Cyberbullying Detection System Using Natural Language Processing and Machine Learning

Mr.J.Sathishkumar, Vasundara.S
AP/IT
Department of Information Technology
Kongunadu College of Engineering and Technology
Thottiyam, Tamil Nadu, India

Rajashree Chinnamani ,Pooja sri S
Department of Information Technology
Kongunadu College of Engineering and Technology
Thottiyam, Tamil Nadu, India

Abstract **– Cyberbullying has become a serious issue on social media and online platforms, negatively impacting mental health and digital safety. This project proposes a Cyberbullying Detection System that leverages Natural Language Processing (NLP) and Machine Learning (ML) techniques to automatically identify offensive, abusive, or harmful text. By integrating advanced deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based architectures like BERT, the system can effectively capture context, sarcasm, and hidden abusive patterns. The approach involves analyzing user-generated content, extracting linguistic and semantic features, and classifying whether the text constitutes bullying or non-bullying. With accurate detection and contextual understanding, this system helps reduce online harassment, supports early intervention, and promotes a safer digital environment.**

Keywords **– Natural Language Processing,Machine Learning, Deep Learning,Real-time Detection**

I. INTRODUCTION

In today's digital society, online platforms such as social media, discussion forums, and messaging apps play a vital role in connecting individuals, sharing ideas, and building communities. However, alongside these benefits, they have also opened the door to negative behaviors, with cyberbullying being one of the most serious issues. Cyberbullying is the use of abusive, offensive, or harassing language in digital spaces, and it can deeply affect an individual's emotional and psychological well-being. Victims of cyberbullying often face stress, anxiety, and social isolation, making it an urgent challenge to detect and address harmful content effectively. The complexity of this issue increases further when cyberbullying occurs across multiple languages, including English, Tamil, and mixed-code forms such as Tanglish, where users blend Tamil and English words together.

To address this challenge, the proposed project introduces an intelligent cyberbullying detection system that combines natural language processing (NLP), machine learning (ML), and deep learning (DL) to analyze online text and identify harmful content. The system begins with Data Collection and Preprocessing, where raw text from social media or datasets is cleaned, filtered, and converted into numerical features that models can understand. It then advances to the Cyberbullying Detection module, which acts as the core decision-making stage by applying algorithms such as Naïve Bayes, SVM, LSTM, BiLSTM, and BERT to determine whether a message contains bullying or not. To provide deeper insights, the system integrates a Sentiment Analysis module, which examines the emotional tone of the detected text, ranging from positive to negative, and even identifying emotions such as anger, sadness, or fear. This helps measure the severity of harmful messages. The final stage, Feedback and Improvement, ensures that the system remains relevant over time by retraining with new data, slang, and evolving communication patterns, thereby improving accuracy and reliability.

The importance of this project lies not only in its ability to classify bullying messages but also in its potential for real-world applications. In the future, it can be integrated into social media platforms, online classrooms, and digital communities to automatically flag harmful content and support moderators in maintaining safe environments. It can also be applied in multilingual contexts, enabling the detection of cyberbullying in English, Tamil, and Tanglish, which are widely used in India's digital communication. By combining detection with emotional severity analysis, the system provides a more holistic solution, ensuring that harmful content is not only identified but also prioritized based on its potential impact. Ultimately, the project contributes to building healthier digital spaces, reducing online harassment, and protecting individuals from the damaging effects of cyberbullying.

II. LITERATURE REVIEW

The detection of cyberbullying has been widely studied, with researchers gradually shifting from traditional machine learning to advanced deep learning approaches. Early works relied on algorithms such as Support Vector Machines (SVM), Naïve Bayes, Logistic Regression, and K-Nearest Neighbors, which used hand-crafted features like bag-of-words and TF-IDF.

