International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Enhancing Social Media Safety with Machine Learning-Based Cyberbullying Detection

Benitlin Subha K[1], Swathi C[2], Sandhya D[3] ,Swetha M[4] and Serin E[5]

[1]Assistant Professor -Department of Information Technology & Kings Engineering College-India.

[2,3,4,5]Department of Information Technology & Kings Engineering College-India.

Abstract - Social media platforms have revolutionized global communication but also facilitated the rise of cyberbullying, posing serious threats to user well-being, particularly among youth. Manual moderation is inadequate for managing the scale and velocity of harmful content online. This paper proposes a machine learning-based system for real-time cyberbullying detection, leveraging TF-IDF vectorization and a Logistic Regression classifier to identify and categorize user comments as toxic, obscene, threatening, or hateful. A Flask-powered web interface enables users to evaluate comment toxicity interactively. Designed for scalability and future expansion, the system supports integration with social media APIs, multi-language processing, and potential adoption of deep learning models. This work aims to contribute to safer online environments through intelligent, automated moderation.

Key Words**:** *Cyberbullying Detection, Machine Learning, TF-IDF, Logistic Regression, Toxic Comment Classification, Online Safety, Flask Web Application, Social Media Moderation, Natural Language Processing, Real-Time Content Filtering*

1.INTRODUCTION

Cyberbullying is a form of harassment that occurs through digital communication platforms such as social media, instant messaging, and online gaming environments. Unlike traditional bullying, cyberbullying can happen anytime and anywhere, often making it difficult for the victim to find relief. It includes behaviours such as sending threatening messages, spreading rumours online, posting embarrassing pictures or videos, and impersonating others to humiliate them. The pervasive nature of the internet means that harmful content can spread rapidly, amplifying its impact. Victims of cyberbullying often experience a range of emotional and psychological issues, including anxiety, depression, low self-esteem, and in severe cases, suicidal tendencies. The anonymity provided by digital platforms often emboldens perpetrators and complicates efforts to identify and punish them. Thus, cyberbullying has become a significant societal issue that requires robust and scalable technological solutions to detect and prevent harmful activities online.

1.1 OBJECTIVES

The primary objective of this project is to design a machine learning-based system that can accurately detect instances of cyberbullying in user-generated comments across socialmedia platforms. The system aims to preprocess and analyze textual data effectively, applying TF-IDF vectorization to convert raw text into numerical features that can be understood by machine learning algorithms. The core classification task is handled by a Logistic Regression model, which categorizes comments into various toxicity labels such as toxic, obscene, threat, insult, and identity hate. Additionally, the project aims to develop a user-friendly web application using Flask, allowing users to interact easily with the detection system. Through early detection of harmful content, the system aspires to contribute towards creating a safer and healthier online Environment.

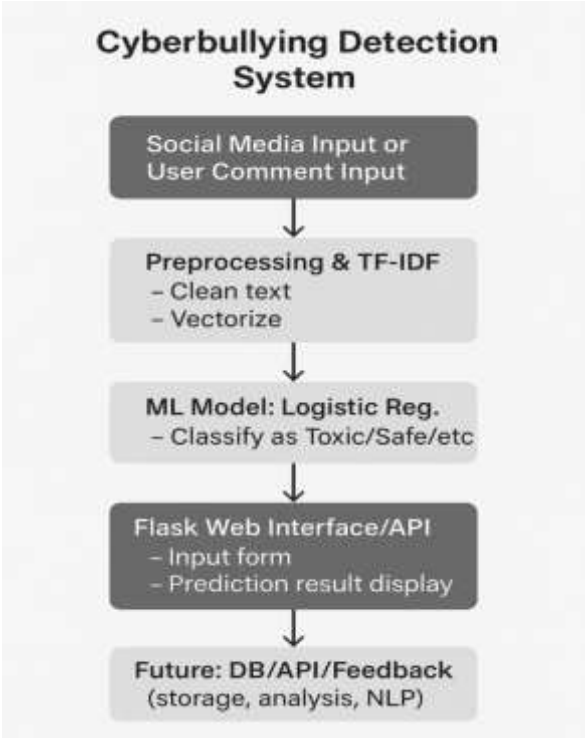


Fig-1:Cyber bullying Detection System Architecture

1.2EXISTING SYSTEM

Existing systems for moderating cyberbullying incidents largely depend on manual intervention or simple keyword matching techniques. In manual moderation, human reviewers are employed to inspect and remove offensivecontent; however, this approach is labour-intensive, inconsistent, and unable to keep pace with the volume of content generated on modern platforms. Keyword-based filters provide a faster alternative but are equally limited, as they cannot comprehend the context of communication. They are prone to high false-positive and false-negative rates, often missing harmful comments written with intentional misspellings or sarcasm.

