International Journal of Leading Research Publication (IJLRP)

E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

SOCIAL MEDIA CYBERBULLYING DETECTOR

G. Naveen Kumar[1], P. Monika[2], K. Soumya[3], J. Haveela[4]

[1]Assistant Professor in Dept. of CSE (AI&ML), Vignan's Institute of Management and Technology for Women, Ghatkesar, Telangana, India

[2,3,4]B. Tech 4th year Students, CSE (AI&ML), Vignan's Institute of Management and Technology for Women, Ghatkesar, Telangana, India

Abstract:

This paper introduces an intelligent Social Media Cyberbullying Detector, a framework designed to strengthen online safety by identifying toxic interactions in real-time. In high-activity digital communities, harmful content can proliferate rapidly, causing significant emotional and psychological distress for users. Traditional manual moderation is increasingly ineffective due to the sheer volume of data, while existing tools often lack the speed required for instantaneous intervention. To address this gap, our system leverages the Reddit API as a live data source, maintaining a continuous stream of user-generated posts and comments for deep-text analysis.

The proposed architecture integrates a real-time data processing pipeline with robust Natural Language Processing (NLP) and machine learning models to distinguish between harmless discourse and genuine harassment. A defining feature of this system is its capacity to recognize offensive language patterns and contextual intent the moment they appear. Upon detection, the system autonomously flags and archives harmful content in a dedicated dataset for further administrative review. To assist moderators in community health assessment, the system also generates live visual analytics, providing a clear statistical breakdown of bullying versus non-bullying content.

By training on labeled cyberbullying datasets, the model is tuned to recognize complex linguistic patterns, ensuring high detection accuracy across various subreddits. This shift from reactive to proactive monitoring enables moderators to mitigate harassment far more effectively than manual oversight. Ultimately, this research provides a scalable, automated solution for continuous surveillance, offering a significant contribution toward maintaining safer and more resilient digital environments.

Keywords: Cyberbullying Detection, Social Media Analysis, Natural Language Processing (NLP), Machine Learning, Reddit API, Real-Time Monitoring, Text Classification, Online Harassment Detection, Offensive Language Identification, LSTM, Logistic Regression, Data Preprocessing, Feature Extraction, Digital Safety, Toxic Comment Detection, Content Filtering, Cyber Safety Systems, AI-based Monitoring.

1. INTRODUCTION:

Social media platforms have become an essential part of daily communication and interaction in the digital age. But as online communities have expanded rapidly, the incidence of harmful behaviours has also sky rocketed including cyberbullying, hate speech and online harassment. Because social media platforms generate a nearly non-stop stream of user data, monitoring such damaging content is difficult, and manual moderation can be slow and inefficient. Traditional content monitoring systems are often limited in detecting context-based abusive language and fail to provide immediate action in real-time environments. To tackle these challenges, this project introduces a Social Media Cyberbullying Detection system that maintains a constant watch on Reddit via its public API. Rather than relying on manual reports, the system automatically pulls user comments and posts, putting them through a pipeline of Natural Language Processing and Machine Learning. The main goal here is real-time classification: identifying where "free

speech" crosses the line into harassment by spotting offensive intent and abusive language patterns as they happen.

On the technical side, the raw text undergoes a cleanup phase—including tokenization and normalization—to ensure the data is ready for a high-accuracy analysis. We utilize trained models like Logistic Regression or LSTM to distinguish between harmless chatter and genuine bullying. Beyond just filtering, it also generates visual breakdowns of community behavior, giving moderators a clear look at the percentage of toxic vs. healthy interactions. Ultimately, this project is about making the internet a safer place by catching harm before it spreads.

Ultimately, this project is a step toward a safer digital culture. By automating the monitoring process on social platforms, we can cut down the response time for catching toxic behavior before it spirals. Combining live data feeds with smart text analysis doesn't just flag mean comments—it provides a scalable way to protect users and support early intervention. It’s an practical solution designed to prioritize digital well-being and keep online spaces healthy for everyone.

2. RELATED WORK:

In recent years, the push to curb cyberbullying has led to a surge in research using Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP). As platforms like Reddit, Twitter, and Facebook have exploded in popularity, so has the volume of toxic content. Traditional moderation—relying on human eyes or basic keyword filters—simply can't keep up with the sheer scale of data. While automated detection has come a long way, these systems still trip up on the nuances of human speech, such as sarcasm, slang, and the way conversations evolve in real-time.

[1] Early efforts often leaned on traditional ML models like Logistic Regression and Support Vector Machines (SVM). These researchers typically used Twitter datasets, relying on TF-IDF for feature extraction. While these models are fast and relatively accurate for short, punchy tweets, they often lose the plot when it comes to longer conversations or deeper context.

[2] To fix that lack of context, some shifted toward Deep Learning using LSTM networks. By analyzing text as a sequence, LSTMs are much better at understanding the relationship between words. However, the trade-off is significant: these models require massive datasets and heavy computing power, making them a challenge to deploy for real-time monitoring.

