WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Virtual Watchdog Identifying Hidden Threats in Online Spaces

Ms. A. Suhana [a], S. Aarthi [b], S. Archana [c], V. Devibhalambigai [d]

[a] *Assistant Professor, Department Of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, India.*
[b, c, d] *Student, Department Of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, India.*

ABSTRACT:

Every individual has the right to express their opinions freely. However, this freedom is sometimes misused to spread hatred and discrimination against individuals or groups based on traits such as race, religion, gender, ethnicity, nationality, disability, or sexual orientation. This form of expression is known as hate speech and can take the form of spoken words, written text, gestures, or visual displays. With the rapid rise of social media and online platforms, the dissemination of hate speech has significantly increased, contributing to real-world hate crimes. While these platforms have facilitated easier communication and information sharing, they have also made it easier to spread harmful content. In response, recent studies have utilized machine learning and deep learning techniques in conjunction with text mining methods to detect and filter hate speech in real-time. This project aims to analyze social media comments using Natural Language Processing (NLP) and a Deep Learning technique known as VADER (Valence Aware Dictionary for Sentiment Reasoning). VADER is used to extract keywords from user-generated content and classify the sentiment as either positive or negative. If a comment is identified as negative, the system blocks it based on the user's preferences and applies predefined threshold rules to block repeated offenders. The proposed model is implemented in a real-time social networking environment with an enhanced notification system that alerts users and administrators of harmful behavior, thereby contributing to safer and more respectful online communication.

Keywords: Cyber security, NLP, Hate Speech Detection, Sentiment Analysis, VADER Algorithm.

INTRODUCTION

Cyber bullying and hate speech have become pressing issues in the digital era, especially with the rapid growth of social networking platforms. While these platforms offer seamless communication and information sharing, they are often exploited to spread harmful content targeting individuals or groups based on race, gender, religion, or personal beliefs. Such content can lead to serious mental and emotional distress, especially when it is left unregulated. This project is designed to counteract this issue by developing an intelligent system that can detect, classify, and manage offensive content in real-time using Natural Language Processing (NLP) and deep learning techniques. The core idea of this system is to enable users to take control of the messages and comments posted on their profiles. Unlike traditional content-based filtering techniques that mainly focus on document classification, this system introduces a hybrid model using the VADER sentiment analysis tool, a deep learning classifier, and a rule-based mechanism to identify and filter inappropriate or harmful messages. The system processes comments posted on social media platforms, classifies them based on sentiment, and applies filtering rules defined by the user or administrator. This ensures a personalized and flexible moderation approach that adapts to the user's preferences and sensitivities. In addition to filtering negative content, the system also includes a mechanism to block malicious users who repeatedly post harmful messages. If a user crosses a set threshold of violations, they are blacklisted, and the system alerts the target user, even when they are offline. This helps in minimizing exposure to toxic content and improves user safety.

2. LITERATURE SURVE

2.1. DEEP LEARNING FOR HATE SPEECH DETECTION: A COMPARATIVE STUDY

Malik et al. (2022) conduct a comparative study of various deep learning models used for hate speech detection, with the goal of identifying the most effective approaches. The paper evaluates multiple models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and transformer-based models like BERT. The authors assess these models based on their accuracy, robustness, and ability to handle imbalanced datasets. One key aspect of the study is the hyper parameter tuning process, where the authors experiment with different configurations to optimize model performance. Additionally, the paper discusses the challenges associated with detecting subtle hate speech and contextual nuances in online communication, such as sarcasm and code-switching. Malik et al. conclude that while deep learning models have shown great promise, further improvements in model generalization are necessary to handle the diversity of language used in online hate speech.