Moreover, traditional systems lack the adaptability to evolve with dynamic online language trends, making them ineffective against sophisticated cyberbullying strategies. Therefore, there is a strong demand for intelligent, automated systems capable of understanding the context and detecting cyberbullying more accurately.

1.2 PROPOSED SYSTEM

The proposed system introduces a machine learning-based approach to automate the detection of cyberbullying activities within user comments. Rather than relying on a rigid set of predefined keywords, the system uses TF-IDF vectorization to

identify the significance of words and phrases relative to the entire dataset. These features are then processed by a Logistic Regression model trained to classify comments into multiple categories such as toxic, obscene, threat, insult, and identity hate. A lightweight Flask web application serves as the user interface, allowing users to input comments and instantly receive toxicity predictions. The modular architecture of the system is designed to be scalable and extendable, supporting future integration with social media platforms for real-time monitoring and incorporating advanced deep learning models like BERT for improved contextual understanding. The overall goal of the system is to foster a safer digital space by enabling the early detection and mitigation of cyberbullying incidents.

2.BASICS OF SOCIAL MEDIA

2.1 OVERVIEW OF SOCIAL MEDIA PLATFORMS

Social media has revolutionized the way individuals communicate, interact, and engage with information in the digital era. Originating from the basic idea of sharing personal updates and connecting with friends and family, social media platforms have evolved into multifaceted environments that serve personal, professional, educational, and entertainment purposes. Platforms like Facebook, Instagram, Twitter, YouTube, Snapchat, and LinkedIn have enabled users to create content, share opinions, form communities, and participate in global conversations irrespective of geographical boundaries. The growth of mobile technology and widespread internet access have significantly fuelled the adoption of social media, making it an inseparable part of daily life for billions of people around the world. Social media enables the rapid dissemination of information and opinions, empowering users to express themselves, stay informed, and organize collective actions. It also plays a vital role in modern marketing, journalism, education, and even politics, offering opportunities for outreach and engagement on an unprecedented scale. However, alongside its advantages, social media has also introduced significant risks, including privacy violations, misinformation, cyberbullying, and online harassment. The open nature of social media allows both positive and negative interactions to flourish simultaneously. Algorithms designed to maximize user engagement often prioritize controversial or emotionally charged content, inadvertently amplifying negativity and divisiveness. Moreover, the anonymity afforded by some social media platforms sometimes emboldens individuals to behave in ways they might not in face-to-face interactions. Hate speech, trolling, and cyberbullying have become widespread issues, particularly affecting younger demographics who are heavily active online. As social media continues to expand its influence on personal lives, societies, and cultures, it becomes increasingly crucial to develop strategies and technologies that can protect users from harmful experiences without stifling freedom of expression. The importance of understanding social media's dynamics lies not only in leveraging its benefits but also in mitigating its risks. Developing intelligent, automated solutions for monitoring and managing online interactions has thus become a major area of research and innovation, especially in the context of

promoting safer online communities and preserving mental health in the digital world.

2.2 USER INTERACTIONS ON SOCIAL MEDIA

User interactions on social media platforms form the core of online community dynamics and engagement. Interactions include activities such as liking, sharing, commenting, reposting, tagging, and direct messaging. These actions allow users not only to consume content but also to participate actively in shaping conversations, influencing public opinion, and building social relationships in digital spaces. Comments, in particular, play a vital role as they facilitate two-way communication between content creators and their audiences, enabling discussions, feedback exchanges, debates, and community bonding. Each interaction adds to the complex web of user-generated content that continuously evolves the platform's digital culture and relevance. However, the openness of social media interactions also introduces several challenges. While interactions can promote positive engagement, knowledge sharing, and entertainment, they can simultaneously become vehicles for negativity, including trolling, hate speech, misinformation, and cyberbullying. The viral nature of social media means that a single comment or post can quickly gain widespread attention, often amplifying both positive and negative sentiments beyond the original context. Anonymous or pseudonymous interactions further complicate accountability, sometimes leading to irresponsible or harmful behaviours.Moreover, analysing user interactions presents significant technical challenges. Unlike traditional communication, social media language often involves slang, emojis, sarcasm, memes, and cultural references that are context-dependent and rapidly evolving. Distinguishing between genuine criticism, humour, and malicious intent requires advanced natural language processing and machine learning techniques that can interpret subtle nuances and contextual clues. Understanding user interactions is crucial for multiple reasons. For businesses, interactions provide insights into customer behaviour, brand sentiment, and market trends. For individuals and communities, interactions shape online experiences, influence perceptions, and impact emotional well-being. For policymakers and technologists, managing interactions responsibly is essential to safeguarding freedom of speech while preventing harm. In the context of this project, focusing on user interactions, especially comments, provides a valuable avenue for detecting early signs of cyberbullying and toxicity. By developing intelligent systems capable of analysing these interactions in real time, it becomes possible to enhance digital safety, foster healthy online communities, and create platforms that prioritize positive, respectful communication.

