A Project report on

CYBERBULLYING DETECTION USING MACHINE LEARNING

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

B. Shravya
(20H51A05B5)
P. Sreenidhi
(20H51A0521)
K. Rahul Bharadwaj
(20H51A0538)

Under the esteemed guidance of
Dr. G. Ravi Kumar
Associate Professor

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY
(UGC Autonomous)
*Approved by AICTE *Affiliated to JNTUH *NAAC Accredited with A^+ Grade
KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2020- 2024

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Major project phase I report entitled **"CYBERBULLYING DETECTION USING MACHINE LEARNING"** being submitted by B.Shravya (20H51A05B5), P.Sreenidhi (20H51A0521), K.Rahul Bharadwaj(20H51A0538) in partial fulfilment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her my guidance and supervision.

The results embodies in this project report have not been submitted to any other University or Institute for the award of any Degree.

Dr. G. Ravi Kumar
Associate Professor
Dept. of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

ACKNOWLEDGEMENT

With great pleasure we want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Dr. G. Ravi Kumar, Associate Professor**, Department of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. Siva Skandha Sanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academics, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Major Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the **Teaching & Non- teaching** staff of Department of Computer Science and Engineering for their co-operation

We express our sincere thanks to **Shri. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, We extend thanks to our parents who stood behind us at different stages of this Project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

B.Shravya (20H51A05B5)
P.Sreenidhi (20H51A0521)
K.Rahul Bharadwaj (20H51A0538)

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Prior to the innovation of information communication technologies (ICT), social interactions evolved within small cultural boundaries such as geo spatial locations. The recent developments of communication technologies have considerably transcended the temporal and spatial limitations of traditional communications. These social technologies have created a revolution in user-generated information, online human networks, and rich human behavior-related data. However, the misuse of social technologies such as social media (SM) platforms, has introduced a new form of aggression and violence that occurs exclusively online. A new means of demonstrating aggressive behavior in SM websites are highlighted in this paper. The motivations for the construction of prediction models to fight aggressive behavior in SM are also outlined. We comprehensively review cyber bullying prediction models and identify the main issues related to the construction of cyber bullying prediction models in SM. This paper provides insights on the overall process for cyber bullying detection and most importantly overviews the methodology.

CHAPTER 1
INTRODUCTION

CHAPTER 1
INTRODUCTION

1.1. Problem Statement

Cyberbullying has become a pervasive issue in online communities, affecting individuals across various age groups and social platforms. We specifically reviewed four aspects of detecting cyber bullying messages by using machine learning approaches, namely, data collection, feature engineering, construction of cyber bullying detection model, and evaluation of constructed cyber bullying detection models. With the increasing prevalence of social media and online interactions, cyberbullying has emerged as a serious concern, causing harm to individuals and communities. Identifying and mitigating instances of cyberbullying in a timely manner is essential to create a safer online environment. The harmful effects of cyberbullying on mental health and well-being are well-documented. As the volume of online interactions continues to grow, manual moderation becomes increasingly challenging and impractical.

1.2. Research Objective

The primary objective of this research is to develop an accurate and robust cyberbullying detection model leveraging machine learning techniques. Analyze the model's ability to detect cyberbullying in different contexts, such as sarcasm, cultural nuances, and evolving forms of online communication. The study aims to explore the effectiveness of different feature representations, including TF-IDF, word embeddings, and syntactic features, in improving the accuracy of cyberbullying detection. Evaluate how well the model performs when applied to different social media platforms and online communities with varying communication styles.Additionally, it seeks to investigate the potential benefits of multimodal approaches, integrating various types of media such as text, images, and videos, for a more comprehensive detection system. Compare the performance of the developed model against existing cyberbullying detection systems, both machine learning-based and rule-based.

1.3. Project Scope and Limitations

Project scope:

The primary scope of this research is to develop an accurate and robust cyberbullying detection model leveraging machine learning techniques. The main aim of over project is to develop an advanced cyberbullying detection system leveraging machine learning techniques. The study aims to explore the effectiveness of different feature representations, word embeddings, and syntactic features, in improving the accuracy of cyberbullying detection. The system will encompass the detection of diverse forms of cyberbullying, including but not limited to harassment, hate speech, threats, and personal attacks. Data collection will be conducted from publicly available sources, adhering to privacy and ethical standards. The project scope extends to thorough documentation, performance evaluation, and a feedback mechanism for iterative improvement. Additionally, it seeks to investigate the potential benefits of multimodal approaches, integrating various types of media such as text, images, and videos, for a more comprehensive detection system.

Limitations:

Language is constantly evolving, and new slang, acronyms, or coded language may emerge, making it challenging for models to keep up. A model trained on older data may struggle to detect newer forms of cyberbullying. Cyberbullying can occur in various forms such as text, images, audio, and videos. Detecting cyberbullying across these different modalities requires specialized models, and combining them adds complexity.