2.2. HCOVBI-CAPS: HATE SPEECH DETECTION USING CONVOLUTIONAL AND BI-DIRECTIONAL GATED RECURRENT UNIT WITH CAPSULE NETWORK

Khan et al. (2022) introduce HCovBi-caps, a hybrid model for hate speech detection that combines Convolutional Neural Networks (CNNs), Bi-directional Gated Recurrent Units (BiGRUs), and a Capsule Network. The authors argue that while CNNs and BiGRUs are effective for capturing local patterns and sequential dependencies in text, respectively, they fail to fully capture the hierarchical relationships and spatial representations of words in hate speech. To address this, they incorporate a Capsule Network, which is designed to preserve positional relationships between objects and provide a more nuanced understanding of text structure. The resulting HCovBi-caps model outperforms traditional methods in terms of both accuracy and interpretability. The paper presents extensive experiments using various publicly available datasets, including Twitter and Reddit, and demonstrates that the Capsule Network enhances the model's ability to detect both explicit and implicit hate speech. The authors conclude that HCovBi-caps provides a promising solution for improving hate speech detection, particularly in environments where text may exhibit complex hierarchical relationships.

3. SYSTEM STUDY

3.1. EXISTING SYSTEM

In the existing content-based filtering systems, each user is treated as an independent entity. Unlike collaborative filtering systems, which rely on the correlation between similar users, content-based filtering works by analyzing the specific content within the message. This approach has been traditionally used in areas like email filtering, news articles, and internet resources. Since the documents in content-based filtering are primarily textual, the task can be modelled as binary classification, separating incoming content into relevant and non-relevant categories. Content-based filtering generally involves the use of machine learning algorithms to create a classifier that can automatically learn from a set of pre-classified examples. These systems often use feature extraction methods like the Bag of Words (BoW) approach, which converts text into a numerical format that represents the frequency of words within the text. Although BoW has been proven to provide good performance in many cases, more sophisticated text representations could potentially offer superior semantics, though they might suffer from lower statistical quality. In addition, recent developments in content-based filtering have extended to multi-label text categorization, where messages are labeled across multiple thematic categories instead of just being classified as relevant or non-relevant. Despite its advantages, the existing systems face several challenges. One of the main limitations is the inability to filter unwanted messages on social networks effectively. Current systems struggle to handle posts that contain images, and they often fail to analyze short text tags accurately. Additionally, automatic blocking of inappropriate content is not a feasible solution with traditional filtering methods. The existing systems lack the ability to adaptively block offensive content based on user-defined preferences and relationships, which results in a more generalized filtering approach that may not align with individual user needs.

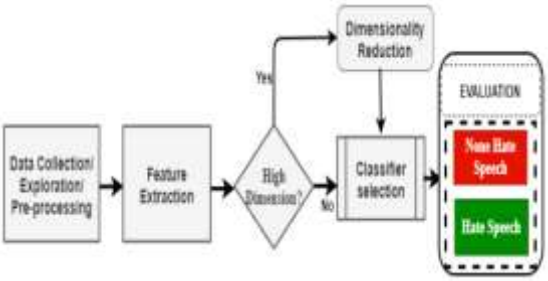

Figure 3.1.1: Existing system

3.2. PROPOSED SYSTEM

Online Social Networks (OSNs) have become an essential part of daily life, providing a platform for communication, information sharing, and social interaction. However, one of the major challenges faced by these platforms is the lack of control users have over the content posted on their walls, leading to the appearance of unwanted or offensive messages. To address this issue, the proposed system introduces a flexible rule-based mechanism that empowers users to control the content displayed on their profiles. By implementing a machine learning-based soft classifier, the system automatically labels messages based on their content, enabling content-based filtering and helping users maintain a cleaner and safer online environment. The system uses Deep Learning (DL) techniques for text classification, where messages are categorized according to their content. The VADER (Valence Aware Dictionary for Sentiment Reasoning) approach is applied to extract and classify sentiment in comments, identifying whether they are positive, neutral, or negative. If a message is flagged as negative, it is immediately blocked or filtered out. Furthermore, the system utilizes a blacklist (BL) mechanism to filter out inappropriate words and block users who repeatedly post harmful or offensive content, ensuring that the user's social media experience remains positive and free from cyberbullying or hate speech. An essential feature of the proposed system is the alert system, which notifies users when a friend continues posting negative content. This alert system uses threshold values to identify users who frequently post unwanted comments and automatically block them after exceeding a certain limit. Additionally, users receive notifications on their mobile devices when such incidents occur, allowing them to take immediate action. This comprehensive solution provides users with direct control over their social media experience, significantly improving the safety and quality of interactions on OSNs.