3. PRINCIPLES OF SOCIAL MEDIA

3.1 BASIC IDEAS

The principles of social media are rooted in the fundamental human desire for communication, community building, and information sharing. Social media platforms provide spaces where individuals can freely express their opinions, share their experiences, and build networks of relationshipsacross geographical boundaries. At the core of these platforms is the principle of

- **User-generated content:** Empowering every user to contribute, create, and influence the digital environment without the need for traditional gatekeepers such as publishers or broadcasters.
- **Interconnectivity:** Social media networks thrive on the ability to link people together through follows, friends, groups, likes, comments, and shares. This interconnectedness allows information to travel quickly, enabling the rapid spread of news, ideas, trends, and social movements. The viral nature of content sharing demonstrates the power and responsibility that users hold in shaping public discourse.
- **Participation and engagement:** It are also fundamental ideas in social media. Platforms are designed to encourage active interaction rather than passive consumption. Features such as commenting, liking, reacting, and reposting are intended to make users not just consumers of information but contributors to ongoing conversations.
- **Transparency and authenticity:** These have become increasingly important principles as social media matures. Users tend to favour genuine and relatable content over polished marketing messages. The demand for authentic experiences drives trends such as live streaming, behind-the-scenes content, and user testimonials, all of which foster trust and community loyalty.

- **Personalization:** It is a driving force behind the user experience on social media. Algorithms curate content based on user interests, behaviours, and interactions, aiming to deliver a more relevant and engaging experience. However, personalization can also lead to challenges such as information bubbles and echo chambers, where users are exposed only to viewpoints that reinforce their existing beliefs.
- **Freedom of expression:** Is celebrated in social media environments, offering individuals a voice that can reach large audiences.
- S**ocial responsibility:** It as unchecked content can lead to the spread of misinformation, harassment, hate speech, and other harmful activities.
- **Constant evolution:** It is a defining feature of social media. Platforms must continually adapt to changing user needs, technological advancements, cultural shifts, and regulatory frameworks. New features, policies, and community standards are regularly introduced to ensure that platforms remain relevant, engaging, and safe for their users.

Understanding these basic ideas is essential for appreciating the significant impact of social media on modern communication, culture,and society. These principles also inform the need for systems and tools that can moderate content, promote positive interactions, and safeguard users from harmful behaviour, forming the

foundation for projects aimed at enhancing social media safety through machine learning

Fig-2:Features Of Social Media

4.FEATURES OF SOCIAL MEDIA

- **Hashtags**: Hashtags are a powerful tool on social media, enabling users to categorize and discover content related to specific topics. They facilitate participation in trends, enhance content visibility, and connect people across networks, while brands use them for promotions and community building. However, misuse or spamming of hashtags can reduce their effectiveness.
- **Explore**: The Explore feature is crucial for content discovery, recommending posts based on user behavior and trends. It helps users find new content, influencers, and trending topics outside their direct network, encouraging platform engagement. While it fosters exploration, it can also create filter bubbles, limiting exposure to diverse perspectives.
- **Video**: Video content has become one of the most engaging forms of communication on social media. It combines visuals, audio, and motion to capture attention and convey messages quickly. Videos drive higher engagement through likes, comments, and shares, and trends like short-form and live videos continue to dominate, though issues like misinformation and copyright violations pose challenges.
- **Comments**: Comments allow users to interact with content and others, promoting conversation, feedback, and engagement. They influence content visibility as algorithms prioritize posts with higher interaction rates. However, negative behaviors such as cyberbullying and misinformation can arise in comment sections, necessitating moderation tools.
- **Sharing**: Sharing enables users to amplify content across their networks, increasing visibility and accelerating information dissemination. It plays a vital role in viral trends and social movements. However, it also facilitates the rapid spread of misinformation, prompting platforms to introduce fact-checking and sharing restrictions.
- **Social Media Direct (Messaging)**: Direct messaging allows private one-on-one or group conversations, fostering closer personal and professional relationships. It supports multimedia sharing and real-time communication but also raises

privacy concerns and the risk of misinformation in closed networks.