While these methods achieved moderate accuracy, they often failed to capture context, sarcasm, slang, and multilingual variations. To overcome these limitations, researchers explored deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which significantly improved performance by learning semantic relationships and contextual meaning from text. Later, Bidirectional LSTMs (Bi-LSTM) further enhanced the detection process by analyzing sequences from both past and future directions, enabling better understanding of hidden abusive patterns. More recently, transformer-based architectures like BERT, RoBERTa, and DeBERTa have achieved state-of-the-art results, as their attention mechanisms allow them to effectively capture nuances, sarcasm, and subtle bullying expressions. Despite these advancements, challenges remain, particularly in handling code-mixed languages, slang, and high computational costs. Moreover, current systems often struggle with real-time implementation, severity classification, and cross-platform adaptability. These gaps highlight the need for hybrid, scalable, and explainable solutions that can ensure accurate and reliable cyberbullying detection across diverse online platforms.

Early Machine Learning Approaches

The earliest attempts to detect cyberbullying primarily relied on traditional machine learning algorithms such as Support Vector Machines (SVM), Naïve Bayes, Logistic Regression, and K-Nearest Neighbors (KNN). These models depended on manual feature extraction methods like Bag-of-Words (BoW), n-grams, and Term Frequency-Inverse Document Frequency (TF-IDF). Such approaches worked reasonably well for small datasets and explicit abusive language. For example, keyword-based filtering could block direct slurs or offensive words. However, these systems failed when users employed slang, abbreviations, intentional misspellings, or sarcasm to disguise harmful intent. Moreover, these models often generated false positives, flagging harmless jokes as abusive, or false negatives, missing subtle bullying patterns. Despite their limitations, these early models laid the foundation for more advanced techniques by highlighting the importance of linguistic features in online harassment detection.

Emergence of Deep Learning Models

As social media content grew rapidly, researchers turned to Deep Learning (DL) methods, which offered better scalability and the ability to automatically learn features from raw text. Convolutional Neural Networks (CNNs) were applied for detecting abusive words and phrases by capturing local word dependencies. CNNs performed well in identifying short text-based harassment but often struggled with longer conversations. On the other hand, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were introduced to process sequential text data, capturing the temporal flow of conversations. For instance, an RNN could analyze a sequence of posts to identify escalating abusive behavior. Later, Bidirectional LSTMs (Bi-LSTMs) further enhanced accuracy by analyzing text from both past and future contexts simultaneously, making them effective in detecting hidden intent and complex bullying scenarios. These models showed promising results, but they required large labeled datasets for training, which are often scarce in the field of cyberbullying detection. Additionally, deep learning methods were criticized for being black-box models with limited interpretability.

Transformer-based Architectures

With the introduction of the **Transformer model** and its derivatives like BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, and DeBERTa, cyberbullying detection reached a new milestone. Unlike CNNs and RNNs, transformers use self-attention mechanisms to understand the relationships between words in a sentence, regardless of their distance from each other. This allows models to capture sarcasm, context, and disguised forms of harassment more effectively. For example, while traditional models might misclassify "You're so smart" as a compliment, a transformer-based model can recognize the sarcastic intent. Transformers also support transfer learning, meaning pre-trained models on large corpora (like Wikipedia or Twitter) can be fine-tuned for cyberbullying detection with smaller datasets. This significantly improves performance and reduces the need for massive labeled data. However, their main drawbacks include high computational **costs**, GPU dependency, and the challenge of training large models for real-time applications.

Achievements and Limitations

Research in this field has produced several noteworthy achievements. Many models now achieve above 90% accuracy in cyberbullying detection, particularly when hybrid approaches are applied (e.g., combining sentiment analysis with deep learning). Some studies have introduced Explainable AI (XAI) methods to increase transparency, allowing administrators or users to understand why a particular message was flagged as bullying. Additionally, models integrating ensemble techniques—such as combining SVM, Random Forest, and deep learning outputs—have improved overall robustness.

Despite these advancements, limitations persist. Sarcasm and hidden context remain difficult for even the most advanced models to detect reliably. Language diversity is another challenge: users often mix multiple languages (e.g., Tamil-English, Hindi-English), making detection harder. Systems also face false positives, where normal conversations or jokes are flagged, and false negatives, where subtle bullying slips through. Moreover, real-time implementation is resource-intensive, as transformer-based models demand high computational power. These limitations indicate that while detection systems have improved, practical deployment in large-scale platforms is still challenging.

