International Journal of Advanced Technology in Engineering and Science
Vol. No. 12, Issue No. 05, May 2024
www.ijates.com
ijates
ISSN 2348 - 7550

SOCIAL MEDIA CYBERBULLYING DETECTION WITH MACHINELEARNING

Mrs. P. Vara Lakshmi[1], Mrs. K. Lakshmi[2]

A. Sukanya[3], B. Sowjanya[4], G. Venkata Shanmukha Sai Teja[5],

A. Venkateswar Reddy[6]

[1]Associate professor, Department of Electronics and Communication Engineering, Tirumala Engineering College,

[2]Associate professor, Department of Electronics and Communication Engineering, Tirumala Engineering College,

[3,4,5,6] UG Student Department of Electronics and Communication Engineering, Tirumala Engineering College.

Abstract:

The proliferation of social media platforms has enabled unprecedented levels of communication and idea-sharing across the globe, with nearly 4 billion users engaging on platforms such as Twitter, YouTube, Instagram, LinkedIn, Facebook, and WhatsApp. However, this connectivity has also facilitated the rise of online abuse, harassment, trolling, and cyberbullying, posing significant risks to individuals' mental and emotional well-being. Victims of cyberbullying often experience depression, engage in self-harm, and tragically, some even resort to suicide. In responseto these alarming trends, the development of effective methods for identifying cyberbullying content on social media has becomeimperative.

This research aims to address this pressing issue by proposing a classification system capable of detecting various forms of cyberbullying, including those based on age, gender, religion, ethnicity, and other pertinent factors. By harnessing advanced natural language processing (NLP) techniques and machine learning algorithms, this system seeks to analyze comments and messages on social media platforms to accurately identify instances of cyberbullying.

Keywords—*Cyber Aggressive, Support Vector Machine Random Forest, Logistic Regression, KNN, Machine Learning.*

I.INTRODUCTION

User Social media is a platform that allows people to post anything like photos, videos, documents extensively and interact with society. Using this media, malevolent users conduct unethical and fraudulent acts to hurt others feelings and damage their reputation. As the digital realm has grown and technology has progressed, cyberbullying has become relatively common, particularly amongst adolescents. We have investigated several machine learning algorithms, including Naive Bayes, Vector Machines for Support, Decision Tree, and Random Forest in the proposed cyberbullying detection model. write the introduction in 350 words ChatGPT Introduction: Social media platforms have emerged as powerful tools for communication and self-expression, facilitating the sharing of photos, videos, documents,and opinions on a global scale.

These platforms have transformed the way individuals interact with society, enabling

unprecedented levels of connectivity and engagement. However, this increased connectivity has also exposed users to new risks, including the proliferation of cyberbullying and online harassment.

Malevolent actors leverage the anonymity and reach of social media to perpetrate unethical and fraudulent acts aimed at harming others' feelings and damaging their reputation.

Cyberbullying, a form of online harassment, has become a prevalent issue in the digital age, with perpetrators using various tactics to intimidate, harass, or humiliate their targets. The consequences of cyberbullying can be severe, leading to emotional distress, psychological trauma, and even suicide in extreme cases. The prevalence of cyberbullying is particularly alarming among adolescents, who are among the most active users of social media platforms. As digital natives, adolescents are deeply immersed in online interactions, making them vulnerable to various forms of online abuse.

By leveraging the power of artificial intelligence and data analytics, these algorithms can analyze vast amounts of text data to identify patterns indicative of cyberbullying and flag potentially harmful content for further review. In our investigation, we have explored several machine learning algorithms, including Naive Bayes, Support Vector Machines,Decision Trees, and Random Forests, in the development of acyberbullying detection model. Each algorithm offers uniquestrengths and capabilities, and our goal is to evaluate their performance and effectiveness in accurately identifying instances of cyberbullying on social media platforms. Through our research, we aim to contribute to the development of robust and efficient tools for combating cyberbullying and creating safer online environments for all users. The proliferation of social media platforms has revolutionized communication and interaction, providing individuals worldwide with unprecedented opportunities to engage with diverse communities and express themselves freely. However, this newfound connectivity has also given rise to significant challenges, particularly concerning the detection and prevention of online harassment, cyberbullying,and discrimination. The proliferation of social media platforms has revolutionized communication and interaction, providing individuals worldwide with unprecedented opportunities to engage with diverse communities and express themselves freely. However, this newfound connectivity has also given rise to significant challenges, particularly concerning the detection and prevention of online harassment,cyberbullying, and discrimination.

