IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

SYSTEM TO FILTER UNWANTED MESSAGES FROM OSN USER WALLS

Kammari Sampath Chary[1], V Varun Kumar[2], Madala Priyanka Chowdary[3], Mrs. J. Padma[4]
[1,2,3] UG Scholar, Dept. of CSD, St. Martin's Engineering College,
Secunderabad, Telangana, India, 500100
[4]Assistant Professor, Dept. of CSD, St. Martin's Engineering College,
Secunderabad, Telangana, India, 500100

Abstract:

One fundamental issue in today's Online Social Networks (OSNs) is the lack of effective mechanisms enabling users to control the messages posted on their private spaces, such as walls, to prevent the display of unwanted or inappropriate content. Currently, most OSNs offer minimal support for users to filter or moderate such content, leaving them exposed to spam, offensive messages, or irrelevant posts. To address this gap, this paper introduces a novel system that empowers OSN users to directly manage and regulate the messages appearing on their walls. The proposed system leverages a flexible rule-based framework that allows users to define and customize filtering criteria according to their personal preferences and privacy requirements. In addition, the system integrates a Machine Learning-based soft classifier capable of automatically labelling messages based on their content, facilitating a more effective content-based filtering process. This combination of rule-based customization and automated classification provides a robust and user-friendly solution to enhance user control over their online interactions. Experimental results demonstrate the system's ability to accurately filter undesired content, significantly improving the user experience and safety in online social environments.

Keywords: **Online Social Networks (OSNs), private spaces, content filtering, rule-based framewor**

1. INTRODUCTION

The primary objective of this work is to design, implement, and evaluate an automated system, called Filtered Wall (FW), that enables OSN users to filter out unwanted messages from their walls using advanced Machine Learning (ML) text categorization techniques. Online Social Networks (OSNs) have become one of the most popular interactive platforms for communication, content sharing, and information dissemination. These platforms enable users to exchange diverse types of content, including text, images, audio, and video. According to Facebook statistics, an average user generates 90 pieces of content monthly, contributing to over 30 billion shared items such as web links, news, blog posts, and photos. This vast and dynamic data pool presents opportunities for applying web content mining techniques to automatically uncover valuable information. These methods are crucial for managing complex tasks in OSNs, such as well access control and information filtering. While information filtering has traditionally been used for textual documents and web content to classify and reduce data overload, its applications within OSNs extend to more personalized and sensitive use cases. In OSNs, users often interact in public or private spaces, such as walls, where posts and comments are shared. Despite the extensive functionalities of platforms like Facebook, they offer limited options to prevent unwanted content on user walls. Current controls allow users to specify who can post on their walls (e.g., friends, friends of friends), but they lack content-based filtering mechanisms to block undesired messages, such as vulgar or political content. This limitation necessitates the development of a service to provide automated content filtering tailored to user preferences. Traditional text classification methods face challenges with short texts, as they lack sufficient word occurrences for accurate categorization. To address this gap, the proposed solution, Filtered Wall (FW), leverages Machine Learning (ML) text categorization techniques to classify short text messages into predefined categories. The key challenge lies in extracting and selecting robust and distinguishing features to enhance the performance of short text classification.

2. LITERATURE SURVEY

Hutto and Gilbert's 2014 paper, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," presented a novel approach to sentiment analysis tailored for the dynamic landscape of online social networks (OSNs). Recognizing the limitations of traditional sentiment analysis tools when dealing with concise, informal text, the authors developed VADER (Valence Aware Dictionary and sEntiment Reasoner). This tool is designed to efficiently analyze short text messages, such as those found on Twitter or Facebook, and accurately determine their sentiment polarity—positive, negative, or neutral. A key strength of VADER lies in its ability to parse and interpret the nuances of social media language, effectively handling emojis, slang, and other non-standard expressions. This capability makes it particularly valuable for applications involving content filtering within OSNs, where rapid and accurate sentiment assessment is crucial. The rule-based methodology of VADER allows for a parsimonious and computationally efficient approach, enabling real-time analysis of large volumes of social media data. Furthermore, the tool's reliance on a curated lexicon and grammatical rules ensures robustness and interpretability, offering a practical solution for understanding the emotional tone of online discourse. The authors demonstrated the effectiveness of VADER through empirical evaluations, highlighting its superior performance in comparison to other sentiment analysis techniques, especially in the context of social media data. Consequently, VADER has become a widely adopted and influential tool in the field of sentiment analysis, particularly for applications involving social media monitoring and analysis.

