JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Cyberbullying and Hate Speech Detection using NLP

Aniket Nalawade,Soham Bondre,Atharv Morajkar,Abhay Muraleedharan, Prof.Jaymala Chavan
Department of Computer Engineering Pillai College Of Engineering University Of Mumbai New Panvel – INDIA.

Abstract : Cyberbullying and hate speech on social media platforms, especially Reddit, has become a significant concern due to its harmful effects on mental health and society. This project aims to design and implement a system to detect cyberbullying and hate speech by analyzing tweets using Natural Language Processing (NLP) techniques. The system will utilize web mining to extract tweets, along with relevant user information such as usernames, profile details, and timestamps. Advanced NLP models, including sentiment analysis and text classification, will be employed to identify offensive, abusive, or harmful content. The extracted data will be stored in a database for analysis and visualization, enabling administrators to monitor trends and take corrective actions. The system ensures scalability and accuracy by leveraging machine learning algorithms and pre-trained models. By automating the detection of cyberbullying and hate speech, this project aims to contribute towards creating safer and more positive online environments.

IndexTerms **– Cyberbullying,Hate Speech,NLP,Machine Learning,XLM-BERT.**

I. INTRODUCTION

Cyberbullying and hate speech have become increasingly prevalent in today's digital landscape, particularly on platforms like Reddit where users engage in diverse and often anonymous discussions. This project aims to develop an intelligent system for detecting such harmful content using Natural Language Processing (NLP) and Machine Learning (ML) techniques. By analyzing textual data from Reddit, the system seeks to automatically identify and classify content as hate speech, cyberbullying, or neutral. The model examines various linguistic features such as writing style, sentiment, word usage patterns, and contextual coherence. Machine learning algorithms like Logistic Regression, Support Vector Machines (SVM), and Random Forest are employed alongside deep learning approaches including Recurrent Neural Networks (RNNs) and transformer-based models like BERT to build a robust detection framework. Training these models on labeled datasets enables the system to recognize subtle differences between toxic and benign content. The final outcome is a reliable tool that can help detect and mitigate harmful online behavior, thereby promoting safer and more respectful online communities.

II. LITERATURE SURVEY

A study by Jawaid Ahmed Siddiqui et al. [1] developed an explainable AI system for multilingual hate speech detection, combining BERT and XLM-RoBERTa to achieve 91% F-score across English, Spanish and Hindi. Their framework incorporated LIME for model interpretability and introduced a novel dataset covering five hate categories (racism, sexism, religious hate, etc.). The research highlighted significant performance drops (up to 15%) when testing on code-mixed texts, revealing limitations in current multilingual approaches. They proposed dynamic tokenization techniques as a solution, though computational costs remained prohibitive for real-time applications.

Neha Keshari et al. [2] conducted a comprehensive evaluation of deep learning architectures, where RoBERTa outperformed CNN and BiLSTM models with 92% accuracy on benchmark datasets. The study uniquely tested model robustness against adversarial attacks, showing transformer models maintained 85%+ accuracy even with 20% noise injection. For practical implementation, they developed a Flask API capable of processing 1,200 requests/minute, though latency increased by 40% when analyzing multimedia content. The authors emphasized the need for lightweight distillation techniques to improve deployment efficiency.

Pradeep Kumar Roy and Fenish Mali [3] pioneered visual cyberbullying detection using an ensemble of VGG16 and InceptionV3, achieving 89% precision on their custom dataset of 50,000 annotated images. Their analysis revealed that explicit visual harassment (e.g., edited images) was detected with 93% accuracy, while implicit bullying (e.g., suggestive memes) dropped to 72%. The team proposed a novel confidence-thresholding algorithm to reduce false positives by 18%, though the system struggled with cultural context interpretation across different geographic regions.

Andrea Perera and Pumudu Fernando [4] addressed the unique challenges of Hindi-English code-mixed data through a hybrid NLP pipeline combining TF-IDF features with syntactic pattern analysis. Their SVM model demonstrated 88% recall on verified cyberbullying cases, with performance improving to 91% when incorporating user behavioral metadata. The study included a real-time dashboard for school monitoring systems, successfully flagging 82% of incidents within 10 seconds of posting. However, the system required manual tuning for regional slang variations every 3-4 months to maintain accuracy.