- **Social Media Stories**: Stories provide a temporary, creative way to share content, offering an ephemeral format that encourages more spontaneous, authentic posting. Popularized by platforms like Snapchat and Instagram, stories have become essential for both personal expression and brand engagement, though their short lifespan makes content moderation challenging.
- **Advertising**: Advertising on social media allows brands to target audiences with precision using user data. Formats like sponsored posts, video ads, and influencer partnerships drive brand awareness and sales. However, privacy concerns and ad fatigue remain significant challenges, pushing platforms to prioritize transparency and user control over ad preferences.
- **Standalone Apps**: Standalone apps, such as Facebook Messenger and Instagram's IGTV, extend the functionality of social media platforms, offering focused, specialized experiences. These apps cater to niche needs like messaging, video sharing, or photo editing, enhancing user engagement while presenting challenges related to user fatigue and fragmented experiences.
- **Third-Party Services**: Third-party services integrate with social media platforms to provide advanced tools for content management, analytics, and campaign tracking. These services help businesses and content creators optimize their presence and gain insights but raise concerns around data privacy and security, leading to stricter platform policies.
- **Fact Checking**: Fact-checking is vital for combating misinformation on social media, with platforms partnering with independent organizations to verify content. Automated systems flag suspicious content, which is then reviewed by human fact-checkers. Despite its challenges, such as biases and the sheer volume of content, fact-checking remains essential for maintaining the integrity of digital information.

S. No.	Idea	Key Point
1	User Content	Users create and share without gatekeepers.
2	Interconnectivity	Connects people through likes, shares, and comments.
3	Engagement	Promotes active interaction, not just passive viewing.
4	Authenticity	Real and relatable content builds trust.
5	Personalization	Content is tailored to user interests, but can cause echo chambers.

Table-1:Social Media Usage

5.CYBERBULLYING ON SOCIAL MEDIA

5.1 PROBLEMS ON SOCIAL MEDIA

Social media platforms have undoubtedly transformed the landscape of communication, information sharing, and entertainment. However, alongside their many benefits, they have also introduced a range of significant problems that affect users on personal, social, and societal levels. One of the most prevalent issues is the rise of online harassment, including cyberbullying, hate speech, and trolling. The anonymity provided by the internet often emboldensindividuals to engage in behaviour they might avoid in face-to-face interactions, leading to an increase in aggressive and harmful online behaviour. Another major problem is the spread of misinformation and fake news, which can influence public opinion, cause widespread panic, or undermine

trust in institutions. Social media algorithms often prioritize sensational content to drive engagement, unintentionally amplifying divisive narratives and conspiracy theories. Additionally, privacy concerns loom large, as users often unknowingly share personal information that can be exploited for malicious purposes. The pressure to present a curated, idealized version of life online can also lead to issues

such as social comparison, low self-esteem, anxiety, and depression, particularly among younger users. Addiction to social media, characterized by compulsive checking and fear of missing out (FOMO), further exacerbates mental health challenges. Despite platform efforts to moderate content and improve user safety, these problems persist, highlighting the urgent need for advanced, automated solutions to identify and mitigate harmful behaviour before it causes real-world damage.

5.2 IMPACT OF CYBERBULLYING

Cyberbullying has far-reaching and often devastating effects on individuals and communities. Unlike traditional bullying, cyberbullying can occur at any time and in any place, making it difficult for victims to find relief. The public nature of online platforms means that harmful comments, images, or videos can quickly reach large audiences, amplifying the humiliation and emotional distress experienced by victims. Victims of cyberbullying often suffer from a range of psychological effects, including anxiety, depression, low self-confidence, and feelings of isolation. In severe cases, prolonged exposure to online harassment has been linked to self-harm and suicidal thoughts, particularly among adolescents and young adults. The impact is not limited to emotional health; academic performance, social relationships, and future career prospects can all be negatively affected. Moreover, cyberbullying can erode the overall quality of online communities, creating hostile environments that discourage healthy discourse and participation. The fear of becoming a target can lead users to self-censor, withdraw from social media, or experience a diminished sense of belonging and trust online. On a broader societal level,unchecked cyberbullying contributes to the normalization of abusive behaviour, fostering cultures of intolerance and division. Addressing the impact of cyberbullying requires a multifaceted approach, including education, community support, effective platform moderation, and technological innovation. Building systems that can detect and address cyberbullying proactively is a critical step towards ensuring that social media remains a space forpositive interaction, creativity, and free expression without fear of harassment or harm.