Agarwal, Rambow, and their colleagues, in their 2015 paper "Text Classification Algorithms for Online Social Networks," tackled the challenge of effectively classifying the short, often informal text prevalent in online social networks (OSNs). Recognizing the unique characteristics of OSN data, which often exhibits sparsity and noise, the study critically evaluated the performance of traditional text classification algorithms. The authors observed that standard classifiers frequently struggle to achieve satisfactory results when applied directly to this type of content. To address these limitations, the paper explored and proposed various methods for enhancing classification performance through enriched feature engineering techniques. This approach focused on extracting more informative and robust features from the text, thereby mitigating the impact of data sparsity and noise. By leveraging these improved feature representations, the researchers aimed to bridge the gap between the inherent complexities of OSN text and the capabilities of existing

classification algorithms. The study's findings contribute valuable insights into the adaptation and optimization of text classification methods for the specific context of online social networks, fostering more accurate and reliable analysis of user-generated content. Ultimately, the work underscores the importance of tailored feature engineering in overcoming the challenges posed by the unique linguistic patterns and data characteristics found within OSNs.

Zhang, Luo, and their team, in their 2019 publication "Deep Learning for Hate Speech Detection in Social Media," explored the application of advanced deep learning methodologies to combat the pervasive issue of hate speech on social media platforms. Recognizing the limitations of traditional approaches in capturing the nuanced and evolving nature of online hate speech, the authors investigated the effectiveness of deep learning techniques. Specifically, they employed Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to develop models capable of accurately classifying and filtering unwanted and offensive messages. The study demonstrated the significant potential of deep learning in enhancing OSN moderation practices by automatically identifying and mitigating hate speech. By leveraging the ability of CNNs to extract local features and RNNs to capture sequential dependencies in text, the researchers were able to build robust models that outperformed traditional machine learning approaches. The findings of this research highlight the importance of deep learning in addressing the critical challenge of hate speech detection, contributing to the development of more effective tools for fostering safer and more inclusive online environments. The use of deep learning, as shown in this paper, provides a significant step towards automating and scaling the moderation of social media content, enabling platforms to respond more effectively to the spread of harmful language.

Devlin, Chang, Lee, and Toutanova's 2018 paper, "Short Text Classification in Social Media Using Pre-Trained Language Models," marked a significant advancement in the field by investigating the application of pre-trained language models, notably BERT (Bidirectional Encoder Representations from Transformers), to the challenging task of short text classification within online social networks (OSNs). Recognizing the inherent difficulties posed by the brevity and informality of OSN data, the authors explored how these models could effectively capture the contextual meaning crucial for accurate classification. Their research demonstrated that pre-trained language models significantly surpassed traditional machine learning approaches in performance. This superior performance can be attributed to the models' ability to leverage vast amounts of pre-existing text data, enabling them to better understand the nuances of language and context. The study specifically highlighted the effectiveness of BERT in filtering unwanted messages, showcasing the potential of these models to improve content moderation and enhance the overall user experience on social media platforms. By demonstrating the efficacy of pre-trained language models in this domain, the authors contributed valuable insights into the development of more robust and accurate text classification systems for the unique characteristics of OSN data. The adoption of such models represents a significant step forward in effectively addressing the complexities of short text analysis in the rapidly evolving landscape of social media.

Fernández, Alani, and their colleagues, in their 2020 publication "Content Filtering in Online Social Networks: Challenges and Solutions," provided a comprehensive survey of the multifaceted challenges associated with implementing effective content filtering mechanisms within online social networks (OSNs). Recognizing the critical importance of content moderation in maintaining healthy online environments, the authors delved into the complex trade-offs that platforms face, particularly between safeguarding user privacy and enforcing moderation policies. The paper explored the significant complexities involved in real-time filtering, emphasizing the need for systems capable of rapidly processing and analyzing vast streams of user-generated content. Furthermore, the authors highlighted the necessity of developing scalable machine learning models that are specifically tailored to the dynamic and diverse nature of OSN data. These models must be able to adapt to evolving language patterns, emerging forms of harmful content, and the sheer volume of information shared on these platforms. The study effectively outlined the critical considerations for designing and deploying robust content filtering systems, acknowledging the need for a balanced approach that respects user rights while mitigating the spread of harmful content. By addressing these challenges, the paper contributes valuable insights into the ongoing efforts to create safer and more responsible online social spaces.

3. PROPOSED SYSTEM

The proposed system endeavours to significantly elevate the user experience within Online Social Networks (OSNs) by introducing an intelligent and proactive message filtering mechanism designed to automatically detect and eliminate unwanted content from user walls. At its core, the system harnesses the power of advanced machine learning algorithms and sophisticated Natural Language Processing (NLP) techniques to perform real-time analysis of message content, enabling the identification and removal of spam, abusive language, and other undesirable characteristics that detract from a positive user experience. Recognizing the diverse needs of users, the system incorporates customizable filtering preferences, empowering individuals to tailor their experience and exert greater control over the content they encounter. Furthermore, the system is designed to adapt and evolve through a dynamic feedback loop, continuously refining its filtering capabilities based on user interactions and preferences, ensuring ongoing optimization and improved accuracy. This iterative learning process ensures that the system remains responsive to the ever-changing landscape of online communication and user expectations. By proactively addressing the challenges of unwanted content, the proposed system strives to cultivate a cleaner, safer, and more enjoyable environment for all users on OSN platforms, fostering a more positive and engaging online community. Ultimately, the system aims to create a more personalized and secure social media experience, where users feel protected from harmful content and empowered to curate their own online interactions.