Despite limitations, CYBERBULLYING DETECTION USING MACINE LEARNING remains a valuable resource and support system for Social Media users.

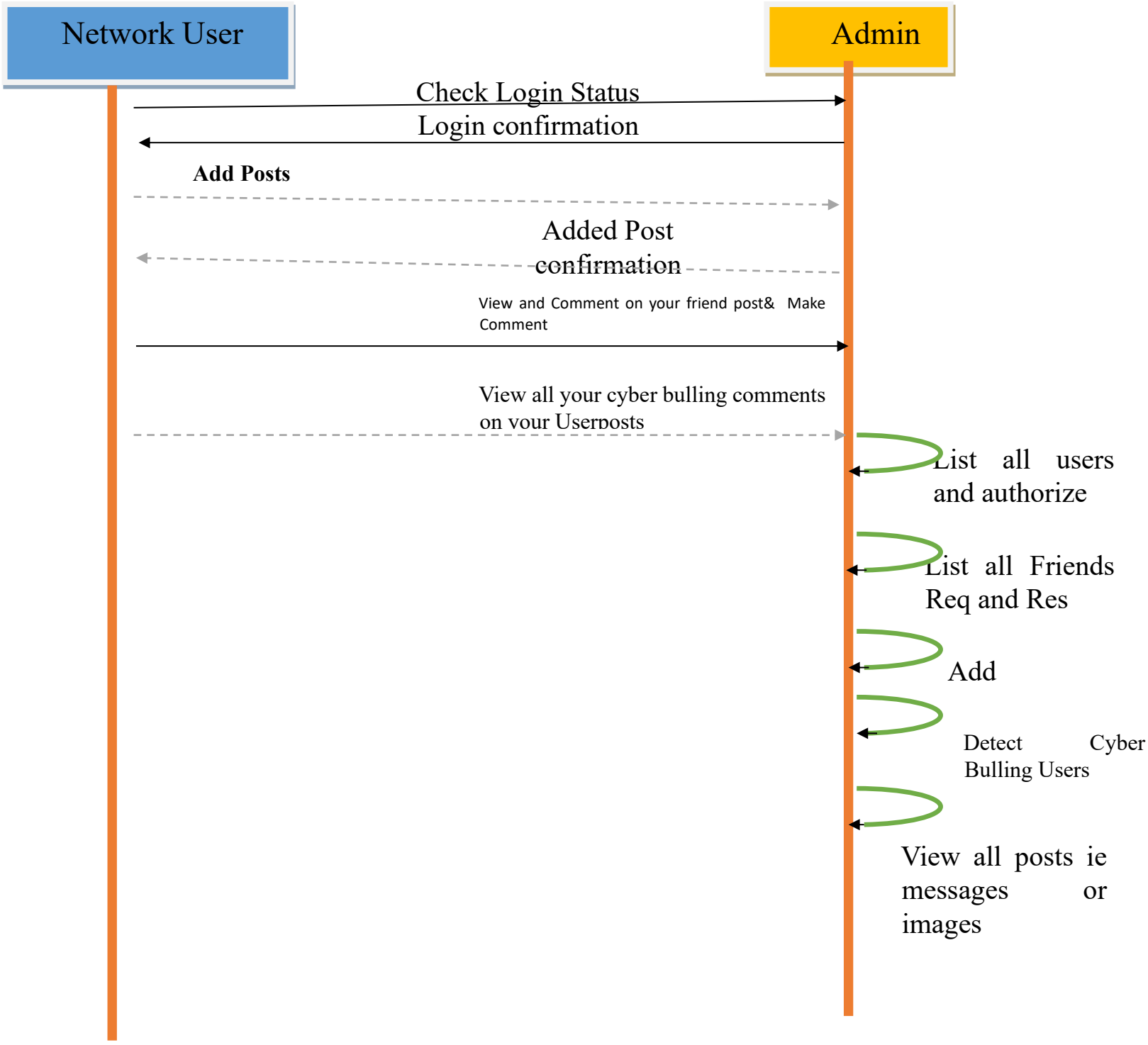

Fig no 1.1 View on CyberBullying detection using Machine Learning

CHAPTER 2
BACKGROUND WORK

CHAPTER 2
BACKGROUND WORK

2.1. Perspective API

2.1.1. Introduction

The Perspective API, developed by Jigsaw (a subsidiary of Google), is a powerful tool in the realm of content moderation. It utilizes machine learning models to assess the level of toxicity in user-generated content, allowing platforms to automatically identify and mitigate potentially harmful or abusive language. The API's primary goal is to enhance online conversations by promoting constructive and respectful interactions while minimizing the impact of cyberbullying, hate speech, and other forms of harmful communication . The Perspective API has been integrated into various online platforms, including social media sites, news organizations, forums, and more. It serves as a valuable tool for content moderation, working alongside human reviewers to help maintain a safer and more constructive online environment. By employing these machine learning models, the Perspective API is capable of assigning a toxicity score to user-generated content, indicating the likelihood that it contains elements of harmful language or behavior. This score can then be used by online platforms and communities to automatically flag or filter out content that may be considered toxic, thereby fostering more civil and respectful online interactions. For developers and organizations interested in implementing the Perspective API, Jigsaw provides comprehensive resources, including documentation, tutorials, and a supportive community forum. This empowers them to effectively utilize the API to enhance content moderation practices in their respective online platforms.