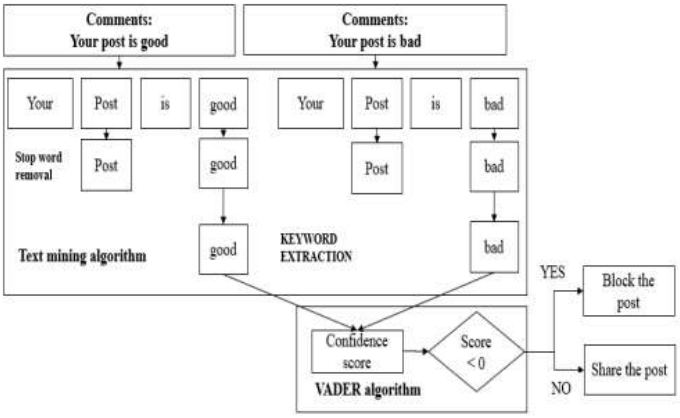

Figure 3.2.1: Proposed system

Feature	Existing System	Proposed System
Processing Type	**High**	**Optimized**
Automated Blocking	**No**	**Yes**

4. METHODOLOGY

4.1 SENTIMENT ANALYSIS USING VADER

The system incorporates the VADER (Valence Aware Dictionary and sentiment reasoner) sentiment analysis tool to analyze the sentiment polarity of user-generated content. VADER is chosen for its effectiveness in capturing sentiment in social media text, including slang, emojis, and capitalization.

4.2 DEEP LEARNING-BASED CLASSIFIER

A trained deep learning classifier is used in conjunction with VADER to improve the accuracy of detecting harmful, offensive, or bullying messages. The model is trained on a labeled dataset of cyber bullying content and can recognize evolving language patterns in user comments and posts

4.3 REAL-TIME CONTENT FILTERING AND MODERATION

The system operates in real-time, evaluating each comment before it is posted. Based on sentiment and context: Acceptable comments are published. Suspicious or harmful comments are flagged or deleted.

4.4 SCALABILITY AND ADAPTABILITY

The architecture is designed to be scalable and adaptable across platforms. It can be integrated into various social media applications and evolves with user behavior and emerging cyber bullying trends.

5. MODULES IMPLEMENTATION

5.1 LIST OF MODULES

- Framework Construction Module
- Read Comments Module
- Classification Module
- Rules Implementation Module
- Alert System Module

5.2 MODULES DESCRIPTION

5.2.1 FRAMEWORK CONSTRUCTION MODULE

Social networking services (SNS) are platforms that allow users to build relationships and interact with others who share similar interests. These platforms provide a medium for users to exchange information, ideas, and personal updates. The framework construction module focuses on designing the graphical

user interface (GUI) for the system. The user interface is designed to enable smooth interactions between users and the system, facilitating actions such as user logins, friend requests, and image sharing. Additionally, the admin interface is designed for managing user data and activities. The GUI allows users to easily interact with the platform and manage their social interactions, ensuring an intuitive and user-friendly experience.

5.2.2 READING COMMENTS MODULE

In modern social media, users engage with content by commenting on posts and sharing thoughts or feedback. The read comments module enables the system to collect and analyze user-generated comments on social media platforms. These comments can vary in format, including text, links, and even short tags. The module is responsible for extracting these comments from users, and it supports different types of comment structures such as uni-grams, bi-grams, and multi-grams. This step involves processing the input data and preparing it for further analysis by the classification module. The system reads the comments in real-time, sending them to the server for further processing and evaluation.

5.2.3 CLASSIFICATION MODULE

The classification module plays a critical role in the proposed system by categorizing user comments based on their content. The goal is to filter out unwanted messages, such as negative or offensive comments. This module uses a back propagation neural network (BPNN) to classify comments as either neutral or non-neutral. It leverages deep learning techniques, particularly the VADER approach, to classify text based on sentiment. The system first identifies neutral sentences and removes them, followed by classifying non-neutral comments into categories of interest. This classification process ensures that only positive or neutral content appears on a user's wall, helping maintain a safe and respectful environment on social media.

