IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Cyberbullying Detection in Social Media using NLP

[1]Prof. Chetana Patil, [2]Srushti Akerkar, [3]Mangali Choudhary, [4]Divyanshi Diva, [5]Sreaya Nair
[1]Professor, Department of Computer Engineering, DPCOE, Pune-412207, India
[2,3,4,5]Student, Department of Computer Engineering, DPCOE, Pune-412207, India

Abstract: The use of the Internet and social media has grown rapidly over the time, becoming an important part of our everyday lives. These platforms help people to express their thoughts, feelings, and ideas and stay connected with each other's. However, as social networking sites have become more popular, the problem of cyberbullying has also increased. Cyberbullying means using technology like social media or messaging platforms to harm or harass or bully someone. The Internet can sometimes spread negative and abusive content, which can cause emotional and mental harm to people. Using offensive or abusive language has become a common issue on social media. Such behavior often causes deep emotional pain, trauma, depression and can also lead to suicide attempts, specially in youngsters. The purpose of this project is to design and develop an automated system to detect and prevent cyberbullying on social networking sites using Natural Language Processing (NLP) and Machine Learning algorithms.

KeyWords : Cyberbullying Detection, Natural Language Processing (NLP), Machine Learning, Text Classification, Sentiment Analysis

1. INTRODUCTION

Social media platforms have become an essential part of daily communication, allowing people to connect and share their ideas easily. However, with the growing use of these platforms, many individuals misuse them for harmful activities such as Cyberbullying. Cyberbullying involves using online platforms to threaten, insult, or emotionally hurt others, often through messages, posts, or comments. Studies show that around 25% of parents reported their children have experienced cyberbullying, which can lead to severe emotional and psychological effects. To address this issue, a cyberbullying detection model using Natural Language Processing (NLP) and Machine Learning (ML) is proposed. In this model, datasets containing online posts and messages are collected and preprocessed using NLP techniques. Then, different ML algorithms such as Logistic Regression, Naive Bayes, SVM, and Random Forest are applied to identify bullying-related content. This approach helps in early detection and prevention, promoting a safer and healthier online environment for users.

2. SYSTEM DESIGN:

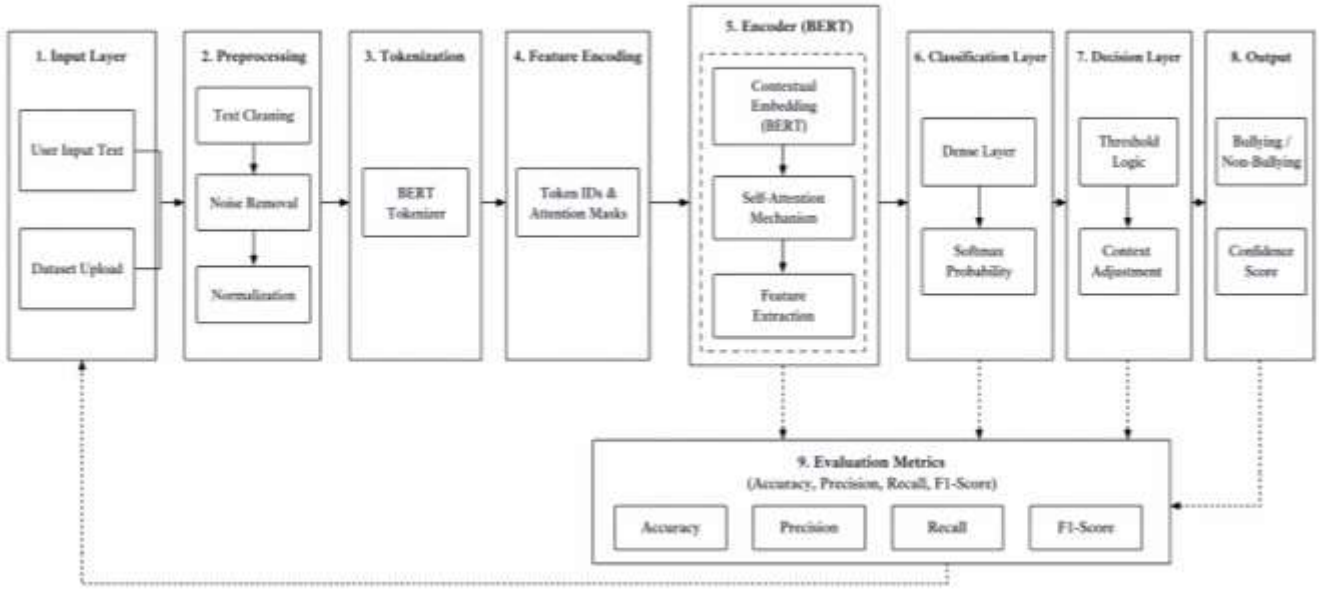

Figure 1. System Architecture

The proposed system for cyberbullying detection in social media is designed using a modular architecture consisting of a user interface, processing layer, and machine learning module.

