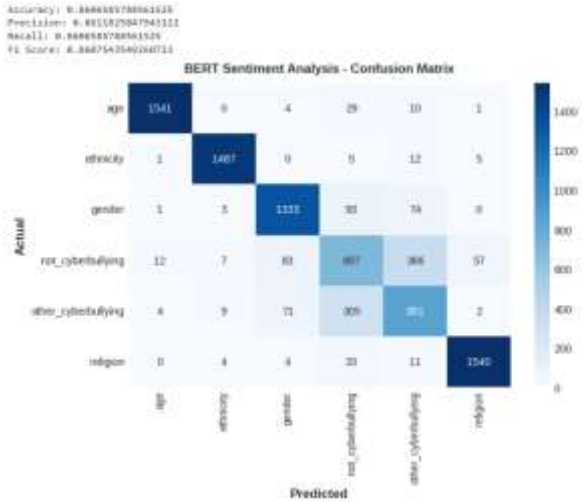
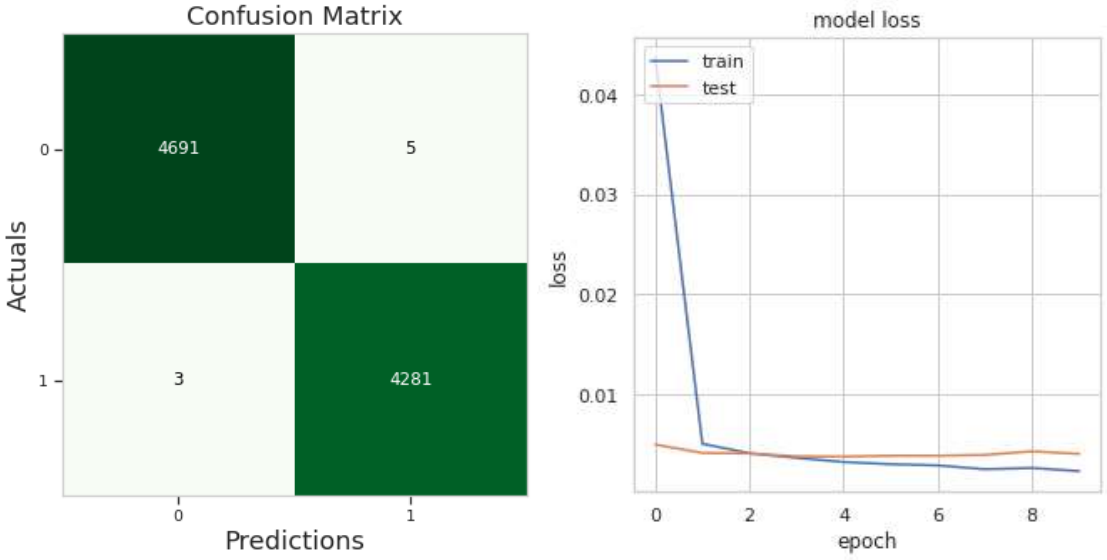


