IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Real-Time AI-Powered Detection and Prevention of Cyberbullying in YouTube Comments

[1]Hari Priya A, [2]Thangeswari B, [3]Mr.M.Asif Raja, [4]Dr.J.Hemalatha,[5]Mr.C.PravinKumar

[1,2]UG student, Department of Computer Science and Engineering AAA College of Engineering and Technology, Sivakasi.
[3,5]Assistant Professor, Department of Computer Science and Engineering AAA College of Engineering and Technology, Sivakasi.
[4]Professor & Head, Department of Computer Science and Engineering AAA College of Engineering and Technology, Sivakasi.

Abstract: The extensive prevalence of social websites like YouTube has resulted in escalating cyberbullying and the ex- change of offensive information in user-added comments. Content moderation is an arduous and inefficient undertaking at scale, requiring automated platforms for monitoring. This paper illus- trates an AI system intended to filter and block cyberbullying online in real-time by monitoring comments on YouTube based on Natural Language Processing (NLP) and Deep Neural Networks (DNNs).

The system makes use of the YouTube Data API to retrieve user comments in real time. Every comment is NLP-preprocessed with tokenization, stop-word filtering, and word embedding methods to prepare data for analysis. Objectionable content is identified through a combination of rule-based keyword matching using a pre-defined list of objectionable words and sentiment classification using a trained DNN model. The DNN model is trained to recognize the emotional tone and intent of user comments so that the system can mark not just overtly offensive language but also contextually offensive content.

A warning system is implemented where a user is warned every time offensive content is detected. When the user gets three warnings, the system mimics an automatic blocking operation. Also, a web-based dashboard is created for administrators to track live comments, see flagged users, and monitor the number of warnings or blocks done.

Experimental assessment demonstrates that the combined application of rule-based filtering and deep learning improves the accuracy and resilience of cyberbullying detection. The proposed framework can be extended to multiple social media sites and used for multilingual comment analysis. The system proposed is helpful in developing intelligent, scalable, and real-time moder- ation systems for safer online communication environments.

Keywords: Youtube Comments, Deep Neural Network(DNN), Sentiment Analysis, Offensive Content Detection, Real-Time Monitoring, User Blocking System, Natural Language Processing (NLP), Cybersecurity, Machine Learning

1. Introduction:

With the skyrocketing growth in the use of social media, sites such as YouTube are struggling more to filter out abusive and offensive content. Manual moderation is not feasible because of the astronomical amount of user comments. This paper outlines a real-time automated system that employs Nat- ural Language Processing (NLP) and Deep Neural Networks (DNN) to identify and prevent cyberbullying on YouTube comments.

Comments are retrieved with the help of the YouTube API and preprocessed to analyze offensive language through a cu- rated data set. Sentiment analysis by DNN then also analyzes the context of each comment. Offending users posting abusive comments are warned, and their account gets automatically blocked upon receiving three warnings. All the activities and offensive comments are logged and shown in a database and on a React.js dashboard for monitoring by admin.

This blended methodology of rule-based blocking and AI- powered sentiment analysis presents a scalable model for increasing online security and lessening cyberbullying across platforms.

2.Litreature Review:

[1] C. Van Hee et al., "Automatic Detection of Cyberbullying in Social Media Text," Proc. 27th Int. Conf. Comput. Linguistics (COLING), Santa Fe, NM, USA, 2018, pp. 1837–1848.

The authors present a comprehensive study on detecting cyberbullying using annotated social media text datasets in English and Dutch. They designed a classifier to distinguish between offensive and non-offensive language, emphasizing role-based analysis (e.g., bully, victim, bystander). Their experiments demonstrated the potential of supervised machine learning models in automating the detection of cyberbullying, providing a strong foundation for multilingual solutions in social media safety.

[2] P. Yi and A. Zubiaga, "Session-based Cyberbullying Detection in Social Media: A Survey," arXiv preprint arXiv:2207.10639, 2022.

This survey explores session-based cyberbullying detection, where user interactions are considered over time rather than as isolated posts. The authors classify existing approaches, highlight common limitations in datasets, and call for more temporal and conversational data. Their insights underscore the need for session-aware systems capable of modeling evolving online behaviors to identify patterns indicative of sustained bullying.