By leveraging advanced natural language processing (NLP) techniques, cross-lingual data analysis, and machine learning algorithms, this multilingual system aims to overcome language barriers and cultural biases to provide robust and comprehensive insights into online behavior.

Frequent use SVM by researches shows that SVM is popular among other classifiers in supervised learning approach. SVM is suitable for high-skew text classification such as to detect cyber bullying using content based features.

II LITERATURE REVIE

The literature survey gives information about the projects thatare done earlier. This literature survey gives various perspectives regarding the project. There are several works on machine learning-based cyberbullying detection. A supervised machine

learning algorithm was proposed using a bag-of-words approach to detect the sentiment and contextual features of a sentence [9]. This algorithm shows barely 61.9% of accuracy. Massachusetts Institute of Technology conducted a project called Ruminate [10] employing support vector machine to detect cyberbullying of youtube comments. The researcher combined detection with common sensereasoning by adding social parameters. The result of this project was improved to 66.7% accuracy for applying probabilistic modelling. Reynolds et al. [11] proposed a language-based cyberbullying detection method which shows 78.5% of accuracy. The authors used the decision tree and instance-based trainer to achieve this accuracy. To improve cyberbullying detection, the author of the paper [12] has used personalities, emotion and sentiment as the feature. Several deep learning-based models werealso introduced to detect the cyberbullying. Deep Neural Network- based model is applied for cyberbullying detection by using real- world data [13]. The authors first analyze cyberbullying systematically then used transfer learning to do the detection task. Badjatiya et al. [14] has presented a method using deep neural network architectures for detecting hate speech. A convolutional neural network-based model has been proposed to detect cyberbullying [15]. The authors employed word embedding where similar words have similar embedding.

Altaf Mahmud et al [4] tried to differentiate between factual and insult statements by parsing comments using semantic rules, but they did not concentrate on comments directed towards participants and non- participants. Another work by Razavi et al [5]used a static dictionary and three level classification approach using bag- of-words features, which involved use of dictionary thatis not easily available. Dadvar et al., [7] analyzed the gender approach within the cyber bullying detection problem, applied to the social network My Space, a platform that offers an interactive, user-submitted community of friends with personal profiles, blogs,groups, etc.

Authors investigated the content of the posts written by the users but regardless of user's profile information. They used an SVM model to train a specific gender text classifier.

The dataset consists of about 381.000 posts. The results obtained by the gender based approach improved the baseline by 39% inprecision, 6% in recall, and 15%inF-measure.At MIT, Dinakar et aapplied different binary and multiclass classifiers on a manually labeled corpus of You Tube comments. We can observe that most of these studies are based on supervised approaches, and usually adopt pre-trained classifiers to solve the problem, typically based on SVM. Data are manuallylabelled using online services or custom applications, and are usually limited only to a small percentage.

NLP techniques are obviously wide adopted in all these works, due to the strict correlation between text analysis and cyber bullying detection. Mostly NLP tasks are For English language, a notable amount of research have been performed in text categorization or cyber bullying detection. The research included YouTube comments, each was manually labelled and then various binary and multiclass classifications were implemented. Among the different classification techniques,SVM gets notable attention due to better performance in varioustext classifications. A recent study reported that the NB Classifier can be used effectively for Indian text classification. Researchers showed that classification results of SVM was better than the NB method for Urdu language. Hence the aim of this research was to explore various machine learning algorithms.