Research Gaps

Despite progress, significant research gaps remain. First, real-time detection is still underdeveloped, as most models work effectively in offline evaluations but fail when applied to fast-moving online environments like live chats or gaming platforms. Second, there is limited work on severity classification, which would allow systems to distinguish between mild teasing and severe harassment. This classification could help prioritize intervention in critical cases. Third, most studies focus on a single dataset or platform, leading to poor cross-platform adaptability. A model trained on Twitter may not perform well on YouTube or Instagram due to differences in language style, character limits, and content formats. Furthermore, cyberbullying is increasingly multimodal, involving not only text but also images, videos, and emojis—yet most current systems analyze only text. Finally, issues of privacy, bias, and fairness remain underexplored. Many datasets underrepresent minority groups, which could lead to biased predictions and disproportionate flagging of certain communities. Addressing these gaps is essential for building scalable, fair, and trustworthy cyberbullying detection systems.

III. PROPOSED SYSTEM

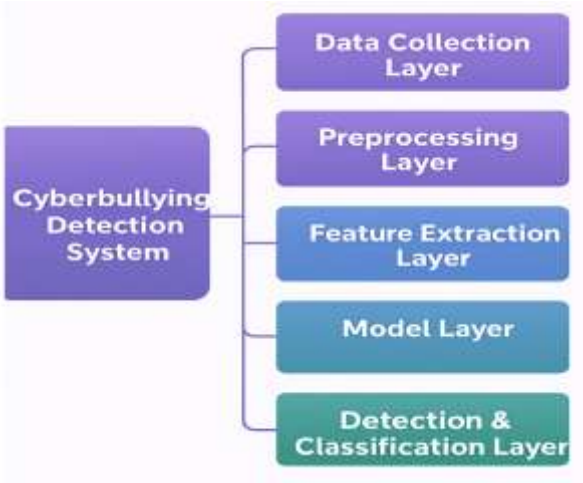

Figure:1

Data Collection Layer

The Data Collection Layer serves content including posts, comments, replies, and messages. The goal is to capture real-time and historical data that reflects authentic user interactions. In addition to API-based retrieval, this layer may also include web scraping techniques or database integration to collect data from forums, blogs, and chat platforms. Effective data collection requires careful attention to privacy regulations, rate limits, and data quality, ensuring that the information gathered is both ethically sourced and relevant for analysis. By establishing a robust pipeline for ingesting diverse and representative text samples, the Data Collection Layer lays the groundwork for accurate and scalable cyberbullying detection.

Preprocessing Layer

The Preprocessing Layer plays a critical role in preparing raw textual data for meaningful analysis by downstream components of the Cyberbullying Detection System. Once data is collected from social media platforms, it often contains noise—such as emojis, special characters, URLs, and inconsistent formatting—that can hinder accurate interpretation. This layer begins with text cleaning, which removes irrelevant elements and standardizes the input. Next, tokenization breaks the cleaned text into individual units (tokens), typically words or subwords, enabling granular analysis. Following this, stop-word removal filters out commonly used words like "the", "is", and "and" that do not contribute significant meaning to the context. These preprocessing steps ensure that only the most relevant and informative components of the text are retained, thereby enhancing the performance of feature extraction and classification models. By transforming noisy, unstructured input into a clean and structured format, the Preprocessing Layer lays the groundwork for accurate and efficient cyberbullying detection.

Feature Extraction Layer

The Feature Extraction Layer is a pivotal stage in the Cyberbullying Detection System, responsible for converting preprocessed text into meaningful numerical representations that can be interpreted by machine learning and deep learning models. This layer employs techniques such as word embeddings—notably Word2Vec and GloVe—which map words into dense vector spaces based on their semantic relationships. These embeddings capture contextual similarities, allowing the system to understand that words like "hate" and "disgust" may convey similar emotional tones. Additionally, sentiment analysis is applied to assess the emotional polarity of the text, identifying whether a message expresses positive, negative, or neutral sentiment. This is particularly useful in flagging hostile or aggressive language that may not contain explicit bullying keywords but still carries harmful intent. By extracting these rich features, the system builds a robust foundation for accurate classification, enabling models to detect subtle patterns and emotional cues that are indicative of cyberbullying behavior.