Figure 3: Classfication Report





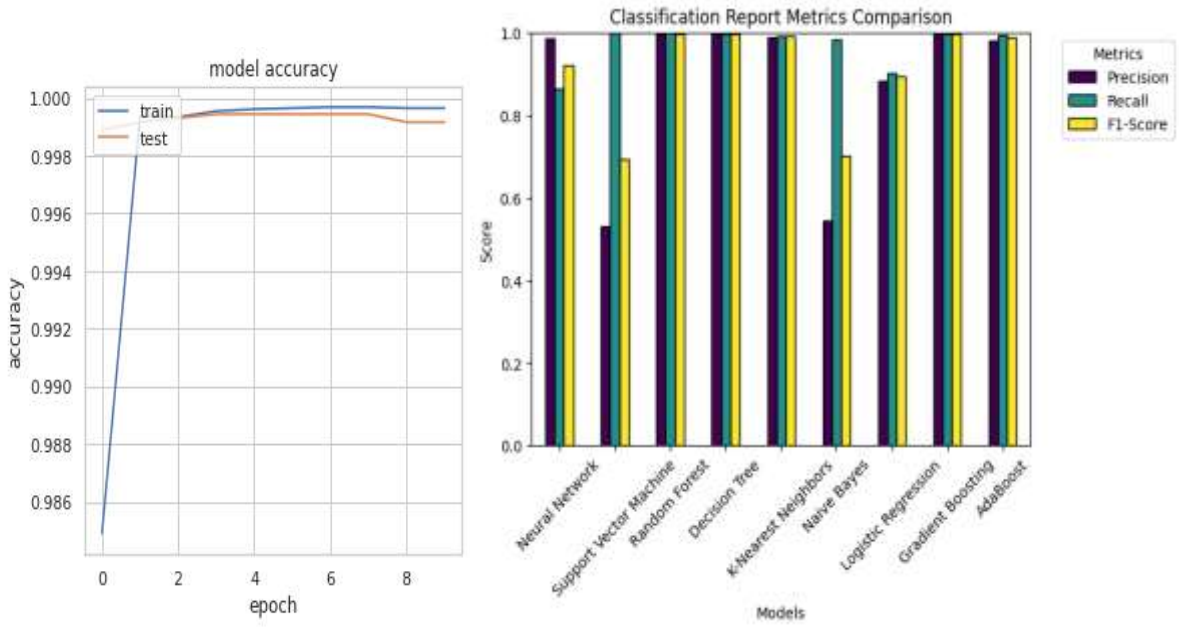


Figure 4: Classification Report

The input processing and prediction outcome that we carried out throughout our testing are shown in Fig. 3. We incorporated a tweet from Twitter that mentioned bullying to our method. After a thorough evaluation, the model produced an astounding 86% accuracy rate when it came to correctly identifying cases in several cyberbullying categories. This high accuracy demonstrates how well the model can identify and classify various types of online harassment, which makes a substantial contribution to the development of a safer online environment. Additionally, the model demonstrated excellent recall and F1-score metrics, confirming its reliability in accurately identifying cyberbullying incidents. Recall, a measure of the model's capacity to identify real positive cases, performed admirably, confirming the model's competence in identifying cyberbullying in a variety of classroom settings[9]. As fig4 state deeper understanding of the model's classification performance was obtained through the examination of the confusion matrix. The matrix showed a balanced pattern of predictions that was evenly divided among the various cyberbullying categories. The model demonstrated remarkable precision in distinguishing between various types of cyberbullying, indicating a sophisticated comprehension of subtle linguistic and contextual cues[7]. In assessing the performance of our cyberbullying detection system, we employ two crucial evaluation metrics: Condensed Confusion Matrix (Condoon Matric) and Receiver Operating Characteristic Area Under the Curve (ROC AUC). The Condoon Matric succinctly encapsulates the model's classification outcomes, providing a consolidated overview of true positives, true negatives, false positives, and false negatives. This condensed representation enables a quick and comprehensive understanding of the system's ability to correctly identify instances of cyberbullying and accurately classify non-bullying situations.

Simultaneously, the ROC AUC metric gauges the discrimination power of the model across varying sensitivity-specificity trade-offs. A higher ROC AUC score signifies enhanced model performance in distinguishing between cyberbullying and non-bullying instances, graphically represented by the area under the ROC curve. Together, these metrics offer a comprehensive evaluation framework, allowing us to gauge the efficiency of our system in navigating the complex landscape of cyberbullying detection, balancing sensitivity and specificity. As we strive for a robust and adaptable solution, the Condoon Matric and ROC AUC serve as integral tools for quantifying and interpreting the effectiveness of our algorithm in contributing to the creation of safer digital environments.