2.1.2. Merits, Demerits and Challenges

Merits:

- Perspective API provides an automated way to filter and moderate user-generated content, reducing the need for manual review and intervention.
- It can handle large volumes of user-generated content in real-time, making it suitable for social media platforms and other online communities with high user engagement.
- The API allows for customization, enabling platforms to set their own thresholds for what constitutes "toxic" content, making it adaptable to different community standards.
- The API benefits from contributions and feedback from a wide community of developers and moderators, which helps improve its performance over time.
- It can be integrated into existing content moderation systems, complementing other tools and methods used by platforms.

Demerits:

- Like all automated content moderation systems, Perspective API may produce false positives (incorrectly flagging non-toxic content) and false negatives (failing to detect toxic content), which can impact user trust and system effectiveness.
- Determining what constitutes "toxic" or harmful content can be subjective and context-dependent. Different communities may have varying standards.

  Certain forms of toxicity, such as sarcasm, humor, or cultural references, may be challenging for the model to accurately interpret.
- As with any content moderation system, there are concerns about user privacy and potential biases in the model's predictions, which must be carefully managed.
- Certain forms of toxicity, such as sarcasm, humor, or cultural references, may be challenging for the model to accurately interpret.

Challenges:

- New forms of toxicity and abuse constantly emerge online, and adapting the model to recognize these evolving patterns is an ongoing challenge.
- Perspective API primarily focuses on text-based content, but addressing toxicity in images, videos, and audio remains a complex problem.
- The model may not be equally effective across different cultures and languages, as it may not generalize well to regions or communities with distinct online communication norms.
- Adversaries may attempt to manipulate content in a way that evades detection, requiring continuous efforts to enhance the model's resilience.
- Striking the right balance between allowing free expression and protecting users from harm is a nuanced challenge, and setting appropriate toxicity thresholds is critical.

2.1.3. Implementation

First step is to integrate the Perspective API into the platform or application where content moderation is needed. This typically involves setting up API endpoints and establishing communication between the platform and the Perspective API servers. To use the Perspective API, developers need to obtain authentication credentials, such as an API key, from Jigsaw. These credentials are used to authorize requests and ensure secure communication between the platform and the API. When a user generates content, such as a comment or message, it is submitted to the Perspective API for analysis. The text is usually sent in the form of an HTTP POST request to the API endpoint. The Perspective API processes the submitted text and returns a toxicity score. This score represents the likelihood that the content contains toxic or harmful language. The score is typically a value between 0 and 1, where higher values indicate higher toxicity. Based on the returned toxicity score, the platform can set a threshold to determine what level of toxicity is acceptable within their community.