5.2.4 RULES IMPLEMENTATION MODULE

The rules implementation module enables users to set filtering criteria for content displayed on their social media walls. This module allows users to create customized rules that define which types of messages should be filtered or blocked. These rules can be based on the user profile attributes, such as age, religious/political views, or work experience. By setting constraints on these attributes, users can fine-tune the filtering process to match their preferences. For example, a user may want to block content from specific profiles or restrict messages from individuals who have previously posted offensive content. This module ensures that content filtering aligns with the user's personal criteria, providing a tailored and controlled social media experience.

5.2.5 ALERT SYSTEM MODULE

The alert system module notifies users whenever offensive or inappropriate content is detected on their wall. When a user receives repeated negative comments, the system triggers an alert based on pre-set threshold values. These thresholds define the number of negative comments a user can post before triggering a block. The alert system also helps users track who is posting offensive content and notifies them via mobile devices when such instances occur. This feature allows users to take immediate action to block or report malicious users. The alert system enhances the user experience by providing timely notifications and ensuring that users stay informed about their social media environment.

SYSTEM ARCHITECTURE

The architecture diagram of the proposed system consists of several key components that work together to provide content-based filtering on social media platforms. At the core of the system, the user interacts with a social networking interface, where they can post comments and content. These comments are then processed through a classification module that uses the VADER sentiment analysis technique to identify whether the content is positive, neutral, or negative. If the comment is classified as negative, it is automatically filtered out or flagged for review. The system also includes a rule-based filtering mechanism that allows users to define custom criteria for blocking certain types of content. Additionally, a blacklist is used to filter out inappropriate words and identify users who repeatedly post negative content.

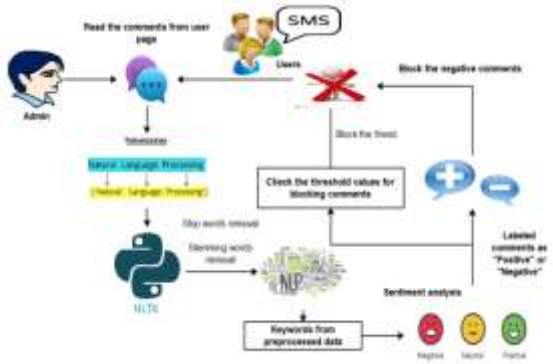

Figure 4.1: System Architecture

EXPIREMENTAL RESULTS

Figure 6.1: Social Media Dashboard

Figure 6.2: Admin Login Page

Figure 6.3: New User Registration Page

Figure 6.4: User Login Page

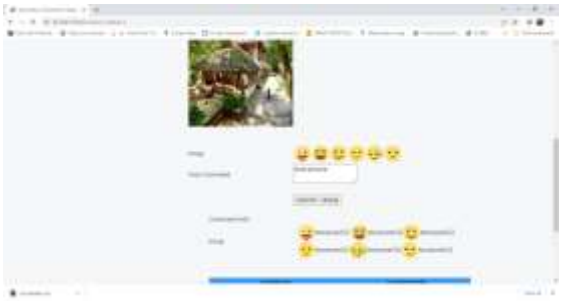

Figure 6.5: Comment Page

Figure 6.6: Second Comment

Figure 6.7: Warning Message

Figure 6.8: Blocked Message

CONCLUSION AND FUTURE ENHANCEMENTS

CONCLUSION

In conclusion, the proposed system effectively addresses the issue of cyberbullying and hate speech on social networking platforms. By integrating machine learning and deep learning techniques, the system automatically classifies and filters offensive content based on sentiment analysis using the VADER approach. The framework allows users to define customized filtering rules, providing them with greater control over the content displayed on their social media walls. The system not only prevents harmful messages but also ensures a positive online environment by blocking malicious users and alerting them when inappropriate content is detected. The system offers significant improvements over existing methods by providing real-time detection and filtering of unwanted content. It also incorporates an alert system that notifies users about negative interactions, which can be crucial for maintaining a respectful and safe digital space. With the ability to block users and filter content based on customizable rules, the proposed system empowers users to protect themselves from cyber bullying and unwanted content while maintaining a secure and engaging social networking experience.