[3] More recently, Transformer-based models like BERT have set the gold standard for accuracy. By using pre-trained embeddings, they can pick up on subtle cues like sarcasm and hidden intent. The catch? They are hardware-heavy. Running these requires high-end GPUs, which isn't always practical for low-resource or budget-friendly systems.

[4] On the simpler side, many studies have focused on fundamental NLP techniques—things like lemmatization, stop-word removal, and sentiment analysis. These are great for cleaning up "noisy" data, but they often struggle with the messy reality of social media, where emojis, slang, and multi-language "code-switching" are the norm.

[5] We’ve also seen a rise in hybrid models that combine the best of both worlds—like pairing TF-IDF with LSTMs or using SMOTE to balance out datasets. These hybrids are highly effective across different platforms, but they come with their own baggage: they are complex to build and need constant retraining to stay relevant as online slang changes.

Looking at the current landscape, it’s clear that while we have models with high accuracy, we haven't quite solved the "real-world" problem. Most existing solutions are either too slow, too expensive to run, or fail to handle the dynamic nature of live social media. This highlights the urgent need for a system that isn't just accurate in a lab, but is fast and efficient enough to monitor live Reddit feeds and stop harassment as it happens.

3. PROPOSED SYSTEM:

A. Overview of the Proposed System:

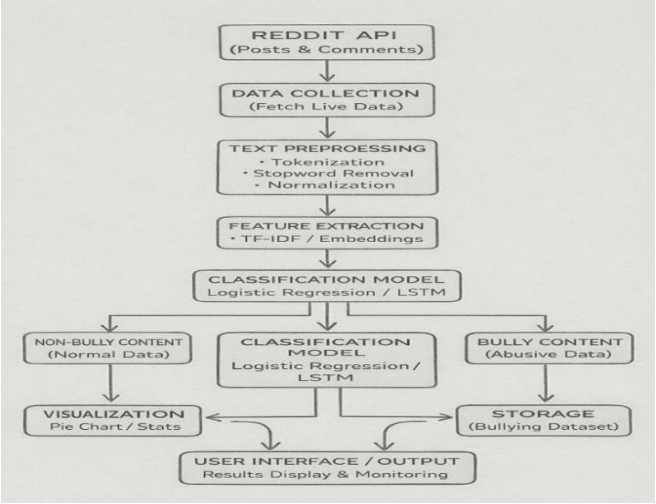
The core of this project is a Real-Time Cyberbullying Detection System, built specifically to make digital spaces safer by catching abusive content as it happens. We've designed the system to tap directly into Reddit's live data feed via its public API, allowing it to pull and analyze a constant stream of posts and comments. By combining Natural Language Processing (NLP) with Machine Learning, the tool doesn't just read text—it understands it, effectively separating harmless discussion from genuine bullying.

What sets this system apart is its instantaneous response. The moment the algorithm identifies offensive or harmful language, it triggers a series of automated actions:

- **Real-Time Classification:** The text is immediately flagged as bullying or non-bullying.
- **Data Isolation:** Any content confirmed as harmful is moved to a dedicated dataset for further review or monitoring.
- **Live Analytics:** The system constantly updates its internal database, tracking the ratio of healthy to toxic interactions.
- **Visual Reporting:** To make this data easy to digest, the results are rendered into live visuals, such as pie charts, giving a clear "health check" of the community.

Ultimately, this tool is designed to act as a proactive watchdog. By giving moderators and users a way to monitor behavior instantly, we can significantly cut down on harassment and ensure that online communities remain a positive space for everyone.

B. System Architecture:



To keep the system efficient and scalable, we've broken the architecture down into six specialized modules. Each plays a specific role in turning raw Reddit data into actionable insights:

1. **Data Collection:** This is the entry point. It interfaces directly with the Reddit API to pull a live stream of posts and comments.
2. **Text Processing:** Before any analysis happens, the raw text is "cleaned." This involves stripping out noise like punctuation and stop-words, performing tokenization, and normalizing the text to ensure the model isn't distracted by formatting.
3. **Feature Extraction:** Since machines can't read English, this module translates cleaned text into a numerical format. We use techniques like TF-IDF or word embeddings to represent the underlying patterns of the language.
4. **Classification:** This is the "brain" of the system. Here, a trained model—typically Logistic Regression or an LSTM network—determines whether a piece of content is bullying or non-bullying.
5. **Storage:** Any content flagged as harmful doesn't just disappear; it's funneled into a dedicated dataset for long-term monitoring and further study.

6. **Visualization:** Finally, the system translates the raw data into something a human can understand at a glance, using live charts and percentage breakdowns to show the overall "vibe" of the community.

C. Data Collection Module:

The collection process is the backbone of the system's real-time capability. By tapping into Reddit's official API, we can target specific subreddits or filter by high-risk keywords. Unlike static datasets, this module works on a continuous loop, extracting new posts and comments as they are published.

The moment data is collected, it is processed and analyzed instantly. If the system catches something abusive, it immediately flags the entry and archives it. This creates a constant "surveillance loop," ensuring that online conversations are monitored 24/7 without the need for manual oversight.

METHODOLOGY:

Step 1: Problem Identification

Cyberbullying is increasing on social media platforms, affecting users mentally and emotionally. Manual monitoring is not sufficient. Therefore, an automated system is required.

