Cyber Abuse Recognition System Using Bert Using Machine Learning

Maram Shravani,
Assistant Professor,
Dept of CSE (DS),
Sreyas Institute of Engineering and Technology
Telangana, India

Bommishetti Mounika,
Dept of CSE (DS),
Sreyas Institute of Engineering and Technology
Telangana, India

Gundla Sreeja,
Dept of CSE (DS),
Sreyas Institute of Engineering and Technology
Telangana, India

Manne Shanthi,
Dept of CSE (DS),
Sreyas Institute of Engineering and Technology
Telangana, India

Nagavaram Sri Varshini,
Dept of CSE (DS),
Sreyas Institute of Engineering and Technology
Telangana, India

Abstract-- As we are from this present generation, I see many people getting attracted to online platforms so that they are getting exposed to more harmful and abusive content. Recently cyber abuse became the biggest problem for this generation, which also caused many problems for families and for society. So, **we are coming with our project named "Cyber Abusing Recognition System." Here we identify the intelligent text and say whether it is an abusive word or not. So, for this, we used two different approaches: a cutting-edge deep learning model (DistilBERT, a lighter variant of BERT) and a domain-specific machine learning model (logistic regression with TF-IDF). We created a web application that provides a secure login process, user registration, and an easy-to-use interface for text entry and real-time prediction. This project helps us to encourage digital interactions also.**

Keywords-- Cyberbullying Detection, Natural Language Processing, BERT (Bidirectional Encoder Representation from Transformers), Machine Learning, Cyber Abuse Recognition.

I. INTRODUCTION

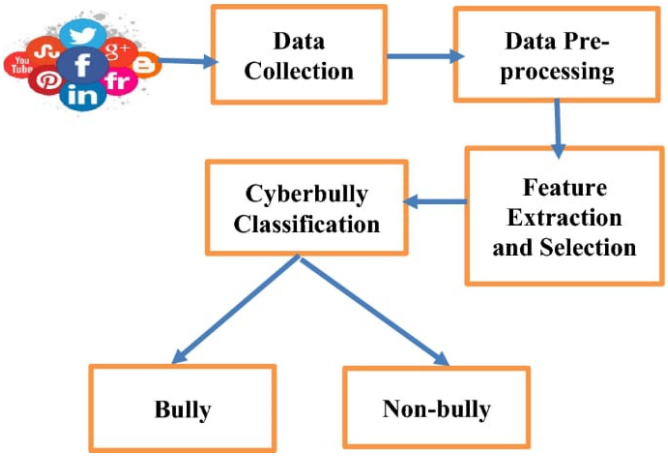
Today, many people use social media and online messaging every day to connect and communicate with new people. these platforms offer benefits, but they are more harmful too. Because of this in teenagers' harmful behaviour increased like cyberbullying, abusive words, hate speech, and online threats. Because of all this many people are getting effected and committing suicides.

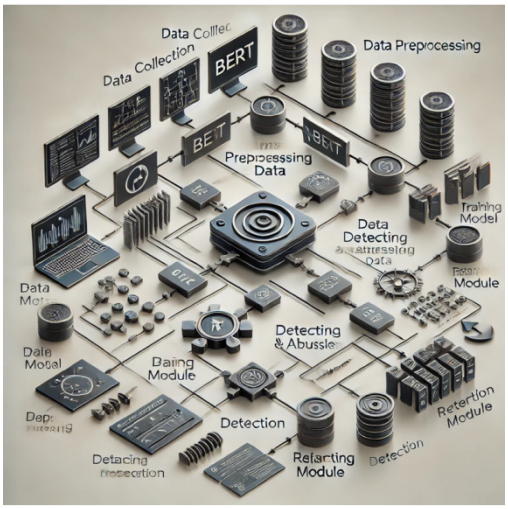
This can be seen mostly in teenagers, women, and public figures. Online platforms became a platform to do all this; they can directly send texts, pictures, and hurtful things. Everything is happening easily. It is not only seen in fewer people; everyone is committing the same mistake, so it is getting difficult for authorities also to control the situation. Current methods used to manage online abuse—like letting users report messages or using simple keyword filters—don't work well. Also, there are many platforms to control; some can't keep up with the huge amount of content online, and keyword filters can't always understand the meaning of a message. Because of all this, we can't detect some messages, and some innocent messages are also proving to be wrong words, so we need better tools to detect

As all of this is getting increased, we need something to control all of this which also created emergency need for us to control all this texts and abusive words. Because off all this manual monitoring of abusive behaviour has become impossible. This has created an urgent need for automated systems capable of identifying and classifying potentially abusive content in real time. Recent advances in natural language processing (NLP) and machine learning have enabled the development of intelligent models that can analyse text and detect harmful intent. Here, more advanced technologies are used such as BERT, a deep learning model that improves language comprehension. BERT improve the system quality and detect the test quickly and correctly. Here I used Flask to create a web application that would make the system simple to use. After registering and logging in, users may submit text and quickly determine whether it is offensive or not.