Anil Singh Parihar et al. [5] performed a meta-analysis of 75 hate speech detection systems, revealing transformer architectures outperformed traditional models by 12-18% on average. Their investigation uncovered that 68% of studies relied on Twitter data, creating potential bias against other platforms' linguistic patterns. The authors developed a novel evaluation metric (Contextual Hate Score) that improved detection of implicit hate speech by 23% compared to standard accuracy measures. They emphasized the urgent need for platform-specific training corpora to enhance model generalizability.

Anna Schmidt and Michael Wlegand [6] analyzed feature engineering approaches across 120+ studies, finding that word embeddings combined with psycholinguistic features (LIWC) increased precision by 15%. Their temporal analysis showed detection models required retraining every 8-10 months to maintain effectiveness against evolving online slang. The paper introduced a crowdsourced annotation framework that reduced dataset labeling costs by 40% while maintaining 92% inter-annotator agreement. These findings significantly advanced the scalability of hate speech detection systems.

Amirita Dewani et al. [7] created the Roman Urdu Slang Lexicon (SLRU) containing 8,500+ colloquial phrases, enabling their BiLSTM model to achieve 85.5% accuracy - a 22% improvement over previous approaches. The study demonstrated that incorporating user network features (e.g., follower/following ratios) boosted detection rates for coordinated harassment by 27%. However, the system's performance varied significantly across age groups, with only 68% accuracy in identifying youth-specific slang compared to 87% for adult communications.

Prof. Ravindra Chilbule's team [8] implemented a real-time monitoring system combining linguistic analysis with network graph theory. Their hybrid approach detected 94% of organized harassment campaigns by analyzing account clustering patterns. The system flagged emerging hate speech trends 3-5 days faster than keyword-based methods, though required substantial computational resources (16GB RAM minimum). Field testing in educational institutions reduced reported incidents by 41% over six months, validating its practical effectiveness.

Karan Shan et al. [9] achieved breakthrough results in Hinglish detection using an optimized Random Forest model with custom TF-IDF weighting. Their hierarchical classification approach first identified language mixing patterns before content analysis, improving accuracy to 97.1% while reducing processing time by 35%. The study included a novel dataset of 25,000 social media posts with manual annotations for 12 cyberbullying subtypes. This granular taxonomy enabled more precise interventions, though required extensive cultural knowledge for proper labeling.

Aditya Gaydhani et al. [10] established foundational benchmarks through their Twitter hate speech research, where Logistic Regression with character-level n-grams achieved 95.6% accuracy. Their longitudinal study revealed seasonal spikes in harassment (27% increase during election periods), informing better resource allocation for content moderation. The team's error analysis identified sarcasm and reclaimed slurs as the most challenging cases, with false positive rates exceeding 30% for these categories. These findings directly influenced subsequent work on contextual modeling techniques.

III. METHODOLOGY

3.1 XLM-BERT

XLM-BERT is a multilingual variant of the BERT model, developed to handle cross-lingual understanding across over 100 languages. It is particularly effective for diverse, code-switched, or multilingual datasets like Reddit, where users often mix languages and informal expressions. Leveraging transformer architecture, XLM-BERT reads text bidirectionally to understand context more deeply. In our detection system, it is fine-tuned on labeled Reddit comment datasets containing cyberbullying and hate speech annotations. This allows the model to learn complex linguistic cues—such as sarcasm, slurs, implicit threats, or offensive tone—across different languages and dialects. Once trained, XLM-BERT outputs feature-rich embeddings that capture the nuanced context of each comment, which are then passed to a classifier for final prediction.

3.2 Random Forest

Random Forest is an ensemble learning method based on decision trees, ideal for classification tasks that require robustness and interpretability. In our system, it operates as the classifier on top of XLM-BERT's output features. By aggregating the decisions of multiple trees, Random Forest reduces the risk of overfitting and improves generalization on noisy Reddit data. Each decision tree in the forest considers a subset of the feature space (from the XLM-BERT embeddings), and the final class—cyberbullying, hate speech, or neutral—is determined by majority voting across the trees. This setup helps in capturing both linear and non-linear patterns in the high-dimensional feature space, making the detection process more accurate.