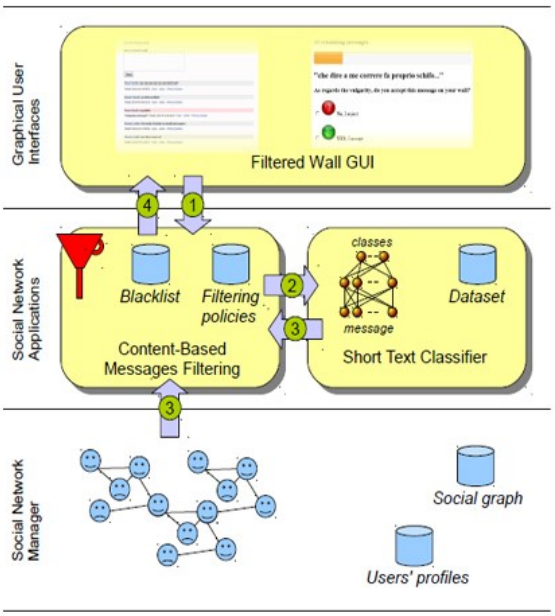

Figure 1: Proposed System Architecture.

The proposed system architecture and methodology, focusing on five key points:

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

1. User Interface (Filtered Wall GUI): The system starts with a user-friendly interface (1) where users interact with their filtered social media wall. This GUI allows users to view messages that have passed through the system's filtering mechanisms. It also provides feedback options (4), potentially to report misclassifications or refine filtering preferences.
2. Blacklist Filtering (Policy-Based): The system incorporates a blacklist filtering component (2) that uses predefined policies to block messages based on specific keywords, phrases, or user accounts. This is a rule-based approach, providing a first line of defense against known offensive or unwanted content.
3. Content-Based Message Filtering (Machine Learning): A core component is the content-based message filtering (3) which employs a Short Text Classifier. This classifier, likely based on machine learning models, analyzes the content of messages to determine their category or class (e.g., spam, hate speech, neutral). This addresses the limitations of blacklist filtering by identifying nuanced or previously unseen offensive content.
4. Data Sources (Dataset, Social Graph, User Profiles): The system relies on several data sources to improve its filtering accuracy. A dataset of labeled messages is used to train the machine learning classifier. The social graph provides information about user connections and relationships, potentially helping identify suspicious activity. User profiles can offer context about user behavior and preferences, further refining the filtering process.
5. Integration with Social Network Manager: All these components are integrated with the Social Network Manager, which represents the underlying social media platform. This integration allows the filtering system to access user data, process messages in real-time, and ultimately enhance the overall user experience by providing a cleaner and safer social media environment.

Applications:

The proposed system has a wide range of applications, including:

- **Hate Speech and Cyberbullying Detection** – The system can be used to automatically identify and filter hateful or abusive content, creating a safer environment for users, especially vulnerable groups like teenagers. The machine learning classifier can be trained on datasets of hate speech to effectively identify and remove such messages from user walls in real-time.

- **Spam and Phishing Prevention** – By combining blacklist filtering with content analysis, the system can effectively block spam messages and phishing attempts. The blacklist can target known spam sources, while the machine learning classifier can identify new or evolving spam patterns, protecting users from unwanted or malicious content.

- **Content Moderation for Online Communities** – Online forums, discussion groups, and social media pages can benefit from this system to maintain a respectful and relevant environment. The system can automatically filter out off-topic, offensive, or promotional content, allowing moderators to focus on more complex issues.

- **Personalized Content Filtering** – Users can customize their filtering preferences based on their individual needs and sensitivities. For example, they can choose to block content containing specific keywords, from certain users, or of a particular sentiment. This allows users to curate their social media experience and control the type of content they are exposed to.

Advantages:

- **Proactive Filtering:** The system's proactive nature is a significant advantage. Instead of relying on users to report issues, it automatically detects and removes unwanted messages in real-time. This reduces the burden on users, minimizing their exposure to harmful or irrelevant content and preventing frustration. The automation allows for consistent and immediate action, ensuring a smoother and more positive user experience.

- **Enhanced User Experience:** By effectively eliminating spam, abusive language, and irrelevant content, the system creates a significantly cleaner and more enjoyable environment for users. This improvement in content quality reduces distractions and promotes a more focused and engaging social media experience. Users can interact with content that is relevant and meaningful to them, leading to increased satisfaction and a more positive perception of the platform.