The front-end of the system is developed using HTML, CSS, Java, and React, providing an interactive and user-friendly interface for inputting and visualizing social media content. The back-end is implemented using a scalable server-side framework that handles data processing, API requests, and communication between components. The system collects textual data from social media platforms such as Instagram and X (Twitter), which is then preprocessed through steps like tokenization, stop-word removal, and normalization.

3. SYSTEM ARCHITECTURE:

The proposed system for cyberbullying detection begins with an input layer that accepts user text or uploaded datasets. The data undergoes preprocessing steps such as text cleaning, noise removal, and normalization to ensure quality input. The processed text is then tokenized using a BERT tokenizer and converted into token IDs with attention masks for feature encoding.

A BERT-based encoder is used to generate contextual embeddings through self-attention mechanisms, enabling effective feature extraction. These features are passed to a classification layer, where a dense layer and softmax function determine the probability of the text being bullying or non-bullying. A decision layer applies threshold logic and context adjustments to refine predictions.

4. WORKFLOW:

The workflow of the system begins with the collection of textual data from social media platforms such as Instagram and X (Twitter). The collected data then undergoes preprocessing, which includes cleaning the text, removing stop words, tokenization, and normalization to prepare it for analysis. After preprocessing, relevant features are extracted and passed to an NLP-based classification model. The model analyzes the text and classifies it as cyberbullying or non-cyberbullying.

The user can either upload a dataset (e.g., CSV file) or manually enter text through the system interface. The input data is then processed by the backend, and then the refined text is passed to the NLP model for classification of cyberbullying content.

Once classification is completed, the results are sent to the backend server, which processes and forwards them to the front-end interface. This workflow ensures efficient processing, accurate detection, and real-time feedback for users. Finally, the results are displayed to the user through the front-end interface.

5.METHODOLOGY:

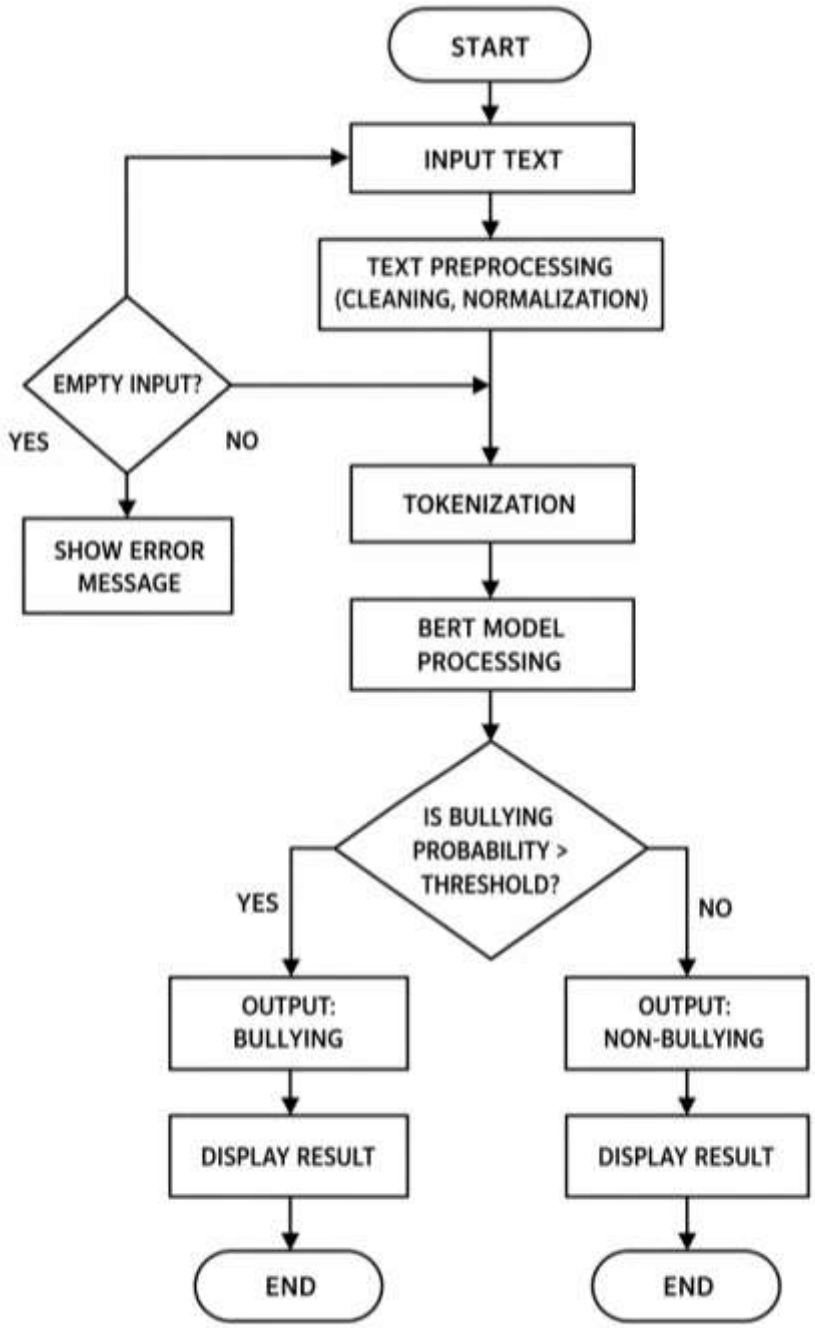

Figure 2. Workflow Diagram

This section includes a set of steps for making an automated cyberbullying detection system. It explains the process for identifying cyberbullying on social media through Natural Language Processing (NLP). It covers several stages, from collecting data to deploying the final system for effective detection of harmful content.