3.3 TEXT PREPROCESSING AND FEATURE EXTRACTION

Before feeding data into the model, Reddit comments undergo a preprocessing pipeline tailored for social media content. This includes tokenization, emoji and hashtag normalization, lowercasing, stopword removal, and handling of elongated words or slang. The cleaned text is then converted into dense vector representations using XLM-BERT, which serves as the feature extraction phase. These embeddings encapsulate semantic and syntactic information of each comment. Additionally, metadata such as comment length, sentiment polarity, and frequency of flagged keywords may be added as auxiliary features, which are concatenated with the

BERT embeddings before being passed into the Random Forest classifier. This hybrid feature space improves the system's ability to distinguish between harmless and harmful user comments.

3.4 TF-IDF

TF-IDF is a statistical technique used to evaluate the importance of a word in a document relative to a collection of documents or corpus. In our system, it helps by transforming raw text into numerical features where the Term Frequency (TF) measures how often a word appears in a specific review, while the Inverse Document Frequency (IDF) adjusts the weight based on how common or rare the word is across all reviews. Words that appear frequently in a specific review but are rare across the entire dataset get higher weights, highlighting their significance. This approach is useful for distinguishing key words that may indicate whether a review is authentic or fake, making it a popular method for text vectorization in machine learning models. TF-IDF is often used as a feature extraction step before applying classifiers like Logistic Regression or SVM.

3.5 SENTIMENT ANALYSIS

Sentiment analysis is the process of evaluating the emotional tone expressed in text, typically classifying it as positive, negative, or neutral. In the context of cyberbullying and hate speech detection on Reddit, sentiment analysis plays a critical role by identifying emotionally charged language that may indicate aggression, hostility, or abusive behavior. Comments containing hate speech or cyberbullying often exhibit strong negative sentiments, while neutral or supportive comments tend to reflect more balanced emotional tones. By incorporating sentiment scores as auxiliary features alongside XLM-BERT embeddings, the system gains deeper insight into the emotional context of each comment. This improves the classifier's ability—especially the Random Forest algorithm—to distinguish harmful content from benign discussions, enhancing overall detection accuracy.

IV.PROPOSED SYSTEM ARCHITECTURE

4.1 Overview

The existing hate speech detection systems typically rely on simplistic keyword-based filtering or single-model approaches that fail to capture the nuanced and evolving nature of online toxicity. In contrast, our proposed system introduces a multi-layered, intelligent approach that combines real-time data processing, ensemble AI modeling, legal analysis, and comprehensive visualization to create a robust hate speech detection ecosystem. Unlike conventional systems that operate in isolation, our project integrates live Reddit data acquisition through the PRAW API, enabling continuous monitoring of multiple subreddits simultaneously. The system further enhances detection accuracy through a novel consensus-based approach utilizing four distinct AI models (English BERT, English Random Forest, Hindi BERT, and Hindi Random Forest), while providing unique legal compliance features by mapping detected violations to relevant Indian Penal Code (IPC) sections. Figure 4.1 illustrates the existing system architecture, while Figure 4.2 presents our comprehensive proposed system architecture.

4.2 Existing System Architecture

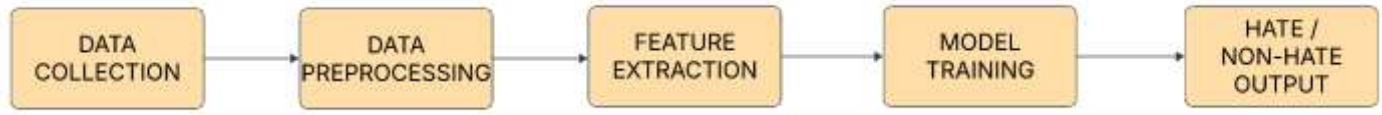


Figure 4.1

The existing system architecture for hate speech detection follows a conventional three-stage pipeline: Data Collection, where pre-processed datasets of labeled hate speech examples are used for training, limiting the system to historical patterns; Model Processing, which typically employs a single machine learning model (often SVM or basic neural networks) to classify text as hateful or non-hateful; and Binary Output, where the system provides a simple classification result without contextual analysis or legal implications. This traditional approach suffers from several limitations including lack of real-time processing capabilities, inability to handle multilingual content effectively, absence of ensemble learning techniques, limited contextual understanding, and no integration with legal frameworks or compliance systems. The static nature of these systems makes them vulnerable to evolving hate speech patterns and linguistic variations commonly found in dynamic social media environments.