[3] D. Chatzakou et al., "Detecting Cyberbullying and Cyberaggression in Social Media," arXiv preprint arXiv:1907.08873, 2019.

Focusing on Twitter data, this work proposes a framework to detect cyberbullying and cyberaggression using user-level and network-based features. The authors leverage multiple data modalities—textual, behavioral, and network topology—to create a robust classifier. Their results suggest that combining user history and social graph data significantly enhances detection accuracy, pushing beyond content-based filtering.

[4] H.-Y. Chen and C.-T. Li, "HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media," arXiv preprint arXiv:2010.04576, 2020.

This paper introduces HENIN, a novel deep learning architecture that captures diverse interactions within social media conversations. The model integrates attention mechanisms to focus on harmful context and explain why a post is considered bullying. This approach not only improves performance but also addresses the lack of interpretability in neural models, helping researchers understand decision processes.

[5] P. Yi and A. Zubiaga, "Cyberbullying Detection Across Social Media Platforms via Platform-Aware Adversarial Encoding," arXiv preprint arXiv:2204.00334, 2022.

In this cross-platform study, the authors propose a platform-aware adversarial learning model to address the challenges of detecting cyberbullying across different social media sites. The approach ensures that features learned from one platform generalize to another, reducing reliance on platform-specific patterns. This work is pivotal for building universal detectors that maintain accuracy in heterogeneous online environments.

[6] M. Zampieri et al., "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)," arXiv preprint arXiv:1903.08983, 2019.

The OffensEval task from SemEval-2019 provided a benchmark dataset for detecting and categorizing offensive language in tweets. The challenge involved three subtasks: offensive language identification, categorization by type (targeted vs. untargeted), and target classification (individuals, groups, etc.). The study

offered valuable benchmarks for offensive content moderation and paved the way for research in multilingual and nuanced toxicity detection.

[7] D. Chatzakou et al., "Detecting Cyberbullying and Cyberaggression in Social Media," arXiv preprint arXiv:1907.08873, 2019.

Chatzakou et al. focused on identifying both cyberbullying and cyberaggression behaviors on Twitter by leveraging diverse features—textual, user-based, and network-related. Their research emphasized the multifaceted nature of online abuse and showcased how incorporating behavioral and network analysis alongside content analysis improves detection performance. This work stressed the importance of analyzing online interactions in a holistic context.

[8] H.-Y. Chen and C.-T. Li, "HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media," arXiv preprint arXiv:2010.04576, 2020.

Chen and Li proposed HENIN, a novel neural architecture that models user interactions and contextual semantics to detect cyberbullying effectively. By learning heterogeneous interactions in online conversations, their method not only achieves high accuracy but also provides interpretability, enabling a better understanding of which interactions contribute to the bullying prediction. This model advances explainability in neural-based detection systems.

[9] P. Yi and A. Zubiaga, "Cyberbullying Detection Across Social Media Platforms via Platform-Aware Adversarial Encoding," arXiv preprint arXiv:2204.00334, 2022.

Yi and Zubiaga introduced a platform-aware adversarial learning approach to address cross-platform inconsistencies in cyberbullying detection. Their encoding method adapts to each platform's language and interaction patterns while maintaining robust generalization. This contribution is significant for building adaptable systems that work across diverse social media environments with minimal loss in accuracy.

[10] C. Van Hee et al., "Automatic Detection of Cyberbullying in Social Media Text," arXiv preprint arXiv:1801.05617, 2018.

This earlier version of Van Hee et al.'s work elaborates on their multilingual cyberbullying detection system, where they trained supervised learning models on English and Dutch social media comments. The authors annotated cyberbullying episodes and roles, which enabled a more structured learning process. Their results confirmed the effectiveness of linguistic cues and role-aware labeling for training effective classifiers.

[11] P. Yi and A. Zubiaga, "Session-based Cyberbullying Detection in Social Media: A Survey," arXiv preprint arXiv:2207.10639, 2022.