International Journal of Advanced Technology in Engineering and Science
Vol. No. 12, Issue No. 05, May 2024
www.ijates.com
ijates
ISSN 2348 - 7550

III PROPOSED METHOD

Our proposed system is designed to tackle the pervasive issue of online harassment by leveraging advanced machine learning techniques to analyze tweet data, predict various attributes such as cyberbullying, gender, religion, age, and ethnicity, and provide actionable insights for platform administrators and users. The system comprises several key components:

Data Collection: We will gather tweet data from various sources, ensuring quality by implementing robust data collectionmethods and filtering out irrelevant or spammy content. Additionally, we will annotate the data with labels, indicating attributes such as cyberbullying and demographic information, to facilitate supervised learning.

Preprocessing: Before training our models, we will preprocessthe tweet data to clean and standardize the text, handle missing values, and balance the data to mitigate biases. This preprocessing step is essential for optimizing model performance and ensuring accurate predictions.

Model Training: We will explore a range of machine learning algorithms, including K-Nearest Neighbors (KNN), Gradient Boosting, Support Vector Machines (SVM), and others, for attribute prediction tasks. Through rigorous experimentation and optimization of hyperparameters, we aim to develop robust and accurate models capable of effectively detecting cyberbullying and predicting demographic attributes

Model Building: For each attribute prediction task, we will develop dedicated models and fine-tune them for better performance.

By training separate models for different attributes, we can tailorthe algorithms to specific prediction tasks and optimize their effectiveness in capturing nuances and complexities within the data.

IV BLOCK DIAGRAM

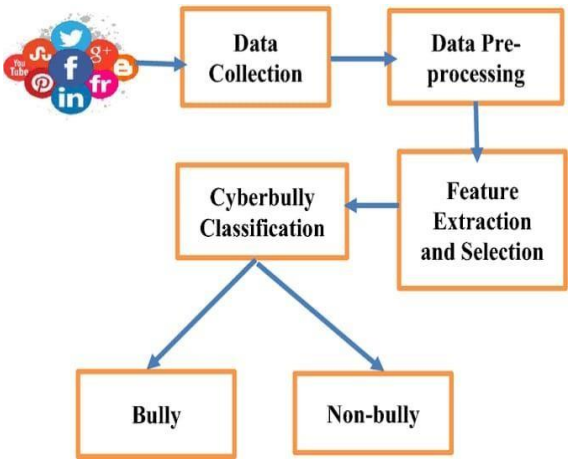


Fig: Block diagram

Website Integration: To make our system accessible and user- friendly, we will create a dedicated website where users can input tweet data for analysis. The website will integrate our trained models for predicting attributes, providing real-time insights into the content. Moreover, to ensure broader accessibility, our proposed system represents a comprehensive approach to combating online harassment by leveraging machine learning and data analytics to analyze tweet data, predict attributes.

Data Collection: Data collection in cyberbullying detection involves gathering information from various online sources where cyberbullying might occur, such as social media platforms, messaging apps, forums, and online communities. This process typically involves accessing public or authorized data through APIs (Application Programming Interfaces) provided by these platforms.

Cleaning: Removing irrelevant characters, such as punctuation and special symbols, and converting text to lowercase to ensure consistency.

Tokenization: Breaking down the text into individual words or tokens to facilitate analysis.
Stop word Removal: Eliminating common words (e.g., "and," "the,""is") that carry little semantic meaning and could add noise to the analysis.

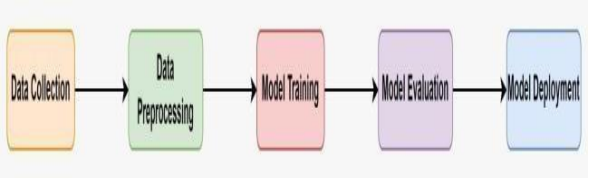
Feature extraction involves identifying and extracting relevant characteristics or patterns from the preprocessed data that can be used to distinguish between cyberbullying and non-cyberbullying instances. Here are some common techniques used for feature extraction.