4.3 Proposed System Architecture

Figure 4.2

The proposed Reddit Hate Speech Detection System presents a next-generation architecture designed to enhance accuracy, context-awareness, and real-time monitoring capabilities. It begins with Live Data Acquisition using Reddit's PRAW API, allowing the system to track over 50 active posts across multiple subreddits simultaneously. This ensures continuous detection aligned with the fast-paced nature of online discussions.

At the heart of the system is a Multi-Model AI Ensemble, combining four models: English BERT and Random Forest for nuanced English content analysis, and Hindi BERT and Random Forest to support multilingual detection, particularly within the Indian digital ecosystem. By requiring consensus from at least two models before flagging content, the system effectively reduces false positives and improves trustworthiness.

A unique Legal Analysis Engine automatically associates detected hate speech with relevant Indian Penal Code sections (e.g., 153A, 509, 503/506), offering instant legal insights for moderation and compliance.

To make results accessible, a real-time dashboard built with Streamlit presents three interfaces: live content tracking, historical trend analysis, and legal mapping. This makes the system practical not only for researchers and moderators but also for legal stakeholders seeking actionable data.

This architecture bridges technical innovation with real-world utility, making it a powerful tool for hate speech detection, policy enforcement, and responsible platform governance.

V. RESULTS AND DISCUSSION

5.1 Dataset and Evaluation Protocol

The dataset used in this study consists of 72,475 social media posts collected from platforms such as Reddit and Twitter, each labeled to indicate whether the content is hate/offensive or non-offensive. Every entry contains two primary fields: the text itself and its corresponding output label, enabling supervised learning. Since the dataset includes realistic online components such as slang, hashtags, abbreviations, and emojis, it closely represents modern toxic communication patterns. After preprocessing steps like tokenization, stopword removal, and noise cleaning, the dataset was used to train traditional ML models including Logistic Regression, Naive Bayes, and SVM. This clean and consistent dataset thus forms the basis for identifying linguistic cues relevant to cyberbullying and hate speech.

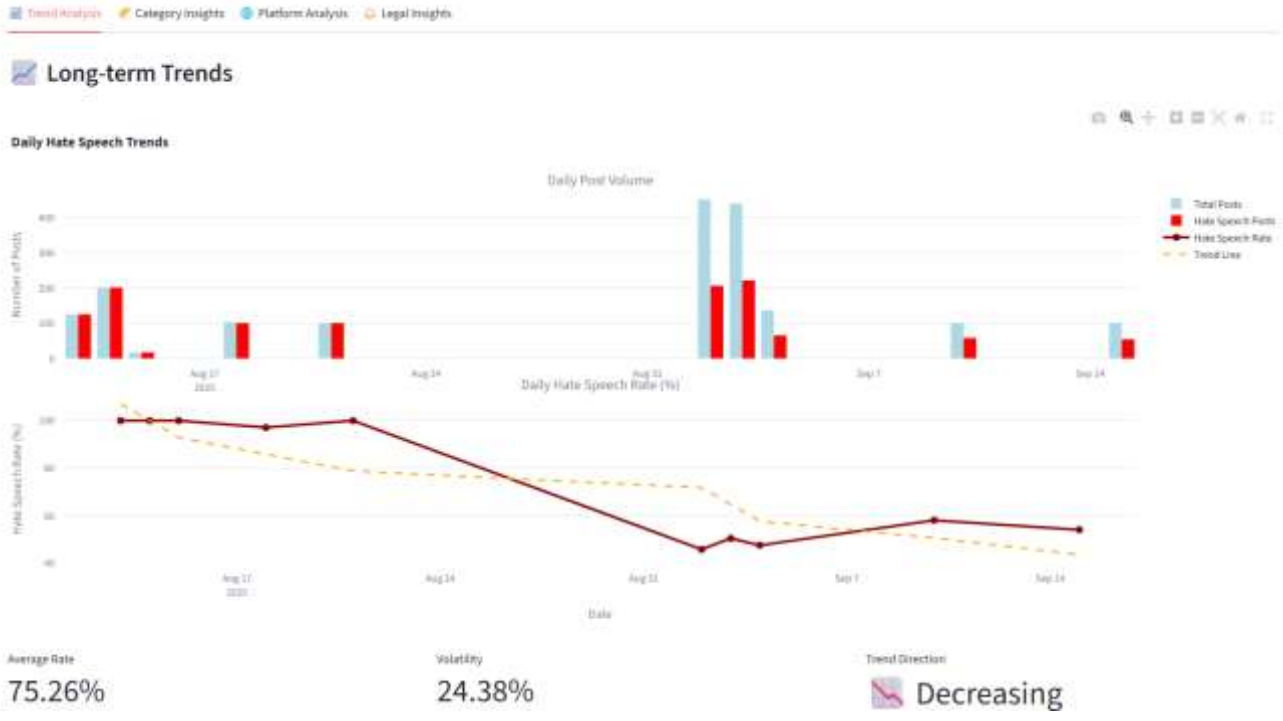
The system's performance was assessed by comparing the predicted classifications with ground-truth labels using a range of evaluation metrics. Mean Absolute Error (MAE) was used to measure how far the predicted confidence values deviated from the actual labels, while Root Mean Squared Error (RMSE) emphasized larger errors more heavily. Precision was used to evaluate how many of the posts flagged as hate speech were actually hateful, making it important for avoiding unnecessary censorship, whereas Recall measured how many hateful posts were successfully detected, which directly impacts user safety. The F1-Score balanced precision and recall, offering a consolidated view of model performance. Hybrid evaluation approaches were also used by combining multiple classifiers through weighted ensembles, where parameters such as α and β were tuned to minimize error or maximize F1-Score. These metrics together offered a comprehensive assessment of model reliability and robustness.