Step 2: Tools and Technologies

Programming Language: Python
Platform: Google Colab
Libraries: NLTK, Pandas, NumPy, Scikit-learn, Matplotlib
Data Source: Reddit API

Step 3: Data Preprocessing

The text is cleaned using:
• Lowercase conversion
• Removal of special characters
• Removal of stop words
• Tokenization
• Lemmatization

Step 4: Feature Extraction

The text is converted into numerical format using:
• TF-IDF Vectorizer

Step 5: Model Training

The model is trained using labeled cyberbullying datasets.
It learns patterns of abusive language and harmful expressions.

Step 6: Real – Time Detection

The trained model analyzes live Reddit data and classifies it as:
• Bullying
• Non-Bullying

Step 7: Visualization

The system generates:
• Pie chart showing percentage of bullying and non-bullying
• Table of detected bullying comments

4. FUTURE SCOPE:

The system can be improved with:
• Use of advanced models like BERT and Transformers
• Multilingual cyberbullying detection
• Real-time alert notifications for moderators
• Integration with multiple social media platforms
• Browser extension for live content filtering

These improvements can make the system more accurate and scalable.

5. IMPLEMENTATION:

A. Project Environment & Libraries

First, we set up the Python environment in a Colab notebook. We use **PRAW** (Python Reddit API Wrapper) for data collection and **Scikit-learn** for the classification "brain" of the system.

- **PRAW**: Connects directly to the Reddit API to pull a live stream of posts.
- **NLTK**: Used for natural language processing, such as stripping "noise" from text.
- **Scikit-learn**: Provides the TF-IDF vectorizer and the Logistic Regression model.

B. Real-Time Data Collection

The backbone of this system is the **Data Collection Module**. Unlike static datasets, this module operates on a continuous loop. In Colab, we provide our API credentials (client ID and secret) to maintain a constant watch on specific subreddits. The moment a user submits a comment, the system pulls it for immediate analysis.

C. Text Preprocessing & Feature Extraction

Raw social media text is often "noisy" with punctuation and slang. To ensure high accuracy, the text undergoes a cleanup phase:

- **Normalization**: Converting all text to lowercase.
- **Cleaning**: Removing special characters and stop-words (common words like "the" or "is").
- **Tokenization**: Breaking the sentences into individual words or "tokens".
- **TF-IDF Vectorization**: Since machines cannot read English, we translate the cleaned text into a numerical format that represents language patterns

D. Classification & Detection

This is the core module where the trained model determines if a comment is "Bullying" or "Non-Bullying".

- **Logistic Regression/LSTM**: These models are used to distinguish between harmless chatter and genuine harassment.
- **Real-Time Flagging**: If the model identifies offensive intent, the system immediately flags the content.

E. Data Isolation & Visualization

Once toxic behavior is caught, the system performs two final actions:

- **Storage**: Harmful content is funneled into a dedicated dataset for further administrative review.
- **Visual Reporting**: The results are rendered into live visuals, such as **pie charts**, in the Colab output cell. This provides a clear "health check" of the community by showing the percentage of toxic versus healthy interactions.

6. CONCLUSION:

This research highlights a practical approach to tackling online toxicity through a Real-Time Cyberbullying Detection System specifically optimized for Reddit. By merging Natural Language Processing with Machine Learning, the proposed framework moves beyond static analysis to catch abusive content as it happens. We have developed a system that is not only efficient and scalable but also capable of the high-speed monitoring required by modern social platforms. By reducing the window between a harmful post and its detection, this tool serves as a vital step toward a safer digital culture. With continued refinement, this architecture offers a viable, large-scale solution for automated moderation across the wider social media landscape.

REFERENCES:

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber *"Automated Hate Speech Detection and the Problem of Offensive Language"* Proceedings of the 11th International Conference on Web and Social Media (ICWSM), 2017.

[2] E. Wulczyn, N. Thain, and L. Dixon,"Ex Machina: Personal Attacks Seen at Scale,"*Proceedings of the 26th International World Wide Web Conference (WWW)*, 2017.

[3] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma,"Deep Learning for Hate Speech Detection in Tweets,"*Proceedings of WWW Conference Companion*, 2017.
[4] Z. Zhang, D. Robinson, and J. Tepper,"Detecting Hate Speech on Twitter Using a Convolutional Neural Network,"*Proceedings of the 15th International Conference on Web and Social Media*, 2018.
[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova,"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"*NAACL-HLT*, 2019.
[6] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi,"Hate Me, Hate Me Not: Hate Speech Detection on Facebook,"*Proceedings of ITASEC*, 2017.
[7] S. Salminen, J. Almerekhi, M. Milenković, et al.,"Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media,"*Proceedings of ICWSM*, 2018.
[8] F. Pedregosa et al.,"Scikit-learn: Machine Learning in Python,"*Journal of Machine Learning Research*, 2011.
[9] S. Bird, E. Klein, and E. Loper,"Natural Language Processing with Python,"*O'Reilly Media*, 2009. [10]Reddit, "Reddit API Documentation,"Available: https://www.reddit.com/dev/api/