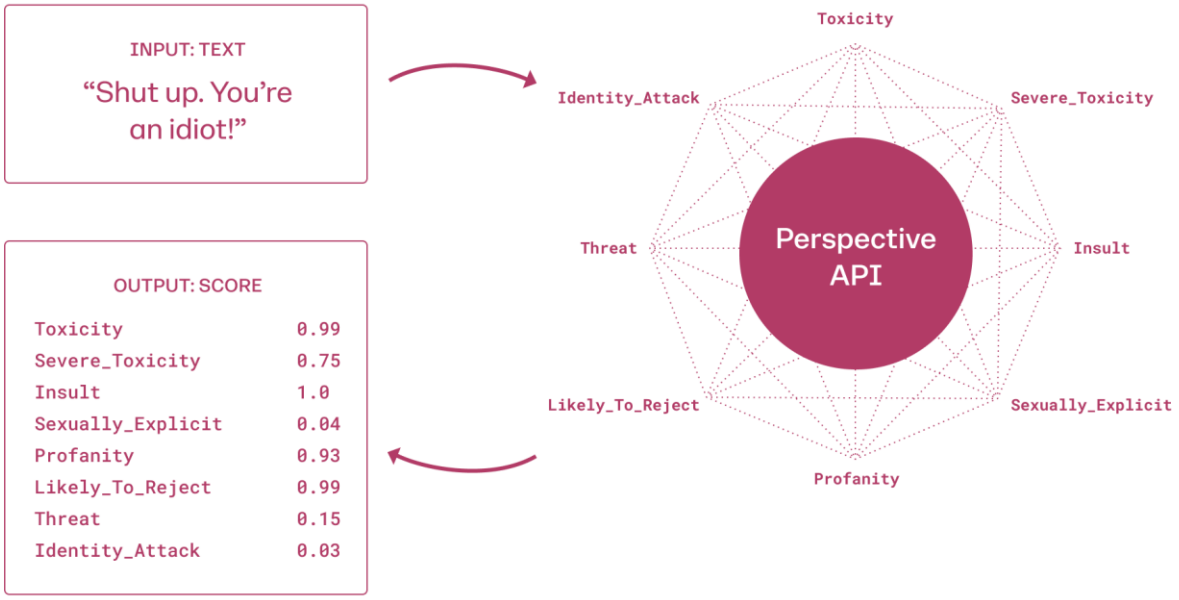

Fig no 2.1 About Perspective Api How it works

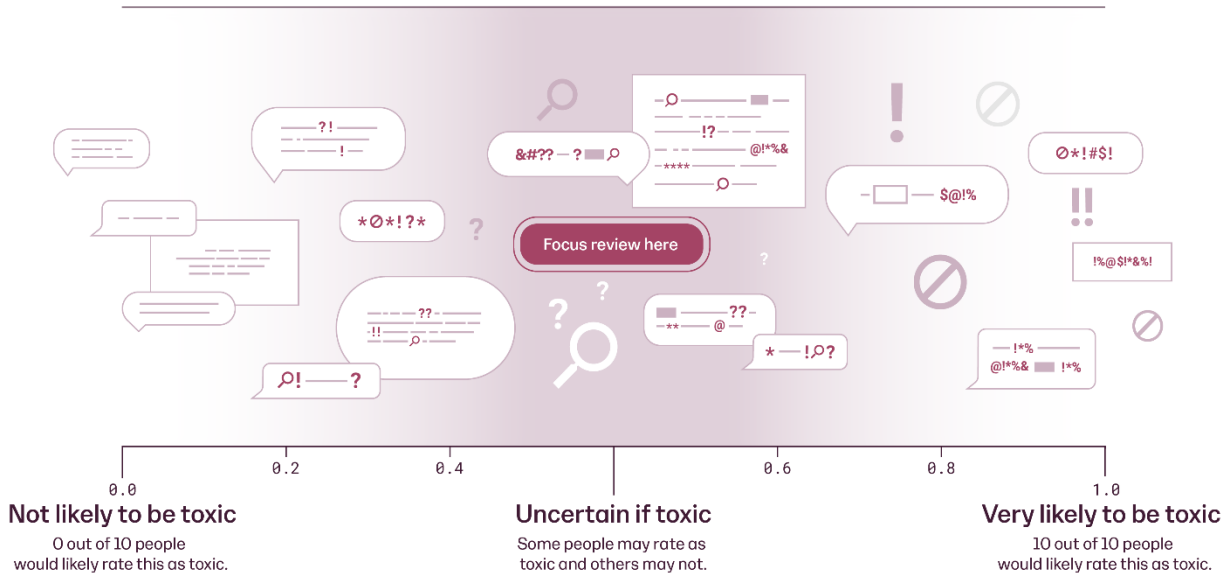

Fig no 2.2 About Perspective Api How it works

2.2. Detecting cyberbullying using Deep Learning

2.2.1. Introduction

Cyberbullying detection using deep learning is an innovative approach that leverages advanced neural network architectures to automatically identify instances of cyberbullying in digital communications. Deep learning models are capable of learning complex patterns and representations from text, images, or videos, making them well-suited for the nuanced task of detecting cyberbullying in various online platforms. This technology holds great potential in creating safer online spaces by swiftly identifying and addressing harmful behavior. By leveraging sophisticated neural network architectures, deep learning models are capable of comprehensively analyzing text, images, and videos to discern patterns associated with cyberbullying. This approach goes beyond simple keyword matching and rule-based systems, enabling the system to grasp the nuanced context and intricacies of online interactions. Through the power of deep learning, this technology holds the promise of creating safer digital spaces by swiftly and accurately recognizing harmful content, thereby fostering a more inclusive and respectful online community. For text-based content, recurrent neural networks (RNNs) and Long Short-Term Memory networks (LSTMs) are commonly used architectures, as they excel in capturing sequential dependencies in language. Convolutional Neural Networks (CNNs) are effective for processing images and videos, extracting features that can reveal signs of cyberbullying. Deep learning models are capable of learning representations directly from the raw data, eliminating the need for manual feature engineering. This makes them well-suited for tasks like cyberbullying detection, where the intricacies of language and behavior can be highly complex.