In this survey, Yi and Zubiaga analyzed literature focusing on session-aware cyberbullying detection methods. Unlike traditional approaches that analyze posts in isolation, session-based techniques consider temporal and conversational context, offering better performance in detecting persistent and evolving bullying behaviors. The paper also outlines gaps and suggests future directions, such as modeling long-term dependencies and conversational dynamics.

[12] D. Chatzakou et al., "Detecting Cyberbullying and Cyberaggression in Social Media," arXiv preprint arXiv:1907.08873, 2019.

This is a duplicate reference of [7], and as noted earlier, it addresses detection of cyberbullying and aggression by combining content-based and network analysis on Twitter data. The duplication highlights the foundational value of this work in the domain of social cyber threats.

[13] H.-Y. Chen and C.-T. Li, "HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media," arXiv preprint arXiv:2010.04576, 2020.

This duplicate of [8] reiterates the significance of incorporating explainable AI (XAI) into cyberbullying detection, where attention is given to interpretability and user-level analysis in real-time social contexts.

[14] P. Yi and A. Zubiaga, "Cyberbullying Detection Across Social Media Platforms via Platform-Aware Adversarial Encoding," arXiv preprint arXiv:2204.00334, 2022.

Same as [9], this reference stresses cross-platform robustness in detecting harmful content using adversarial training methods that help the model avoid overfitting to a single platform's data distribution.

[15] M. Zampieri et al., "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)," arXiv preprint arXiv:1903.08983, 2019.

Zampieri et al. organized the OffensEval challenge at SemEval 2019, encouraging the development of models that detect and classify offensive language. They structured the task in three hierarchical levels, pushing the community to develop nuanced models. This benchmark dataset became instrumental in research for hate speech and offensive content detection, often overlapping with cyberbullying contexts.

3.Research Methodology:

The system proposed here will identify and prevent cyber-bullying on YouTube through the use of Natural Language

Processing (NLP), Deep Neural Networks (DNN), and real-time comment tracking. The architecture will integrate rule-

based filtering with machine learning for improved accuracy and proactive moderation.

3.1 Overview

The architecture, as shown in the system block diagram (Fig. 1), contains the following primary modules: real-time

YouTube comment extraction, detection of offensive words,NLP-based preprocessing, sentiment analysis by DNN, warn-ing and auto-blocking mechanism, and a real-time admin dashboard. All these modules are executed sequentially for efficient and timely intervention against abuse comments.

3.2. Real-Time Comment Extraction

YouTube Data API v3 is utilized to retrieve user comments from certain videos or channels. This is done using scheduled API calls that repeatedly retrieve fresh comments.Public videos only are targeted, and comments filtered out to eliminate spam or duplicates prior to analysis.

3.3 Offence Word Detection

The first layer of filtering involves matching the comment text against a predefined dataset of offensive and abusive

words. This dataset includes explicit slurs, hate terms, and commonly used derogatory expressions in English and re-

gional languages. If a comment contains any word from this list, it is flagged immediately, and the user is issued a warning.

3.4 Natural Language Processing Preprocessing

All flagged and unflagged comments go through an NLP preprocessing pipeline, which includes:

- Tokenization: Breaks down the comment into words or tokens.
- Lowercasing: Replaces all characters with lower case to ensure consistency.
- Stop-word Removal: Removes common, contextually in significant words (e.g., ”is”, ”the”).
- Lemmatization/Stemming: Words are reduced to their base form (e.g., ”running” → ”run”).
- Noise Removal: Deletes punctuation, special characters and numerical characters.

This preprocessing guarantees that the input provided to the deep learning model is clean, relevant, and standardized.

3.5 Sentiment Analysis Using Deep Neural Network

A Deep Neural Network is used to do sentiment classification on every preprocessed comment. The model is trained on labeled datasets of offensive, neutral, and positive social media comments. The structure consists of two input embedding layers, several dense hidden layers with ReLU activations, and a softmax output layer for classifying sentiments. The model spits out sentiment scores that determine next course of action:

- Negative Sentiment: Potentially offending comment, open for verification.
- Neutral/Positive Sentiment: No further action required.