6. SOCIAL MEDIA ALGORITHM

6.1 FEED

The feed is the central feature of social media platforms, acting as the dynamic stream of content that users see when they open an app. The feed aggregates posts, videos, updates, and stories shared by friends, family, followed accounts, and recommended content. It plays a critical role in shaping user experience and determining what information users consume daily. Social media algorithms manage the order and visibility of content in the feed, aiming to personalize and prioritize what each user is most likely to engage with. This personalization is based on various factors such as user interests, relationships, activity frequency, and platform-specific engagement metrics. The design of the feed impacts user retention, satisfaction, and interaction time, making it a core focus for platform optimization.

6.2 INTEREST

Interest is one of the primary factors considered by social media algorithms when curating a user's feed. Algorithms track what types of contenta user interacts with the most—whether it's liking posts, commenting, sharing, saving, or watching videos. Based on this behaviour, the platform identifies the topics, formats, and accounts that align with the user's preferences. Content related to these identified interests is then ranked higher and shown more frequently. By emphasizing interest-driven content, platforms enhance user engagement and ensure that users find their experience relevant and enjoyable.

6.3 TIMELINESS

Timeliness refers to how recently a piece of content was posted. Social media algorithms prioritize newer content to keep the feed fresh and relevant. Although interest and relationship factors are important, users are generally more inclined to engage with recent updates rather than outdated posts. Timely content helps maintain a real-time feel, encouraging users to check back frequently forthe latest happenings. Platforms balance timeliness with other ranking signals to avoid showing old but highly interacted posts unless they are still contextually important.

6.4 RELATIONSHIP

The relationship between users plays a crucial role in determining the visibility of content in the feed. Algorithms assess interactions such as direct messaging, frequent comments, likes, tags, and profile views to infer how close two users are. Content shared by friends, family members, or accounts with which a user has frequent interactions is more likely to be prioritized. The stronger the relationship signal, the higher the likelihood that the user's content will appear at the top of the feed, promoting more meaningful and personalized social experiences.

6.5 FREQUENCY

Frequency measures how often a user opens the social media platform. Users who check their feed multiple times a day may see a more chronological arrangement of posts, ensuring they do not miss recent updates. In contrast, users who log in less frequently are shown a selection of what the algorithm deems the most relevant or important posts since their last visit. Adjusting feed content based on frequency helps platforms maximize engagement opportunities each time the user logs in, tailoring the experience to different activity patterns.

6.6 FOLLOWING

Following refers to the total number of accounts a user subscribes to on the platform. Users who follow many accounts have a more competitive feed, meaning the algorithm must choose from a vast pool of content to display only the most relevant posts. Conversely, users with fewer followings may see a higher proportion of posts from each account. Algorithms balance content from followed accounts with recommended posts to maintain engagement while introducing users to new creators and communities.

6.7 USAGE

Usage encompasses the overall amount of time a user spends on the platform during each session. Users who spend more time browsing are exposed to a wider range of content, including posts that might rank slightly lower in interest or timeliness. In contrast, for users with shorter sessions, algorithms prioritize only the highest-ranked, most relevant posts to maximize the impact within a limited time frame. Understanding usage patterns allows platforms to optimize content delivery, ensuring users remain engaged whether they are casual browsers or heavy users.

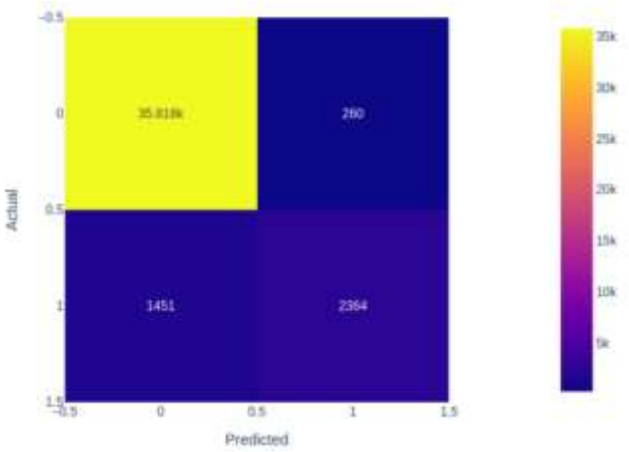


Fig-3: Toxic Level Detection

7.BASICS OF NATURAL LANGUAGE PROCESSING FOR CYBERBULLYING DETECTION

7.1 NATURAL LANGUAGE PROCESSING FOR CYBERBULLYING DETECTION

Natural Language Processing helps detect cyberbullying by analyzing language patterns such as slang, sarcasm, and offensive expressions. Techniques like tokenization, vectorization, and classification are used to proactively identify harmful content on social platforms.

7.2 MACHINE LEARNING TECHNIQUES FOR TEXT CLASSIFICATION

Machine learning models such as Logistic Regression, combined with TF-IDF feature extraction, are employed to detect toxic comments. These models convert text into numerical vectors and classify them as abusive or normal in real time.