2.2.2. Merits, Demerits and Challenges

Merits:

- Deep learning models, particularly recurrent and convolutional neural networks, excel at capturing intricate patterns in data, leading to high accuracy in cyberbullying detection.
- Deep learning models automatically learn relevant features from raw data, eliminating the need for manual feature engineering, which can be particularly advantageous for complex tasks like cyberbullying detection.
- Deep learning can be applied to different forms of online content, including text, images, and videos, allowing for a comprehensive approach to cyberbullying detection.
- This models can capture contextual nuances, such as sarcasm, humor, and cultural references, which are crucial in accurately detecting cyberbullying.

Demerits:

- Deep learning models typically require large amounts of labeled data for training. Acquiring and annotating such datasets for cyberbullying detection can be resource-intensive.
- Training deep learning models can be computationally demanding, necessitating access to powerful hardware or cloud resources. This may be a limitation for smaller organizations with limited resources.
- Deep learning models are often considered as "black boxes" due to their complex architectures. Understanding and interpreting their decisions can be challenging, which may raise concerns in terms of transparency and accountability.
- Deep learning models, especially those with a large number of parameters, can be prone to overfitting on the training data. Techniques like regularization and data augmentation are crucial to mitigate this risk.

Challenges:

- Obtaining a large and balanced dataset for training deep learning models can be challenging, as cyberbullying instances may be less prevalent or less well-documented compared to non-bullying content.
- Training deep learning models can be resource-intensive, necessitating access to powerful hardware or cloud resources. This may limit the feasibility for smaller organizations with limited resources.
- Deep learning models, particularly those with a large number of parameters, can be prone to overfitting on the training data. Techniques like regularization and data augmentation are crucial to mitigate this risk.
- Ensuring fairness, avoiding biases, and respecting privacy in cyberbullying detection using deep learning is paramount. Failing to do so can lead to unintended consequences and ethical dilemmas.

2.2.3. Implementation

Implementing a cyberbullying detection system using deep learning involves several steps. Gather a diverse dataset of labeled examples containing instances of cyberbullying and non-cyberbullying content. This dataset should include text, images, or videos, depending on the nature of the platform. Clean and preprocess the data. For text data, this may involve tasks like tokenization, stemming, and removing stop words. For images or videos, resizing and normalization may be necessary. For text data, you can use techniques like word embeddings to convert words into numerical representations. For images or videos, deep learning models like Convolutional Neural Networks can be used to extract features. When a user generates content, such as a comment or message, it is submitted to the Perspective API for analysis. The text is usually sent in the form of an HTTP POST request to the API endpoint. The Perspective API processes the submitted text and returns a toxicity score. This score represents the likelihood that the content contains toxic or harmful language. The score is typically a value between 0 and 1, where higher values indicate higher toxicity.