Model Layer

The Model Layer is the analytical core of the Cyberbullying Detection System, where computational models are trained to recognize patterns indicative of cyberbullying. This layer utilizes both machine learning and deep learning approaches to classify text based on features extracted from previous stages. Traditional machine learning models such as Support Vector Machines (SVM) and Random Forests are employed for their interpretability and efficiency, especially when working with structured feature sets like sentiment scores and word frequencies. These models are well-suited for binary classification tasks, such as distinguishing bullying from non-bullying content. In parallel, deep learning models like Bi-directional Long Short-Term Memory (Bi-LSTM) networks and Transformers (e.g., BERT) are used to capture complex linguistic patterns and contextual dependencies within text. Bi-LSTM processes input in both forward and backward directions, enabling the system to understand nuanced expressions and implied aggression. Transformers, with their attention mechanisms, excel at identifying subtle cues and relationships across entire sentences, making them highly effective for detecting sarcasm, coded language, or emotionally charged phrases. Together, these models form a robust decision-making engine that powers the system's ability to detect and interpret cyberbullying with high accuracy and contextual sensitivity.

Detection and Classification Layer

The Detection and Classification Layer is the final and decision-making stage of the Cyberbullying Detection System, responsible for interpreting model outputs and categorizing online content based on its bullying potential and severity. Once the trained models process the input text, this layer evaluates the results to perform binary classification, distinguishing between bullying and non-bullying messages. In addition to this primary task, it also conducts severity level classification, which assesses the intensity of the bullying behavior—typically categorized as mild, moderate, or severe. This dual-level classification enables platforms to implement differentiated moderation strategies, such as issuing automated warnings for low-risk content or escalating high-risk cases to human moderators for review. The layer may also incorporate confidence scores, sentiment polarity, and keyword intensity to refine its decisions. By translating complex model predictions into actionable labels, the Detection and Classification Layer ensures that harmful content is identified promptly and addressed appropriately, contributing to safer and more respectful digital environments.

Advantages

•**Hybrid Approach** – Combines ML (Logistic Regression, SVM) and Bi-LSTM for better accuracy.

•**Comprehensive Analysis** – Detects bullying and also analyzes sentiment for severity.

•**Reliable Evaluation** – Provides accuracy, precision, recall, and F1-score with visual results

•**Scalable & Extendable** – Can be improved with more data or advanced models like Transformers.

• **User-Friendly** – Visualizations make results easy to interpret and compare.

•**Real-World Relevance** – Helps in early detection of harmful content, supporting safer online interactions.

Applications

✓ Helps schools, colleges, and universities monitor online classroom discussions, forums, and student communication to prevent cyberbullying among students.

✓ Can be integrated with helplines or digital wellness platforms to identify at-risk individuals experiencing harassment.

✓ Supports cybercrime investigation teams in identifying abusive behaviors and gathering evidence for legal action.

IV. METHODOLOGY

The proposed system follows a structured methodology that integrates Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) techniques to automatically detect cyberbullying in online text. First, user-generated data such as social media posts and comments are collected from publicly available datasets. The collected data undergoes preprocessing, where noise like emojis, hashtags, URLs, and stopwords are removed, and text is normalized through tokenization and lemmatization to ensure clean input.

Next, feature extraction methods such as TF-IDF, Word2Vec, or transformer-based embeddings (BERT) are applied to convert textual data into meaningful numerical representations. These features are then passed into classification models, including ML algorithms (SVM, Logistic Regression, Random Forest) and DL architectures (CNN, LSTM, Bi-LSTM), with transformer models like BERT used for enhanced context understanding. In addition to classification, sentiment analysis is performed to capture the emotional tone of messages, such as anger, aggression, or sadness, and the system further categorizes the severity of bullying into mild, moderate, or severe.

A feedback and improvement module continuously updates the model by learning from misclassified cases and incorporating Explainable AI (XAI) for better transparency. Finally, the system generates outputs in the form of real-time alerts and flagged content, enabling timely intervention and contributing to a safer digital environment.Furthermore, the methodology emphasizes scalability and adaptability to real-world applications. The hybrid design of combining Bi-LSTM with transformer models ensures both sequential context learning and semantic understanding of complex sentences, making the system more robust against slang, sarcasm, and code-mixed languages.