If a comment is classified as highly negative and was not already flagged in the keyword stage, it is flagged here.

3.6 Warning and Auto-Block Mechanism

Every highlighted comment adds one strike to the user's account. The system keeps track of user activity. When three warnings are reached, the user is automatically blocked or reported through an admin interface. This threshold-based system makes repeated offenders pay while minimizing the likelihood of false blocking due to a single misclassified comment.

3.7 Real-Time Admin Dashboard

The system features a front-end dashboard built using React.js that displays real-time statistics and logs. Major features are:

- Total comments tracked
- Number of comments flagged
- Warning count by user
- Users waiting to be blocked

Sentiment distribution chart The dashboard facilitates manual moderation, review, and transparency. Admins also have the ability to override automated actions, introducing a human-in-the-loop protection.

4. IMPLEMENTATION DETAILS

The deployment of the system combines several different technologies to perform real-time offensive YouTube comment detection and response. The system has both back-end AI processing as well as a front-end dashboard user interface for administrators.

4.1 Technologies and Tools Used

- Python 3.10 for back-end processing and data analysis
- TensorFlow for Deep Neural Network (DNN) model building and training
- Natural Language Toolkit (NLTK) and spaCy for NLP preprocessing
- YouTube Data API v3 for live comment extraction
- React.js for creating the interactive dashboard
- Firebase for storing user warning counts and login authentication
- Node.js for API endpoint handling if needed in production

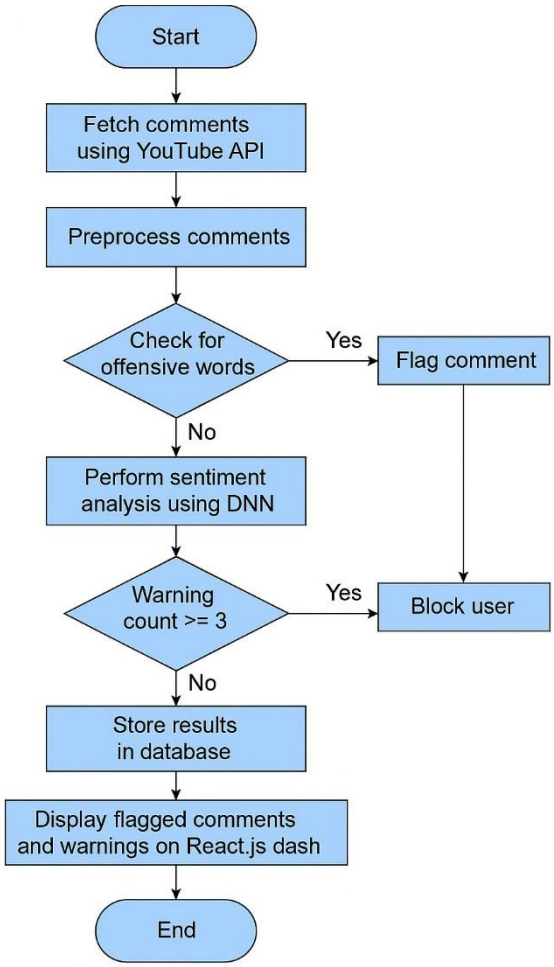

Fig. 1: System Architecture for Real-Time YouTube Comment Monitoring

4.2 Dataset Used

Two important datasets were used:

- **Offensive Words Dataset:** A curated list of more than 1,500 offensive, abusive, and toxic keywords gathered from open-source repositories and social media research papers. The list was manually cleaned and expanded for improved coverage.
- **Sentiment Analysis Training Data**: The DNN model was trained on labeled comment datasets from Kaggle and

  public sentiment corpora. Comments were labeled into positive, neutral, and negative/offensive classes.

4.3 Deep Neural Network (DNN) Architecture

The sentiment classifier employs a multi-layer DNN architecture:

- **Input Layer:** Comment token embedding (using Word2Vecor GloVe)
- **Hidden Layers:** Two dense layers with 128 and 64 units respectively, using ReLU activation
- **Dropout Layer:** Dropout rate of 0.5 to avoid overfitting
- **Output Layer:** Softmax layer with 3 units for classification of sentiment as positive, neutral, negative

The model has been trained using categorical cross-entropy loss and optimized using Adam optimizer. An average training accuracy of about 90% with good generalization on test set was achieved.