5.1. Data Collection

The first step is collecting data from social media platforms like Twitter from Kaggle. The dataset contains text, posts, comments, and messages labeled as either bullying or non-bullying. This data helps the system learn how online bullying looks in real-life communications. Using publicly

available datasets ensures that the model is trained on real-world examples while maintaining privacy and ethical norms.

5.2. Text Preprocessing

Once the data is collected, it needs to be cleaned and prepared for analysis. It is the process of cleaning, and organizing raw data into a structured format for analysis. Text preprocessing includes way similar as removing unwanted characters, stop words, punctuation, and special symbols. Tokenization and lemmatization are applied to break sentences into meaningful words and reduce them to their base forms. This process helps the machine understand the true meaning of each text.

5.3. Model Selection

After the text is preprocessed, different Machine Learning algorithms are tested to find which one performs best. Algorithms like Naive Bayes, Support Vector Machine (SVM), Logistic Regression, and Random Forest are considered. These models are chosen because they're extensively used in text classification tasks and perform well in detecting patterns in words and phrases related to bullying.

5.4. Model Training and Testing

In this stage, the prepared dataset is divided into two parts: 80 for training and 20 for testing. The training data helps the model learn how to identify the bullying patterns, while the testing data checks how well the model performs on unseen data. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to measure how effective the model is. Among the tested models, Logistic Retrogression frequently gives better accuracy and balanced results.

5.5. System Deployment

After the model performs well, it's deployed as a web-based or social media monitoring system. The system takes new user inputs such as comments or posts and predicts whether the text contains bullying. This helps in early detection and prevention of harmful online behavior. The final system can help parents, educators, and platform moderators in maintaining a safer digital environment.

This section gets into the "how" of our Intelligent placement portal. We'll talk about the specific AI techniques we used, the kind of data that makes it all work, and the models we built to turn that data into useful predictions.

6. RESULTS AND EVALUATION:

6.1. Evaluation Metrics

- **Accuracy:** Measures overall correctness of predictions.
- **Precision:** Indicates how many predicted bullying instances are actually correct.
- **Recall:** Measures the ability to detect actual bullying cases.
- **F1-Score:** Harmonic mean of precision and recall, showing balanced performance.

6.2. Performance Analysis

- Achieved high F1-score, indicating a good balance between precision and recall.
- High recall ensures most bullying content is detected (important for safety).
- Precision is also strong, reducing false alarms.
- Confusion matrix analysis shows low false negatives (fewer missed bullying cases).

6.3. Comparative Evaluation

- Outperforms traditional machine learning models due to context-aware embeddings.
- Better handles complex sentence structures and language variations.
- Provides more reliable predictions in real-world social media data.

7. FUTURE SCOPE:

The problem of cyberbullying is growing as more people use social media every day. Although the current model can detect bullying effectively, there is still room for improvement in the future.

- **Real-Time Detection Systems:** One major area of development is to create a real-time detection system that can instantly identify harmful or abusive content as soon as it is posted. This will help in taking quick action to prevent further harm.

- **Multilingual and cross-Cultural Detection:** The system can also be trained on multilingual datasets, so it can detect cyberbullying in different languages like Hindi, Telugu, and regional Indian languages, not just English.

- **Integration with Social Media Platforms:** Cyberbullying detection tools will be directly integrated into platforms like Instagram, Facebook, and X (Twitter) for automated moderation and user safety.

8. CONCLUSION:

Cyberbullying has become a serious issue in today's digital world. This study presents a model for detecting cyberbullying using Natural Language Processing (NLP) techniques. Several machine learning algorithms were tested to identify which one performs best for text classification. The experiments were carried out using a global Twitter dataset. Among all the models, Logistic Regression (LR) achieved the highest accuracy, showing that it performs better than the other classifiers. It was also observed that Logistic Regression gives more accurate results and faster predictions as the amount of data increases.

9. REFERENCES:

1. Aditya Desai, Shashank Kalaskar, Omkar Kumbhar, Rashmi Dhumal. **"Cyberbullying Detection on social media using Machine Learning"**, ITM Web of Conferences 40, 03038, 2021.

2. Fawzya Ramadan Sayed, Eman Hassan Elnashar, Fatma A Omara. "**Cyberbullying Detection in social media Using Natural Language Processing"**, Scientific African, e02713, 2025.

3. Stephen Afrifa, Vijayakumar Varadarajan. **"Cyberbullying Detection on Twitter using Natural Language Processing and Machine Learning Techniques"**, International Jornal of Innovative Technology and Interdisciplinary Sciences 5 (4), 1069-1080, 2022.

4. Andrea Perera, Pumudu Fernando. Accurate **"Cyberbullying Detection and Prevention on Social Media"**, Procedia Computer Science 181, 605-611, 2021.