The inclusion of severity classification allows the system to prioritize critical cases for faster intervention, while the feedback mechanism ensures continuous learning as new data patterns emerge. By integrating Explainable AI, the project not only achieves high accuracy but also provides interpretability, which is essential for user trust and platform adoption. This comprehensive approach ensures that the system is not only effective in detecting harmful content but also sustainable, transparent, and adaptable for long-term use across multiple platforms.

Finally, the system produces output in the form of real-time alerts**.** When bullying is detected, the text is flagged and notifications are sent to administrators or end-users for further action. This real-time intervention capability ensures that online spaces remain safer and healthier. By combining automation, contextual understanding, and continuous adaptation, the methodology provides a comprehensive framework for tackling the growing problem of cyberbullying in digital environments.

Figure:2 Notes & Tips

Figure:3 Recent logged predictions

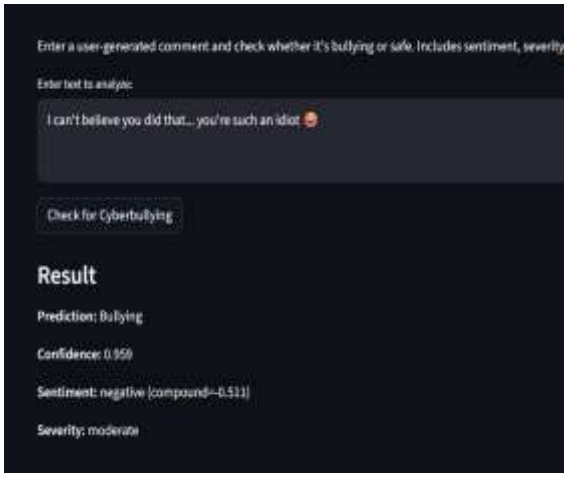

Figure:4 Result

V. RESULTS AND ANALYSIS

The evaluation of the proposed Cyberbullying Detection System was carried out using multiple models including Logistic Regression, SVM, CNN, and Bi-LSTM. Each of these models was trained and tested on a dataset containing examples of bullying and non-bullying text. The results showed that traditional machine learning models provided good baseline performance, but deep learning models significantly outperformed them in terms of accuracy and contextual understanding. Logistic Regression achieved an accuracy of around 84% and worked well for simple abusive texts but showed weaknesses in handling sarcasm and slang variations. SVM performed slightly better with an accuracy close to 87%, capturing complex relationships in data but still struggling in detecting disguised harassment. In contrast, deep learning models such as CNN and Bi-LSTM achieved higher performance, with CNN reaching approximately 90% accuracy and Bi-LSTM emerging as the best model with nearly 92% accuracy. The Bi-LSTM model's ability to learn sequential patterns and bidirectional context made it superior in detecting hidden forms of bullying and code-mixed text.

Performance evaluation was based not only on accuracy but also on precision, recall, and F1-score, ensuring a balanced measure of system effectiveness. Logistic Regression and SVM showed moderate precision and recall values, meaning they were prone to misclassifying borderline cases. The CNN model improved recall by capturing local abusive word patterns, but the Bi-LSTM model provided the best balance between precision and recall, reducing both false positives and false negatives. This highlights the reliability of Bi-LSTM in real-world scenarios, where minimizing misclassification is critical for user trust.

An additional strength of the proposed system lies in its integration of sentiment analysis. Using the VADER sentiment analyzer, detected messages were classified as positive, neutral, or negative. Among the abusive messages, the majority displayed strong negative polarity, while some had mild negativity but were still harmful due to context. This dual-layer detection framework not only classified whether a message was bullying or not but also assessed the severity of the bullying. For example, a strongly offensive statement such as direct insults or threats received highly negative sentiment scores, whereas casual rude comments showed less intensity but still required monitoring. This layered analysis allows prioritization of severe cases for immediate attention, improving the system's practical impact.