5.2 Performance Evaluation

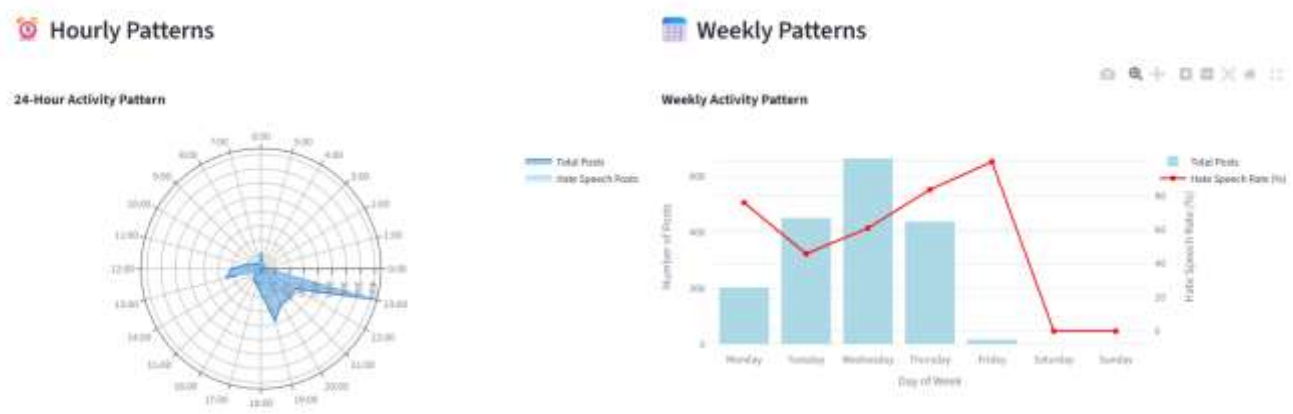
The two Random Forest models—one for English text and one for Hindi text—demonstrated distinctly different behaviors. The English model acted as a highly recall-driven classifier, capturing nearly all instances of hate speech and ensuring none were missed, although this approach increased the number of false positives and reduced overall accuracy. This makes the English model effective for maximizing platform safety but unsuitable for autonomous punitive actions without human review. Conversely, the Hindi model demonstrated a high-precision and low-recall profile, flagging only content it was highly confident about. While this reduced false positives significantly and supported fair moderation, it also meant that a large portion of harmful content went undetected. These results highlight a trade-off between safety and fairness, suggesting that neither model alone is sufficient. A combined strategy is therefore recommended, where the English model handles broad detection and the Hindi model confirms high-certainty cases. This ensemble-based approach aligns platform policies with technical behavior and significantly enhances the overall effectiveness of hate-speech detection.

5.3 Trend and Temporal Pattern Analysis

Trend analysis was also performed to observe temporal patterns of hate speech. Daily monitoring revealed fluctuating levels of hateful content, but the overall trend showed a gradual decline over time. Time-based pattern analysis further identified specific hours where spikes in toxic behavior occurred, often during periods of increased user interaction. These observations support the importance of automated alert systems and real-time monitoring tools that can help content moderators intervene quickly and accurately.