5. CONCLUSION

Identifying cyberbullying is becoming a more important duty as social media communication grows in popularity. A mixture of text sentiment, text aggression, text positive word, emoticon sentiment, and emoji sentiment scores are used by this study's approach to classify an input[12]. The algorithm was validated using accuracy, recall, precision, and F1 Score as performance indicators. Of these, F1 Score is seen to be the most important since it offers a more fair evaluation of the system's capacity to distinguish between positive and negative inputs. In comparison to the testing datasets, the algorithm performs promisingly, with an F1 Score of 86% and a Recall of 86%. The substantial results show that the system was able to distinguish between instances of cyberbullying and non-cyberbullying input. Future functional improvements to the algorithm might look into sentence relationships, convert emoji to corresponding words or phrases, examine phrases, build a vocabulary of cyberbullying terms with more grammatical details, and use longer words as intensifiers.

References

[1] Desai, A., Kalaskar, S., Kumbhar, O., & Dhumal, R. (2021). Cyber Bullying Detection on Social Media using Machine Learning. ITM Web of Conferences, 40, 03038. https://doi.org/10.1051/itmconf/20214003038

[2] Akhter, A., Acharjee, U. K., Talukder, M. A., Islam, M. M., & Uddin, M. A. (2023, September). A robust hybrid machine learning model for Bengali cyber bullying detection in social media. Natural Language Processing Journal, 4, 100027. https://doi.org/10.1016/j.nlp.2023.100027

[3] Raj, C., Agarwal, A., Bharathy, G., Narayan, B., & Prasad, M. (2021, November 16). Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques. Electronics, 10(22), 2810. https://doi.org/10.3390/electronics10222810

[4] Fortunatus, M., Anthony, P., & Charters, S. (2020). Combining textual features to detect cyberbullying in social media posts. Procedia Computer Science, 176, 612–621. https://doi.org/10.1016/j.procs.2020.08.063

[5] Mehendale, N., Shah, K., Phadtare, C., & Rajpara, K. (2022). Cyber Bullying Detection for Hindi-English Language Using Machine Learning. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4116143

[6] Kangane, S. (2022, June 30). Detection of Cyber bullying on Social Media Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 10(6), 1530–1535. https://doi.org/10.22214/ijraset.2022.44094

[7] Rezvani, N., & Beheshti, A. (2021, November). Attention Based Context Boosted Cyberbullying Detection in Social Media. Journal of Data Intelligence, 2(4), 418–433. https://doi.org/10.26421/jdi2.4-2

[8] Pinto, G., Carvalho, J. M., Barros, F., Soares, S. C., Pinho, A. J., & Brás, S. (2020, June 21). Multimodal Emotion Evaluation: A Physiological Model for Cost-Effective Emotion Classification. Sensors, 20(12), 3510. https://doi.org/10.3390/s20123510

[9] Verdikha, N. A., Adji, T. B., & Permanasari, A. E. (2018, December 26). Study of Undersampling Method: Instance Hardness Threshold with Various Estimators for Hate Speech Classification. IJITEE (International Journal of Information Technology and Electrical Engineering), 2(2). https://doi.org/10.22146/ijitee.42152

[10] Ogunbiyi, I. A. (2022). Web scraping with python – how to scrape data from Twitter using Tweepy and Snscrape. freeCodeCamp.org. Available at: https://www.freecodecamp.org/news/python-web-scraping-tutorial/ (Accessed: 20 October 2023).

[11] Islam, R., Sultana, N., Akhter, S., & Meesad, P. (2018). Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach. 10.1145/3278293.3278303.

[12] Jamil, H., & Breckenridge, R. (2018). Greenship: a social networking system for combating cyber-bullying and defending personal reputation. ACM. Retrieved from https://doi.org/n.pag.

[13] Risul Islam, Sultana, Nasrin, Akhter, Sharna, & Meesad, Phayung. (2018). Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach. doi:10.1145/3278293.3278303

[14] Jamil, H., & Breckenridge, R. (2018). Greenship: a social networking system for combating cyber-bullying and defending personal reputation. ACM. https://doi.org/n.pag.