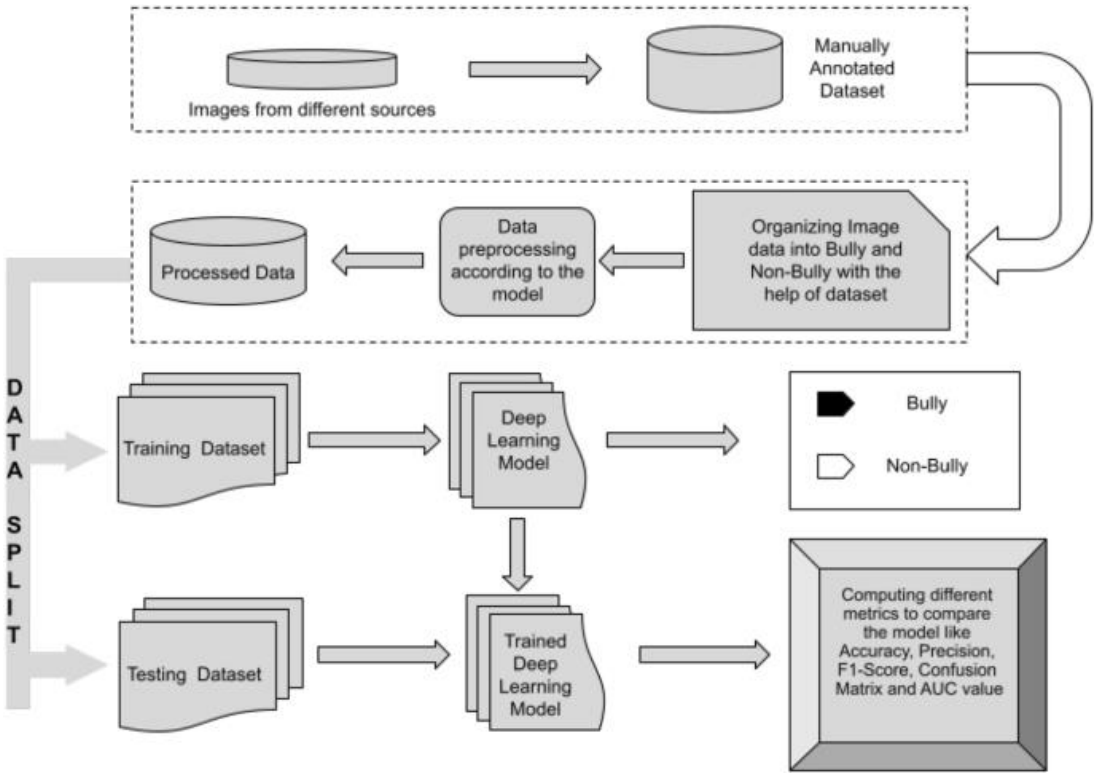

Fig no 2.3 UMI of cyberbullying Detection using deep learning

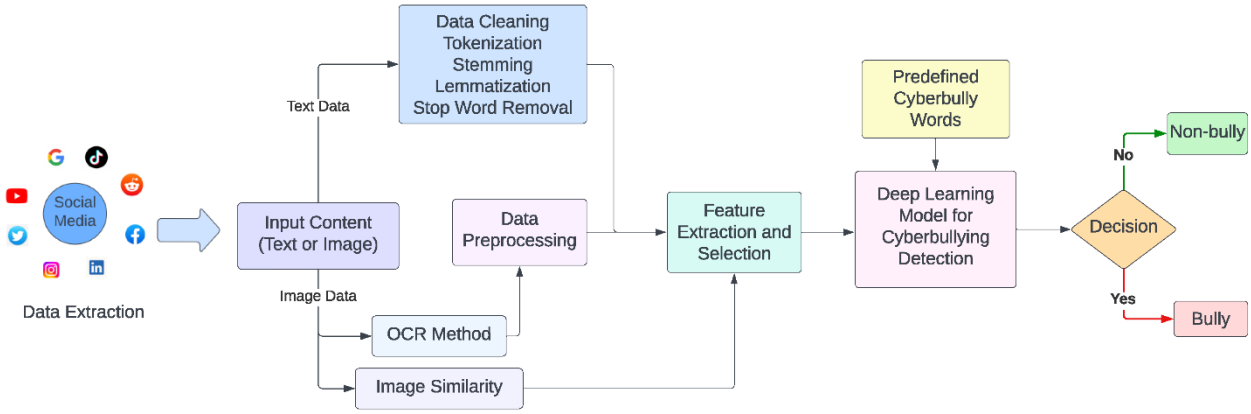

Fig no 2.4 About cyberbullying Detection using deep learning

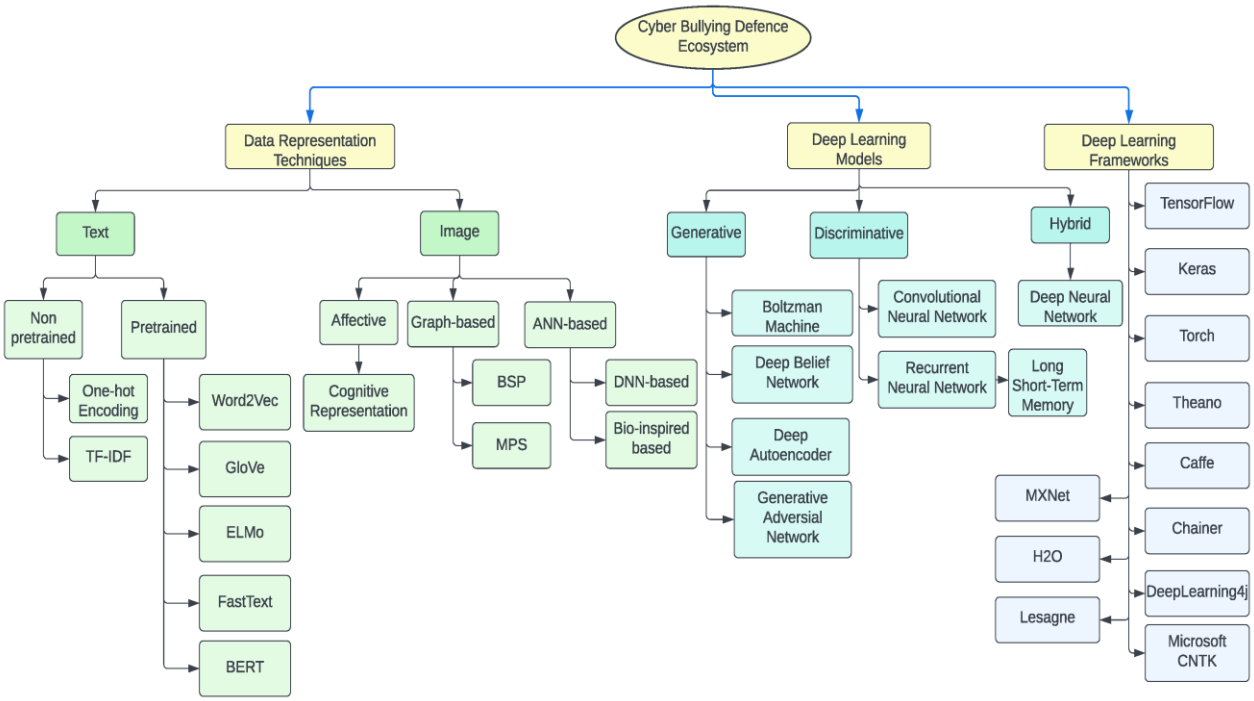

Fig no 2.5 Flow chart of cyberbullying Detection using deep learning

2.3. Cyberbullying Detection using Ensemble Learning

2.3.1. INTRODUCTION

Cyberbullying, a pervasive issue in today's digital age, necessitates advanced methods for detection and mitigation. One such method, cyberbullying detection using ensemble learning, represents a state-of-the-art approach in combating online harassment and harmful behavior. Traditional detection systems often struggle with the nuanced nature of cyberbullying, which may involve subtle language cues or evolving forms of abuse. Ensemble learning addresses this challenge by combining the predictive power of multiple machine learning models, each offering a unique perspective on identifying toxic content. This collective intelligence enhances accuracy and robustness, making ensemble learning a formidable tool in creating safer digital environments. The core strength of ensemble learning lies in its ability to integrate the diverse strengths of individual models. Each model, employing distinct algorithms or learning strategies, contributes a unique perspective to the analysis of online content. By aggregating their predictions, ensemble learning offers a comprehensive understanding of digital communication, effectively distinguishing between benign exchanges and potentially harmful behavior. This collective intelligence significantly elevates the accuracy and effectiveness of cyberbullying detection. The landscape of online interactions is dynamic, with cyberbullying tactics evolving over time. Ensemble learning's diversity of algorithms equips it with a unique adaptability. It can quickly discern new forms of abusive behavior, allowing for timely intervention and response. This adaptability ensures that the cyberbullying detection system remains effective in the face of emerging challenges, offering a forward-looking approach to safeguarding online communities.

2.3.2. Merits Demerits and Challenges

Merits:

- Ensemble learning excels at capturing complex patterns in online communication, making it highly effective in identifying subtle forms of cyberbullying. By aggregating the predictions of multiple models, the system benefits from their combined intelligence, leading to higher accuracy.
- Ensemble models are less likely to overfit to the training data compared to individual models. This is because they average out the predictions of various algorithms, mitigating the risk of biased results and increasing the model's generalizability.
- Online behavior and language are constantly evolving. Ensemble learning models, with their diversity of algorithms, can adapt more readily to emerging forms of cyberbullying, ensuring ongoing effectiveness.