Algorithms used:

- Random forest
- Logistic regression
- Support vector machine
- Gradient boosting
- Ada boosting (Adaptive boosting)
- KNN (k-nearest neighbors)

Machine learning, a fascinating blend of computer science and statistics, has witnessed incredible progress, with one standout algorithm being the Random Forest. Random forests or Random Decision Trees is a collaborative team of decision trees that work together to provide a single output.

Methodology:



Data Collection:

We gather a dataset of messages from various social media platforms, ensuring each message is labeled as either a bully message or a non-bully message. This dataset serves as the foundation for training and evaluating our classification model.

Preprocessing:

We preprocess the text data by cleaning it to remove noise,such as special characters, URLs, and stop words. Additionally, we perform stemming or lemmatization tonormalize the text andreduce word variations. This step helps standardize the text datafor more effective feature extraction and modeling.

Feature Extraction:

We extract relevant features from the preprocessed text data to represent each message. Common techniques include TF-IDF (Term Frequency-Inverse Document

International Journal of Advanced Technology in Engineering and Science
Vol. No. 12, Issue No. 05, May 2024
www.ijates.com
ijates
ISSN 2348 - 7550

Frequency), which captures the importance of words in a document relative to a corpus, word embeddings, which represent words as dense vectors in a continuous space, and n-grams, which capturesequences of words. Additionally, we may incorporate features related to message length, sentiment analysis, and lexical diversity to enhance the representation of messages.

Model Selection:

We choose a suitable machine learning algorithm for classification based on the nature of our dataset and the desired performance metrics. Common choices include Logistic Regression, Support Vector Machines, Random Forest, and Neural Networks. We may experiment with multiple algorithmsto identify the one that yields the best results.

Evaluation:

We evaluate the performance of our classification model using metrics such as accuracy, precision, recall, and F1 score on the testing data. These metrics provide insights into how wellour model is able to classify messages as bully or non bully is performed and it is having high evaluation.

RF Classification Report:

	precision	recall	f1-score	support
0	0.94	0.98	0.96	1602
1	0.96	0.89	0.93	1636
2	0.86	0.81	0.84	1514
3	0.51	0.45	0.48	1624
4	0.49	0.60	0.54	1594
5	0.93	0.94	0.94	1562
accuracy			0.78	9532
macro avg	0.78	0.78	0.78	9532
weighted avg	0.78	0.78	0.78	9532

Fig: Visual Representation after Text preprocessing

Deployment:

Once we are satisfied with the performance of our model, we deploy it to classify new messages in real-time on social media platforms. We integrate the model into a monitoring system to automatically flag or filter out bully messages, thereby helping to create a safer and more.

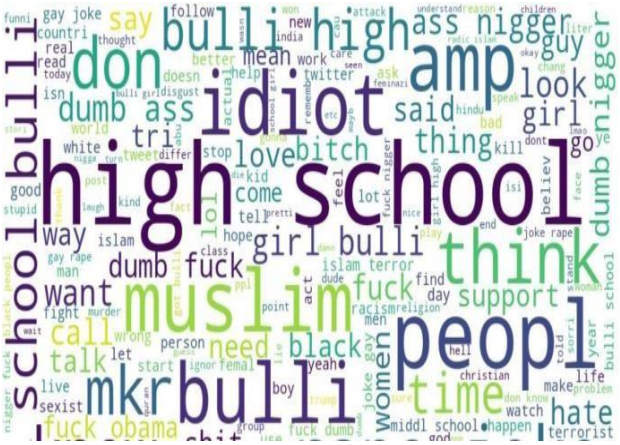


Fig: RF Classification Report

V Results and Discussion:

This code generates a word cloud visualization based on thefrequencies of words in the variable all_words. The Word

Cloud object is configured with specific parameters such as width,height, background color, and minimum font size.