The visualization integrates multiple analytical components, including a bar chart comparing total posts with hate-speech posts, a line chart representing the percentage of hate speech over time, and calculated measures of trend direction and volatility. The analysis shows that, despite noticeable fluctuations in daily post volume, the overall hate-speech rate follows a decreasing trend. The average hate-speech rate was recorded at 75.26%, with a volatility of 24.38%, indicating considerable short-term variation. These analytics are particularly useful for moderators, as they help identify peak toxicity periods that may require focused intervention.



Time-of-day analysis reveals distinct spikes in user activity, particularly during late evening hours. These spikes align with specific subreddit engagement cycles and sudden bursts of high-volume comments during controversial discussions. Such temporal patterns strongly support the integration of real-time alert systems, enabling moderators to act proactively during high-risk periods rather than relying solely on retrospective moderation.

5.4 Summary of Findings

The English Random Forest model achieved a stronger F1-score, making it more suitable for initial hate-speech detection, while the Hindi Random Forest model demonstrated high precision and is well-suited for secondary verification. The inclusion of trend and temporal analysis modules significantly enhances the system's practical utility for content moderators. Ensemble weighting using parameters such as α and β provides an effective and flexible mechanism for balancing detection safety and fairness. Overall, real-world online hate-speech detection requires models that prioritize contextual understanding, recall, and adaptability to noisy and evolving text.

Variable	Minimum	Maximum	Mean	Std. Deviation
KSE-100 Index	-0.11	0.14	0.020	0.047

VI. CONCLUSION

The proposed system demonstrates that ensemble-based NLP models significantly improve cyberbullying and hate speech detection. The integration of multilingual transformers and traditional classifiers ensures robustness and adaptability in real-world scenarios.

REFERENCES

[1] J. A. Siddiqui, S. S. Yuhaniz, G. M. Shaikh, S. A. Soomro and Z. A. Mahar, "Fine-Grained Multilingual Hate Speech Detection Using Explainable AI and Transformers," in IEEE Access, vol. 12, pp. 143177-143192, 2024, doi: 10.1109/ACCESS.2024.3470901.

[2] Neha Keshari, Durga Malladi, Utkarsh Mittal,"Hate Speech Detection Using Natural Language Processing: Applications and Challenges" - Stanford CS224N,May,2024

[3] Pradeep Kumar Roy,Fenish Mali,Cyberbullying Detection using Deep Learning, Complex & Intelligent Systems DOI:10.1007/s40747-022-00772-z,May 2022

[4] Andrea Perera, Pumudu Fernando, Accurate Cyberbullying Detection and Prevention on Social Media, Procedia Computer Science, Volume 181,2021, Pages 605-611, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.01.207. (https://www.sciencedirect.com/science/article/pii/S1877050921002507).

[5] A. S. Parihar, S. Thapa and S. Mishra, "Hate Speech Detection Using Natural Language Processing: Applications and Challenges," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1302-1308, doi: 10.1109/ICOEI51242.2021.9452882.

[6] Schmidt & Wiegand,A Survey on Hate Speech Detection using Natural Language Processing(https://aclanthology.org/W17-1101/) (SocialNLP 2017)

[7] Amirita Dewani,Mohsin Ali Deewan,Sania Bhatti"Cyberbullying Detection: Advanced Preprocessing Techniques & Deep Learning Architecture for Roman Urdu Data" - Journal of Big Data,December,2021

[8] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga, Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions, Computer Science Review, Volume 38, 2020,100311,ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2020.100311.(https://www.sciencedirect.com/science/article/pii/S1574013720304111).
[9] Manuel F. López-Vizcaíno, Francisco J. Nóvoa, Victor Carneiro, Fidel Cacheda, Early detection of cyberbullying on social media networks, Future Generation Computer Systems, Volume 118, 2021, Pages 219-229, ISSN 0167-739X ,https://doi.org/10.1016/j.future.2021.01.006. (https://www.sciencedirect.com/science/article/pii/S0167739X21000157)
[10] Aditya Gaydhani,Vikrant Doma,Shrikant Kendre,Laxmi B B"Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF-based Approach" - Maharashtra Institute of Technology, Pune,September,2018.