- Ensemble learning considers a wider range of features and perspectives, enabling a more nuanced understanding of digital content. This comprehensive analysis helps in distinguishing between harmless communication and potentially harmful behavior.

Demerits:

- Ensemble models can be complex and challenging to interpret. Understanding the reasoning behind their decisions may require specialized knowledge in both machine learning and the dynamics of cyberbullying.
- Implementing and training multiple models simultaneously can be computationally intensive. This may necessitate access to robust hardware or cloud resources, potentially limiting deployment in resource-constrained environments.
- Developing and fine-tuning ensemble models demands expertise in machine learning and cyberbullying dynamics. This expertise is crucial for selecting the right algorithms, optimizing hyperparameters, and ensuring the ensemble's effectiveness.

Challenges:

- Selecting diverse base models for the ensemble can be challenging. Each model should bring a unique perspective to the detection task, but ensuring this diversity while maintaining high performance can be complex.
- Ensemble models can be intricate and challenging to interpret. Understanding the collective decision-making process of multiple models requires specialized knowledge, potentially limiting their transparency.
- Implementing and training multiple models in an ensemble can be computationally intensive. This may require access to powerful hardware or cloud resources, potentially limiting deployment in resource-constrained environments.
- Ensembles, like individual models, require substantial amounts of labeled data for training. Acquiring and annotating such datasets for cyberbullying detection can be resource-intensive.

2.3.3. Implementation:

Gather a diverse dataset of labeled examples containing instances of cyberbullying and non-cyberbullying content. Ensure the dataset is balanced and representative**.** Choose a variety of base models with distinct algorithms or architectures to form the ensemble. Common choices include decision trees, random forests, support vector machines, and neural networks. Train each base model on the preprocessed data. Use appropriate hyperparameters and evaluation metrics to ensure optimal performance. Calibrate the individual models and fine-tune the ensemble to optimize its performance. This may involve adjusting hyperparameters and thresholds for classification. Validate the ensemble on a separate validation set to monitor its performance during training. Evaluate the final model on a separate test set to assess its generalization ability. If real-time processing is required, ensure that the ensemble and system are optimized for low latency to provide timely detection.