FUTURE ENHANCEMENTS

- **Integration with Image and Video Filtering:** Future versions of the system can be enhanced to detect and filter harmful content in images and videos shared on social networks, expanding beyond text-based content.
- **Multi-Language Support:** The system can be enhanced to support multiple languages, allowing users from different linguistic backgrounds to benefit from hate speech detection and filtering.
- **Advanced User Profiling:** Enhancements can include the ability to analyze user behavior and interactions more deeply, improving the accuracy of the classification and detection of harmful content.
- **Real-Time Speech Analysis:** Integrating speech recognition systems to detect cyber bullying or hate speech in voice messages or live streaming could further improve the effectiveness of the system.

- **Cross-Platform Integration:** Future versions can be adapted to work across multiple social media platforms, providing a uniform filtering solution for different environments.
- **Deep Learning Model Improvement:** The system could utilize more advanced deep learning models, like transformers, for better understanding and classification of nuanced hate speech in diverse contexts.
- **Community Moderation Tools:** Developing additional tools for community-based moderation could allow users to collaborate and contribute to the identification and reporting of harmful content.
- **Enhanced Reporting Mechanism:** Future versions could provide more detailed reports to users, showing specific metrics and trends in offensive content, as well as actions taken against offenders.

REFERENCES

[1] Roy, Pradeep Kumar, et al. "A framework for hate speech detection using deep convolutional neural network." IEEE Access 8 (2020): 204951-204962.

[2] Aluru, Sai Saketh, et al. "Deep learning models for multilingual hate speech detection." arXiv preprint arXiv:2004.06465 (2020).

[3] Mullah, NanlirSallau, and Wan Mohd Nazmee Wan Zainon. "Advances in machine learning algorithms for hate speech detection in social media: a review." IEEE Access 9 (2021): 88364-88376.

[4] Cao, Rui, Roy Ka-Wei Lee, and Tuan-Anh Hoang. "DeepHate: Hate speech detection via multi-faceted text representations." Proceedings of the 12th ACM Conference on Web Science. 2020

[5] Khan, Shakir, et al. "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection." Journal of King Saud University-Computer and Information Sciences 34.7 (2022): 4335-4344.

[6] Mozafari, Marzieh, Reza Farahbakhsh, and Noel Crespi. "A BERT-based transfer learning approach for hate speech detection in online social media." Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8. Springer International Publishing, 2020.

[7] Rabiul Awal, Md, et al. "AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection." arXiv e-prints (2021): arXiv-2103.

[8] Alkomah, Fatimah, and Xiaogang Ma. "A literature review of textual hate speech detection methods and datasets." Information 13.6 (2022): 273.

[9] Malik, Jitendra Singh, Guansong Pang, and Anton van den Hengel. "Deep learning for hate speech detection: a comparative study." arXiv preprint arXiv:2202.09517 (2022).

[10] Khan, Shakir, et al. "HCovBi-caps: hate speech detection using convolutional and Bi-directional gated recurrent unit with Capsule network." IEEE Access 10 (2022): 7881-7894.

[11] Toraman, Cagri, Furkan Şahinuç, and Eyup Halit Yilmaz. "Large-scale hate speech detection with cross-domain transfer." arXiv preprint arXiv:2203.01111 (2022).

[12] Patil, Hrushikesh, Abhishek Velankar, and Raviraj Joshi. "L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models." Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022). 2022.

[13] Gravano, Agustín, et al. "Assessing the Impact of Contextual Information in Hate Speech Detection." IEEE Access, vol. 11, pp. 30575-30590, 2023, doi: 10.1109/ACCESS. 2023.3258973. (2023).

[14] Velankar, Abhishek, Hrushikesh Patil, and Raviraj Joshi. "A review of challenges in machine learning based automated hate speech detection." arXiv preprint arXiv:2209.05294 (2022).

[15] Akuma, Stephen, TyosarLubem, and Isaac Terngu Adom. "Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets." International Journal of Information Technology 14.7 (2022): 3629-3635.