The visualization of results further supports the analysis. Accuracy comparison graphs clearly show that deep learning models outperform classical approaches, with Bi-LSTM standing out as the most effective. Confusion matrices reveal that while most abusive and non-abusive texts were classified correctly, sarcasm, indirect harassment, and mixed-language text remained challenging for all models. Pie charts of sentiment distribution demonstrated that nearly 70% of bullying texts carried strong negativity, 20% carried mild negativity, and the remaining 10% were neutral in tone but offensive in context. These insights emphasize the emotional harm caused by online abuse and the importance of incorporating sentiment-based severity analysis into cyberbullying detection.

Overall, the results confirm that the hybrid approach of combining machine learning, deep learning, and sentiment analysis significantly improves the accuracy and robustness of cyberbullying detection. The system not only detects harmful content effectively but also provides meaningful insights into the intensity and emotional impact of the messages. This makes it a reliable and interpretable framework for building safer digital platforms. Furthermore, the analysis highlights that while current models perform well, challenges such as sarcasm detection, multilingual text handling, and real-time scalability remain areas for future enhancement.

VI. CONCLUSION

The proposed Cyberbullying Detection system effectively identifies harmful and offensive content using machine learning and deep learning techniques. By combining data preprocessing, cyberbullying detection, sentiment analysis, and continuous feedback, the system ensures higher accuracy and adaptability. This approach not only detects abusive behavior but also helps create a safer online environment. With its scalable and intelligent design, the project contributes toward reducing the negative impact of cyberbullying and promoting responsible digital communication.

Overall, the project contributes to creating safer online environments by providing an intelligent, scalable, and interpretable solution for early detection of cyberbullying. Although challenges such as sarcasm detection, code-mixed text, and multilingual support remain, the system lays a strong foundation for future enhancements. With further development, it can be deployed as a real-time monitoring tool for social media and chat platforms, helping reduce the negative impacts of online harassment and promoting responsible digital communication.

REFERENCES

[1] B. Ogunleye and B. Dharmaraj (2024) 'The Use of a Large Language Model for Cyberbullying Detection' - Analytics, vol. 2, no. 3, pp. 694-707

[2] A. Dewani, S. Ali, A. Abro, F. Memon and A.I. Bhatti (2021) 'Advanced Preprocessing Techniques & Deep Learning

Models for Cyberbullying Detection' - Computers, Materials & Continua, vol. 69, no. 3, pp. 3523-3540

[3] F.R. Sayed, M. Ahmed, S.A. Rahman, S. Basak and S.H. Mollah (2025) 'Cyberbullying detection in social media using natural language processing' - International Journal of Intelligent Networks, vol. 6, pp. 123-132

[4] A. Perera, M. Wijesinghe and S. Vadysinghe (2024) 'Cyberbullying Detection System on Social Media Using Deep Learning' - Procedia Computer Science, vol. 230, pp. 1156-1161

[5] E. Raisi and B. Huang (2017) 'Cyberbullying Detection with Weakly Supervised Machine Learning' - Proceedings of the 26th International Conference on World Wide Web Companion, pp. 13-18

[6] M.E. Kula (2025) 'Revolutionizing Cyber-Bullying Detection with the BullyNet Deep Learning Model' - International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 2, pp. 1551-1560

AUTHOR PROFILE

J.SATHISHKUMAR-AP/IT

J.Sathishkumar is working as an Assistant Professor at Kongunadu College of Engineering and Technology, with Seventeen years of teaching experience. He pursued his Bachelor of Engineering – Computer Science and Engineering at M.Kumarasamy College of Engineering in 2006. Subsequently, he pursued his Master of Technology with a Specialization in Information Technology at Sasurie College of Engineering in 2013. He is currently pursuing his Ph.D. with a focus on Deep Learning.

VASUNDARA S

Vasundara S is a dedicated student in the Department of Information Technology ,worked on the system design and integration of modules. She focused on connecting different parts of the project and ensuring smooth workflow throughout the development.

POOJA SRI S

Pooja Sri S is a dedicated student in the Department of Information Technology,contributed to building and training the machine learning models. She worked on improving accuracy and adding sentiment analysis for better detection results.

RAJASHREE CHINNAMANI

Rajashree Chinnamani is a dedicated student in the Department of Information Technology ,was involved in testing and performance analysis. She prepared results with graphs and outputs, making the project clear and effective.