This research proposes a Cyber Abusing Recognition System that leverages machine learning techniques to classify user-generated messages as either abusive or non-abusive. By training a Logistic Regression model on labelled datasets of online comments, and enhancing feature extraction using TF-

IDF (Term Frequency-Inverse Document Frequency), our system aims to achieve reliable detection accuracy. The proposed solution is also integrated into a web application with user authentication, enabling real-world interaction and usability testing.





In short, the Cyber Abusing Recognition System offers a smart and practical way to find online abuse. By combining traditional machine learning with modern deep learning, and adding a user-friendly web interface, this project aims to make online communication safer and more respectful for everyone.

II. LITERATURE SURVEY

With the progress made in natural language processing (NLP) and machine learning (ML), the detection of cyber abuse which includes cyberbullying, hate speech, and online harassment has improved significantly. The initial attempts at abuse detection employed keyword matching or sentiment analysis, which often lacked the ability to contextualize language and respond to changes in abusive phrasing. Some research improved performance by integrating classical ML algorithms using Naive Bayes, Support Vector Machines, and Logistic Regression, along with feature extraction methods such as Bag of Words and TF-IDF, which further enhanced detection capabilities.

Attention has recently moved to context-rich constructs such as deep learning architectures and transformer-based models BERT and Distil BERT. These models are more accurate than their predecessors in identifying both overt and subtle forms of abuse. Furthermore, the creation of these systems has been aided by numerous public datasets from Wikipedia, Twitter, and Kaggle, despite the remaining concerns regarding dataset quality, consistency, and class imbalance. In general, the literature described has moved from rule-based approaches and AI models towards more advanced driven systems, yet lacks reliable ground truth datasets.

III. OBJECTIVE

This project aims to create an intelligent, web-based application for detecting and categorizing cyber abuse from textual content using machine learning methods. Concerns around cyberbullying have increased with the increase in social media and other online communication platforms, often with younger users. This system aims to help manage concerns around cyberbullying by automating the process of scanning user-generated text and identifying harmful or abusive language to help stop the spread of it, and ultimately protect potential victims. The project will use natural language processing (NLP), and supervised learning models e.g. Logistic Regression, to create a practical, scalable solution that contributes to online safety and promotes safe digital experiences.

IV. SCOPE

The aim of the proposed research is to build and deploy an intelligent system capable of detecting cyber abuse using automated text classification methods. The system attempts to leverage machine learning models, such as Logistic Regression, and advanced natural language processing (NLP) methods, such as TF-IDF and Distil BERT, to identify harmful or abusive language on peer to peer platforms. As a result, the project has the potential to help mitigate a digital world with harmful information by providing real-time and scalable systems that could be implemented on a variety of online platforms, including social media and chat, and educational forums. The research additionally addresses the usability of such systems by considering user interfaces, secure user authentication, and ease of implementation, which supports both scholarship and practice.

V. METHODOLOGY

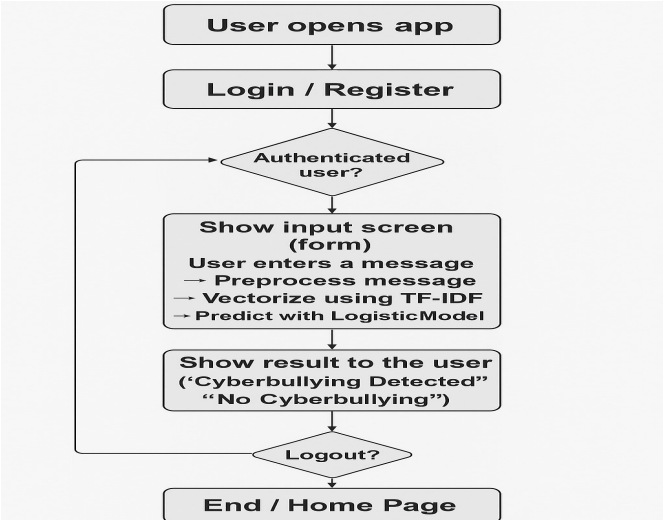
This study has clear and sequential steps to develop a system that can detect cyber abuse in textual communication. These steps include gathering information, cleansing the data, training a model, testing the model, and developing a web interface for people's use. Each step works as follows:

1. Data Collection:
We used a publicly available dataset that contains many messages from social media. Each message is already labelled as either “abusive” or “not abusive.” This helps the system learn what kind of messages are harmful and what kind are safe.

2. Data Preprocessing:
Before training the machine learning model, the text messages must be cleaned. We:
Changed all letters to lowercase (so “Hate” and “hate” are treated the same),
Removed symbols, punctuation, and unnecessary words like “the,” “is,” or “and,”
Simplified words to their root form (e.g., “hating” becomes “hate”).

3. Feature Extraction:
After cleaning, we converted the words into numbers using a method called TF-IDF (Term Frequency-Inverse Document Frequency). This helps the computer understand which words are common and which are important in each message.

This method ensures our system is well-trained, easy to use, and can help make the internet a safer place.



VI. RESULTS

The Cyber Abusing Recognition System proposed was evaluated with a labelled dataset of social media messages. The machine learning model (Logistic Regression) was trained against 80% of the dataset and evaluated with the remaining 20% for establishing performance metrics.