Tabular Representation for Algorithms:

S.NO	ALGORITHM	PRECISION	RECALL	ACCURACY	F1 SCORE
1	RANDOM FOREST	0.94	0.98	0.78	0.96
2	LOGOSTUC REGRESSION	0.91	0.95	0.78	0.93
3	SUPPORT VECTOR MACHINE	0.96	0.97	0.80	0.97
4	GRADIENT BOOSTING	0.95	0.97	0.79	0.96
5	ADA BOOSTING	0.85	0.94	0.73	0.89

Table: Precision, Recall, Accuracy and F1 Score

VI . FUTURE S COPE AND CONCLUSION:

FUTURE SCOPE:

The future scope of the project involves embracing cutting-edge techniques and technologies to further enhanceperformance and address the evolving challenges of online communication.

CONCLUSION:

In conclusion, the project "Cyberbullying Detection Using Machine Learning" addresses a critical need to combat the pervasive and harmful phenomenon of cyberbullying across the internet. Cyberbullying poses serious threats to individuals' mental and emotional well-being, contributing to tragic outcomes such as suicides and depression. Recognizingthe urgency of this issue, the project has aimed to develop an automated system capable of accurately classifying commentsand messages as bullying or non-bullying and removing harmful content from web applications.

REFERENCES:

[1] Rice, Eric, et al. “Cyber bullying perpetration and victimization among middle-school students.” AmericanJournal of Public Health (ajph), pp. e66-e72, Washington, 2015.

[2] Bangladesh Telecommunication Regulatory Commission, http://www.btrc.gov.bd/content/internet-subscribers- Bangladeshjanuary-2018, [Last Accessed on 18 Mar 2018].

[3] Mandal, Ashis Kumar, Rikta Sen. "Supervised learning methods for Bangla web document categorization." International Journal of Artificial Intelligence & Applications,
IJAIA, Vol 5, pp. 5, 10.5121/ijaia.2014.5508

[4] Dani Harsh, Jundong Li, and Huan Liu, “Sentiment Informed Cyberbullying Detection in Social Media” JointEuropean Conference on Machine Learning and Knowledge Discovery in Databases. Spinger, Cham, 2017

[5] Dinakar, Karthik, Roi Reichart, and Henry Lieberman. “Modeling the detection of Textual Cyberbullying.” The Social Mobile Web

[6[K. Dinkar, R. Reichart and H. Liebernman, “Modeling the Detection of Textual Cyberbullying,” MIT. International Conference on Weblog nd Social Media. Barcelona, Spain, 2011.

[7] M. Dadvar and F.de Jong. 2012.”Cyberbullying detection: step toward a safer internet yard”. In Proceedings of the 21st International Conference on World Wide Web(WWW ’12 Companion).ACM, New York, NY, USA, 121-126

About Authors:

Mrs. P Varalakshmi is Currently working as Assistant Professor in Tirumala Engineering College. She received her MTech (Electronics & Communications) Degree from Chundi Ranganayakulu Engineering College.

A.sukanya currently studying B.Tech (Electronics & Communication Engineering) in Tirumala Engineering College, JawaharlalNehru Technological University Kakinada, Andhra Pradesh in the year 2024. She completed her Diploma in Sai TirumalaNVR Engineering College

B. Sowjanya currently studying B.Tech(Electronics & Communication Engineering) in Tirumala Engineering College, Jawaharlal Nehru Technological University Kakinada,
Andhra Pradesh in the year 2024. She completed her Diploma in Sai Tirumala NVR Engineering College.

G. Venkata Shanmukha Sai Teja currently studying B.Tech (Electronics & Communication Engineering) in Tirumala Engineering College, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh in the year 2024. He completed his
intermediate in Sri Chaitanya Junior College.

A. Venkateswar Reddycurrently studyingB.Tech (Electronics & Communication Engineering) in Tirumala Engineering College, Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh in the year 2024. He completed his intermediate in Junior College.