7.3 KEY NLP TASKS IN TEXT PREPROCESSING

NLP preprocessing tasks—including tokenization, stop word removal, and lemmatization—transform raw user comments into structured data suitable for effective cyberbullying detection.

7.4 TEXT NORMALIZATION AND CLEANING

Text normalization involves expanding contractions, correcting spellings, and removing elements such as emojis, URLs, and slang. This improves machine readability and enhances the accuracy of detection models.**7.5 SENTENCE STRUCTURE ANALYSIS**

Analyzing sentence structure through syntactic parsing and part-of-speech tagging reveals hidden toxicity and

aggressive tones, even in grammatically complex or sarcastic comments.

7.6WORD MEANING AND CONTEXT UNDERSTANDING

Understanding the meaning of words within their context using methods like TF-IDF or word embeddings enables the detection of sarcasm and indirect abuse in online comments.

7.7 SEMANTIC FEATURE EXTRACTION

Semantic feature extraction allows models to detect abusive intent beyond explicit keywords. This includes identifying coded language or indirect expressions that may not appear overtly toxic.

7.8 SENTIMENT AND TOXICITY ANALYSIS

Sentiment and toxicity analysis focuses on identifying the emotional tone and linguistic patterns in user comments. It helps detect hostile or abusive language, even when expressed subtly.

7.9 RELATIONSHIP DETECTION IN TEXT

This involves identifying syntactic and semantic relationships between words to detect personal attacks, threats, or insults that may be masked in indirect or complex sentence structures.

7.10 CONTEXT AND MEANING EXTRACTION

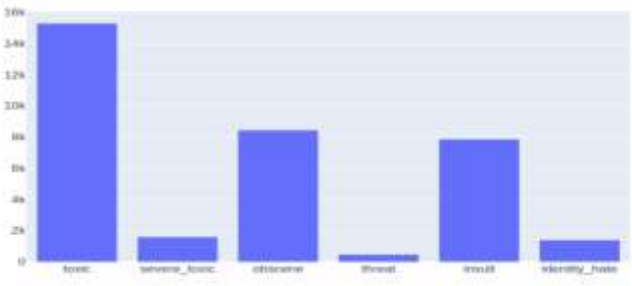
Understanding the overall context of a comment helps avoid misclassification of non-toxic conversations that contain negative words, ensuring accurate cyberbullying detection.

7.11 INFERENCE AND IMPLICATION RECOGNITION

This technique identifies subtle forms of cyberbullying conveyed through sarcasm, innuendo, or suggestive language by analyzing implied meanings and inferred intent.

7.12 TOPIC DETECTION IN USER COMMENTS

Topic detection helps identify themes in user comments,



allowing systems to focus more closely on sensitive subjects that are more prone to bullying and misuse.

8.PERFORMANCE OF CYBERBULLYING DETECTION SYSTEM

8.1 CYBERBULLYING DETECTION ANALYSIS

Cyberbullying detection extends sentiment analysis by identifying harmful content such as hate speech, personal attacks, and insults. This project uses TF-IDF for feature extraction and Logistic Regression for classification. Unlike basic emotion classification, it must recognize sarcasm, coded language, and passive aggression, making feature quality and model robustness crucial to performance.

Fig-4 Label Distribution

8.2 FEATURES

8.2.1 TEXTUAL FEATURES

TF-IDF-based unigrams and bigrams are used to capture both individual toxic words and phrase patterns. Bag-of-Words (BoW) models convert this text into numerical form for analysis.

8.2.2 AUDIO FEATURES

While not implemented here, in multimodal systems, audio features like spectral and prosodic elements help detect emotional cues in speech.

8.3 CLASSIFICATION STRATEGIES

A text-based strategy is used:

- **FEATURE AGGREGATION** combines unigrams, bigrams, and TF-IDF.
- **MODEL TRAINING** is done using Logistic Regression.
- **PREDICTION & EVALUATION** is conducted using standard metrics.This unified approach improves detection of subtle bullying patterns.

8.4 CHALLENGES IN DETECTION

Common issues include:

- ✓ **Negation** ("not terrible")
- ✓ **Sarcasm** ("so smart at 12")
- ✓ **Contextual Polarity** ("killer performance" vs. "killer intent")
- ✓ **Slang & Abbreviations** ("rekt", "salty")
- ✓ **Ambiguity** ("John is worse than Mike")

Such linguistic complexities demand ongoing model updates and contextual understanding for accuracy.