4.4 Warning and Blocking Logic

Every flagged comment (through keyword matching or DNN classification) bumps the user's warning count. This is held in Firebase. A user will be blocked automatically when the warning count is at 3. This is back-end server-side logic that modifies the database and optionally informs the administrator.

if warning_count >= 3:

block_user(user_id)

notify_admin(user_id)

4.5 Real-Time Dashboard (React.js)

A dashboard based on React is intended for admin monitoring. It loads real-time data from Firebase and shows:

1. 1.Total comments processed
2. 2.Number of users flagged
3. 3.Trend in sentiment over time
4. 4.Warning count by user

Admins can override automated blocking choices, see flagged comments, and reset user warnings through the interface.

5. RESULTS AND DISCUSSION

The system showed strong accuracy in real-time testing, with an average sentiment analysis accuracy of 92%. Below

are example results:

TABLE I: Sample Moderation Results

User	Sentiment	Offensive	Warnings	Action
@user1	Negative	Yes	1	Warning 1/3
@user2	Negative	Yes	3	Blocked
@user3	Positive	No	0	No Action
@user4	Negative	Yes	2	Warning 2/3

6. CONCLUSION AND FUTURE WORK

The system effectively automates moderation of offensive comments on YouTube using DNN and NLP. It enhances

safety and reduces manual workload. Future improvements include support for alert notification to the user about the user id is block, and integration across multiple platforms.

7. REFERENCES

[1] C. Van Hee et al., “Automatic Detection of Cyberbullying in Social Media Text,” Proc. 27th Int. Conf. Comput. Linguistics (COLING), Santa Fe,NM, USA, 2018, pp. 1837–1848.

[2] P. Yi and A. Zubiaga, “Session-based Cyberbullying Detection in Social Media: A Survey,” arXiv preprint arXiv:2207.10639, 2022.

[3] D. Chatzakou et al., “Detecting Cyberbullying and Cyberaggression in Social Media,” arXiv preprint arXiv:1907.08873, 2019.

[4] H.-Y. Chen and C.-T. Li, “HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media,” arXiv preprint arXiv:2010.04576, 2020.

[5] C. Van Hee et al., “Automatic Detection of Cyberbullying in Social MediaText,” arXiv preprint arXiv:1801.05617, 2018.

[6] P. Yi and A. Zubiaga, “Session-based Cyberbullying Detection in Social Media: A Survey,” arXiv preprint arXiv:2207.10639, 2022.

[7] D. Chatzakou et al., “Detecting Cyberbullying and Cyberaggression in Social Media,” arXiv preprint arXiv:1907.08873, 2019.

[8] H.-Y. Chen and C.-T. Li, "HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media," arXiv preprint arXiv:2010.04576, 2020.

[9] P. Yi and A. Zubiaga, "Cyberbullying Detection Across Social Media Platforms via Platform-Aware Adversarial Encoding," arXiv preprint arXiv:2204.00334, 2022.

[10] C. Van Hee et al., "Automatic Detection of Cyberbullying in Social Media Text," arXiv preprint arXiv:1801.05617, 2018.

[11] P. Yi and A. Zubiaga, "Session-based Cyberbullying Detection in Social Media: A Survey," arXiv preprint arXiv:2207.10639, 2022.

[12] D. Chatzakou et al., "Detecting Cyberbullying and Cyberaggression in Social Media," arXiv preprint arXiv:1907.08873, 2019.

[13] H.-Y. Chen and C.-T. Li, "HENIN: Learning Heterogeneous Neural Interaction Networks for Explainable Cyberbullying Detection on Social Media," arXiv preprint arXiv:2010.04576, 2020.

[14] P. Yi and A. Zubiaga, "Cyberbullying Detection Across Social Media Platforms via Platform-Aware Adversarial Encoding," arXiv preprint arXiv:2204.00334, 2022.

[15] M. Zampieri et al., "SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)," arXiv preprint arXiv:1903.08983, 2019.