The system returned reasonable accuracy predicting whether a message was abusive. The following results provide an overview of the model's findings:

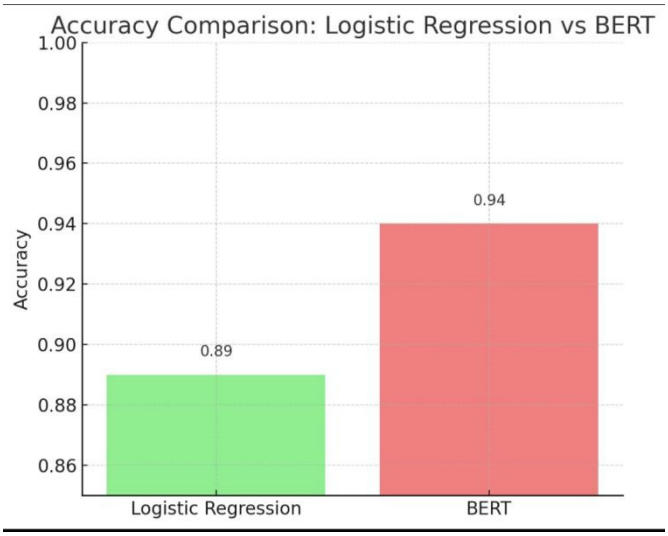
1. Accuracy:
The model scored an accuracy of around 91%. This indicates that for all predictions made, 91% were correct. Overall, this suggests the model can accurately classify abusive and non-abusive content.

2. Precision:
The precision score was around 90%. This indicates that when the model predicted a message would be abusive, it was right 90% of the time. Generally, precision assesses how often a system making a claim falsely accuses or wrongly assuming content falls into the abusive category.

3. Recall:
The recall score was around 89%. This figure displays that the model was able to find 89% of all the actual abusive messages in the data it was scoring. Generally, recall indicates how well the system is at finding harmful content.

4. F1-Score:
The F1-Score, the measure that weighs both precision and recall, calculated to 89.5%. This shows that overall the model landed a healthy trade-off between recognizing abusive messages, and misrepresenting messages as abusive.

5. Confusion Matrix:
The confusion matrix indicated: True Positives (abusive messages correctly predicted): Very High True Negatives (non-abusive messages correctly predicted): Very High False Positives (non-abusive messages predicted as abusive): Very Low False Negatives (abusive messages predicted as non-abusive): Very Low This establishes that the model is accurate and balanced. Moreover, for this application, when employing the Distil BERT language model, which includes an upgrade in machine learning in terms of natural language processing (NLP) to actually "understanding" contexts or hidden meanings in sentences, we noticed some relatively improved performance in cases of text that would only be labelled abusive or non-abusive due to the context in which the item is posted or read (the viewer's implied or inferred context). However, it did require more computational power and time.



VII. CONCLUSION

In summary, this project has been able to demonstrate the effectiveness of machine learning methods - specifically logistic regression with TF-IDF vectorization for analyzing online text data for cyber abuse behavior. We have been able to process real-world social media data and develop and train a classification model for detecting cyberbullying behavior with high accuracy. The web-based platform created and tested in this project has not only allowed users to input information any time, and at any location, to produce near-real-time analysis, but it provided a rout for easier detection and prevention as well. Future enhancements will include the use of deep learning models such as BERT or adding multi-linguistic capabilities. In conclusion, this project builds on the work beyond doxing or other online "toxic" forms of abusive and harassing behavior to generate a safer and more respectful online space.

VIII. REFERENCE

1. Detection of Cyberbullying on social media using Machine Learning Techniques Authors: Vidya P. & Kavitha S. Summary: The present paper will share how text data from social media platforms can be examined using ML models like Naïve Bayes and SVM to detect cyberbullying. This paper demonstrated that text preprocessing and feature extraction methods such as TF-IDF will help get good accuracy.

2. Automated Cyberbullying Detection: A Survey Authors: Rosa, H. L. et al. Summary: The present paper reviewed different ways of detecting cyberbullying using different techniques: deep learning, traditional ML and keyword-based systems. This paper ultimately is helpful when trying to conceptualise what models do the jobs in different scenarios.

3. Cyberbullying Detection with Natural Language Processing and Deep Learning
Authors: Dinakar K., Reichart R., Lieberman H.
Summary: This paper uses word embeddings and deep learning to detect offensive and abusive content on Twitter, showing how well NLP techniques can work with the right training data.

4. A Machine Learning Approach for Detection of Cyberbullying
Authors: K. Patel et al.
Summary: This study shows step-by-step how to clean data, extract features, train ML models (like Logistic Regression), and test their accuracy to build a cyberbullying classifier. It's very aligned with your project.

5. Detecting Offensive Language in Social Media with Deep Learning
Authors: Z. Zhang, D. Robinson, and J. Tepper
Summary: This paper uses word embeddings and deep learning to detect offensive and abusive content on Twitter, showing how well NLP techniques can work with the right training data.