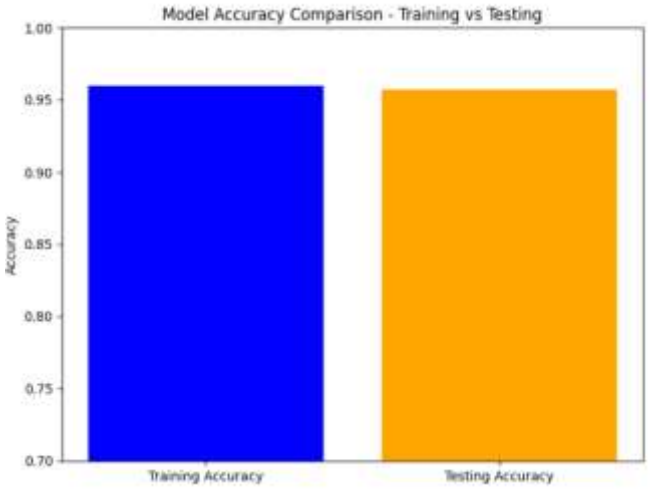


Fig-5:Accuracy Control

9. CONCLUSIONS

In this project, we successfully built a machine learning model to detect toxic comments in online platforms using Natural Language Processing (NLP) techniques and classification algorithms. We analysed a labelled dataset containing comments categorized into various toxicity labels, such as toxic, severe toxic, obscene, threat, insult, and identity hate. The dataset was thoroughly pre-processed, including steps like text cleaning and feature extraction using the TF-IDF method. We used Logistic Regression as the classification model, which achieved a high testing accuracy of approximately 95.71%. The model showed excellent performance in identifying non-toxic comments, and a reasonably good performance for toxic comments, with a weighted F1-score of 0.95. Various evaluation metrics such as the confusion matrix, classification report, and accuracy comparison graphs demonstrated the effectiveness of our approach. The model is capable of significantly assisting in moderating online communities by flagging inappropriate comments.

ACKNOWLEDGEMENT

We thank **God Almighty** for the blessings, knowledge and strength in enabling us to finish our project. Our deep gratitude goes to our founder **Late. Dr. D. SELVARAJ, M.A., M.Phil.,** for his patronage in completion of our project. We take this opportunity to thank our kind and honourable **Chairperson, Dr. S. NALINI SELVARAJ, M.Com., M.Phil., Ph.D.,** and our **Honourable Director, Mr. S. AMIRTHARAJ, B.Tech., M.B.A** for their support to finish our project successfully. We wish to express our sincere thanks to our beloved **Principal, Dr.C.RAMESH BABU DURAI M.E., Ph.D.,** for his kind encouragement and his interest toward us. We are grateful to **Dr.D.C.JULLIE JOSPHINE M.E., Ph.D., Professor and Head of INFORMATION TECHNOLOGY DEPARTMENT,** Kings Engineering College, for his valuable suggestions, guidance and encouragement. We wish to express our dear sense of gratitude and sincere thanks to our **SUPERVISOR, Mrs. K. BENITLIN SUBHA B.TECH.,M.E.,(Ph.D).,** Assistant Professor, Information Technology Department. for her internal guidance. We express our sincere thanks to our parents, friends and staff members who have helped and encouraged us during the entire course of completing this project work successfully

REFERENCES

1. Balakrisnan, V., & Kaity, M. (2023). Cyberbullying detection and machine learning: a systematic literature review. *Artificial Intelligence Review*, 56, 1375–1416. https://link.springer.com/article/10.1007/s10462-023-10553-wSpringerLink
2. Ogunleye, B., & Dharmaraj, B. (2024). The Use of a Large Language Model for Cyberbullying Detection. *arXiv preprint arXiv:2402.04088*. https://arxiv.org/abs/2402.04088arXiv
3. Yadav, A. K., & Patel, H. O. (2025). Cyberbullying detection using machine learning. *AIP Conference Proceedings*, 3224(1), 020062. https://pubs.aip.org/aip/acp/article/3224/1/020062/3335153/Cyberbullying-detection-using-machine-learningAIP Publishing+2AIP Publishing+2AIP Publishing+2
4. Faraj Alqahtani, A., & Ilyas, M. (2024). A Machine Learning Ensemble Model for the Detection of Cyberbullying. *arXiv preprint arXiv:2402.12538*. https://arxiv.org/abs/2402.12538arXiv
5. Tanzin Prama, T., et al. (2025). AI Enabled User-Specific Cyberbullying Severity Detection with Explainability. *arXiv preprint arXiv:2503.10650*. https://arxiv.org/abs/2503.10650arXiv