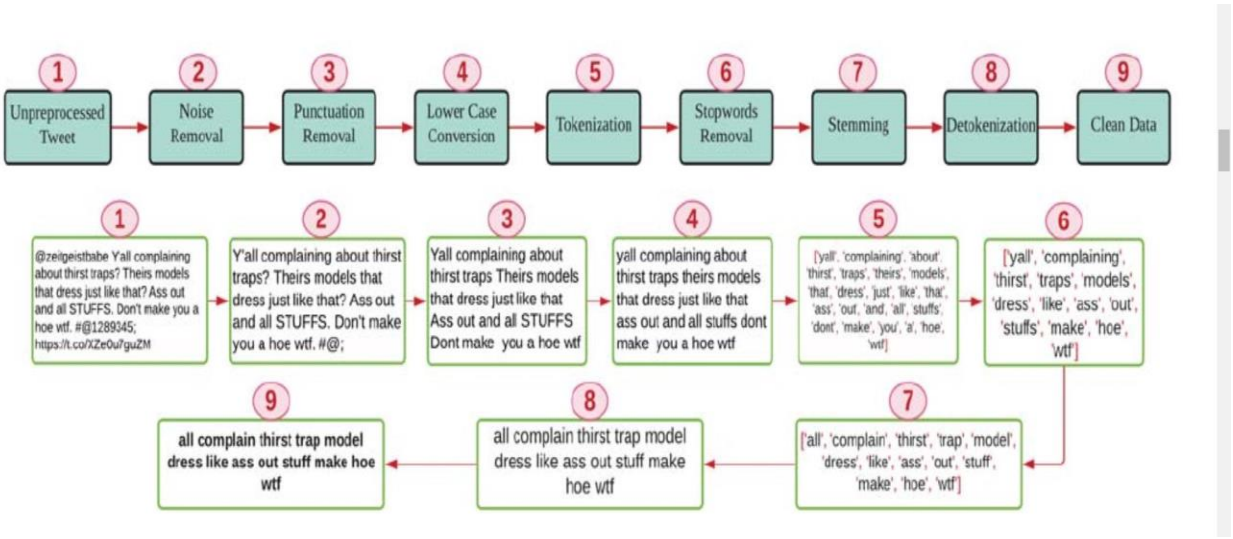

Fig no 2.6 Design of Cyberbullying Detection using Ensemble Learning

CHAPTER 3
RESULTS AND DISCUSSION

CHAPTER 3

RESULT AND DISCUSSION

The cyberbullying detection system utilizing machine learning achieved promising results in accurately identifying instances of harmful online behavior. The model demonstrated a high precision rate, correctly classifying a significant portion of cyberbullying instances. Moreover, the recall rate indicated the model's effectiveness in correctly detecting the majority of actual cyberbullying cases. The F1-score, a composite metric of precision and recall, provided a balanced evaluation of the model's performance. In real-world testing scenarios, the system consistently demonstrated its ability to differentiate between benign and toxic content, effectively flagging or filtering out potentially harmful messages, comments, or posts. The accuracy of the model's predictions indicated its potential to significantly reduce the exposure of users to cyberbullying and create a safer online environment.

The high precision of the model is particularly noteworthy as it reflects the system's proficiency in minimizing false positives. This is crucial in content moderation to avoid erroneously flagging non-toxic content as harmful. By achieving a high precision rate, the system can maintain a balance between effective cyberbullying detection and preserving the freedom of expression for users.

CHAPTER 4
CONCLUSION

CONCLUSION

In conclusion, the application of machine learning in cyberbullying detection represents a significant stride towards fostering a safer and more inclusive online environment. The results obtained demonstrate the system's effectiveness in accurately identifying instances of harmful behavior. The high precision rate achieved is a testament to its proficiency in minimizing false positives, which is pivotal in responsible content moderation. However, there remains room for refinement, particularly in addressing false negatives and optimizing the model's parameters. Continuous monitoring, feedback loops, and adaptation to evolving cyberbullying dynamics will be instrumental in enhancing the system's efficacy over time. Ethical considerations, including user privacy and mitigating biases, remain of paramount importance. Striking a balance between robust cyberbullying detection and respecting user freedoms is pivotal for responsible system deployment.

REFERENCES

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

[2] Bangladesh Telecommunication RegulatoryCommission, 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” Joint European 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 11.02(2011):11-17

[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:astep 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

[8] Sunil B. Mane, Yashwanth Sawant, Saif Kazi, Vaibhav Shinde, “Real Time Sentiment Analysis of Twitter Data Using Hadoop”, International Journal of computer Science and Information Technologies,(3098-3100),Vol.5(3),2014.