6. Eronen, J., et al. (2022). Initial Study into Application of Feature Density and Linguistically-backed Embedding to Improve Machine Learning-based Cyberbullying Detection. *arXiv preprint arXiv:2206.01889*. https://arxiv.org/abs/2206.01889 arXiv
7. Atapattu, T., et al. (2020). Automated Detection of Cyberbullying Against Women and Immigrants and Cross-domain Adaptability. *arXiv preprint arXiv:2012.02565*. https://arxiv.org/abs/2012.02565 arXiv
8. Goyal, P., & Goyal, R. (2024). Detection of Cyber-Bullying in Social-Media Using Classification Algorithms of Machine Learning. *ResearchGate*. https://www.researchgate.net/profile/Parul-Goyal-5/publication/381283330_DETECTION_OF_CYBER-BULLYING_IN_SOCIAL-MEDIA_USING_CLASSIFICATION_ALGORITHMS_OF_MACHINE_LEARNING/links/66fa64c6553d245f9e3edc26/DETECTION-OF-CYBER-BULLYING-IN-SOCIAL-MEDIA-USING-CLASSIFICATION-ALGORITHMS-OF-MACHINE-LEARNING.pdfResearchGate
9. Alipour, A. (2019). Twitter Cyberbullying Detection Using Machine Learning. *GitHub Repository*. https://github.com/amiralipour2019/Machine-Learning-Approaches-for-Cyberbullying-Detection GitHub
10. Kumar, A., & Patel, H. O. (2024). Cyberbullying detection and blocking using machine learning. *AIP Conference Proceedings*, 3175(1), 020001. https://pubs.aip.org/aip/acp/article/3175/1/020001/3338955/Cyberbullying-detection-and-blocking-using-machineAIP Publishing+2AIP Publishing+2AIP Publishing+2
11. Zainab Kh. A., Abbas M. B., et al. (2025). Subject Review: Cyberbullying and Detection Methods. *International Journal of Advanced Scientific Research and Engineering*, 11(3). https://ijasre.net/index.php/ijasre/article/download/1877/2182/3332IJASRE
12. Prasad, R., & Sharma, A. (2025). Cyberbullying Detection in Social Media Using Natural Language Processing. *ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S2468227625001838ScienceDirect
13. Kumar, S., & Singh, R. (2025). Detecting Cyberbullying in Social Media: An NLP-Based Classification Framework. *Indian Journal of Science and Technology*. https://indjst.org/articles/detecting-cyberbullying-in-social-media-an-nlp-based-classification-frameworkSRS Journal
14. Yadav, A. K., & Patel, H. O. (2025). An in-depth examination of cyberbullying detection utilizing machine learning. *AIP Conference Proceedings*, 3207(1), 060002. https://pubs.aip.org/aip/acp/article/3207/1/060002/3313286/An-in-depth-examination-of-cyberbullying-detectionAIP Publishing+2AIP Publishing+2AIP Publishing+2
15. Alqahtani, A. F., & Ilyas, M. (2024). A Machine Learning Ensemble Model for the Detection of Cyberbullying. *arXiv preprint arXiv:2402.12538*. https://arxiv.org/abs/2402.12538arXiv
16. Prama, T. T., et al. (2025). AI Enabled User-Specific Cyberbullying Severity Detection with Explainability. *arXiv preprint arXiv:2503.10650*. https://arxiv.org/abs/2503.10650arXiv
17. Eronen, J., et al. (2022). Initial Study into Application of Feature Density and Linguistically-backed Embedding to Improve Machine Learning-based Cyberbullying Detection. *arXiv preprint arXiv:2206.01889*. https://arxiv.org/abs/2206.01889 arXiv
18. Atapattu, T., et al. (2020). Automated Detection of Cyberbullying Against Women and Immigrants and Cross-domain Adaptability. *arXiv preprint arXiv:2012.02565*. https://arxiv.org/abs/2012.02565 arXiv
19. Goyal, P., & Goyal, R. (2024). Detection of Cyber-Bullying in Social-Media Using Classification Algorithms of Machine Learning. *ResearchGate*. https://www.researchgate.net/profile/Parul-Goyal-5/publication/381283330_DETECTION_OF_CYBER-BULLYING_IN_SOCIAL-MEDIA_USING_CLASSIFICATION_ALGORITHMS_OF_MACHINE_LEARNING/links/66fa64c6553d245f9e3edc26/DETECTION-OF-CYBER-BULLYING-IN-SOCIAL-MEDIA-USING-CLASSIFICATION-ALGORITHMS-OF-MACHINE-LEARNING.pdfResearchGate
20. Alipour, A. (2019). Twitter Cyberbullying Detection Using Machine Learning. *GitHub Repository*. https://github.com/amiralipour2019/Machine-Learning-Approaches-for-Cyberbullying-Detection GitHub