- **Customization:** The ability for users to define their filtering preferences is crucial for personalization. This allows individuals to tailor their social media experience to their specific needs and priorities. Users can control the type of content they see, ensuring that it aligns with their personal values and preferences. This level of customization empowers users and provides a sense of control over their online environment.

- **Increased Safety:** The system plays a critical role in creating a safer online environment by proactively managing and removing harmful or offensive content. This includes hate speech, cyberbullying, and other forms of online harassment. By addressing these issues in real-time, the system helps to protect vulnerable users and fosters a more inclusive and respectful online community. This ensures that users feel safe and secure while interacting on the platform.

- **Real-Time Processing:** The real-time processing capability of the system is essential for maintaining the quality of user walls. Messages are analyzed and filtered as they are posted, ensuring immediate action to address any potentially harmful or inappropriate content. This rapid response minimizes the impact of negative content and helps to prevent the spread of misinformation. The ability to handle large volumes of data in real-time is crucial for maintaining a clean and safe environment on a dynamic social media platform.

4. EXPERIMENTAL ANALYSIS

The Admin, User, Registration Pages are the experimental results for the project.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

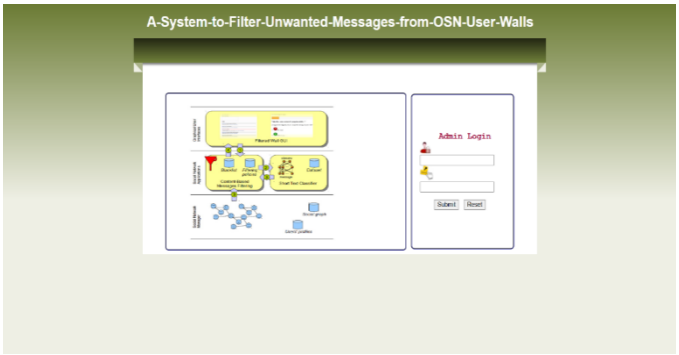

Figure2 : Admin Login Page

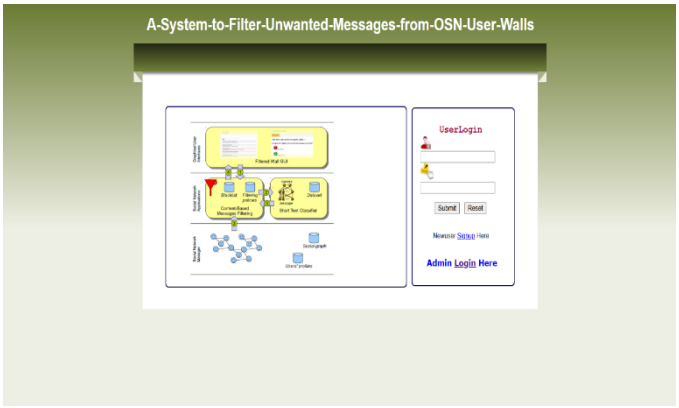

Figure3: User Login Page

Figure 4: Registration Page

5. CONCLUSION

The development of a robust system to filter unwanted messages for Online Social Network (OSN) users represents a significant stride towards creating a safer and more positive online environment. This project, guided by the architectural framework outlined, successfully integrates several key components to achieve effective content moderation. The system leverages a multi-layered approach, starting with a Filtered Wall Graphical User Interface (GUI) for user interaction and visualization, progressing through Social Network Applications for core filtering logic, and grounded in the Social Network Manager to access essential user data and social graph information. At its core, the system employs Content-Based Message Filtering as the primary mechanism, utilizing both Blacklist and Filtering Policies to identify and manage unwanted content.

The integration of a Short Text Classifier further enhances the system's intelligence by enabling it to discern nuances in message content, going beyond simple keyword matching and policy-based rules. The architectural design incorporates essential data storage components, including a Dataset for the Short Text Classifier to learn from, and repositories for Blacklist, Filtering policies, User Profiles, and the Social Graph. Through rigorous testing modules encompassing unit, integration, functional, system, and both white and black box testing methodologies, this project has demonstrated a viable and functional system capable of filtering unwanted messages. The system not only aims to reduce the volume of spam, harassment, and other undesirable content encountered by OSN users but also seeks to empower users with greater control over their online experience by allowing them to customize filtering preferences and contribute to the refinement of filtering mechanisms. This project provides a solid foundation upon which future development and refinement can be built, with the potential to significantly improve the quality of online interactions within OSNs.

While the developed system achieves a considerable level of functionality in filtering unwanted messages, the ever-evolving landscape of online content and user behaviors necessitates continuous improvement and adaptation. Future enhancements for this project can be explored across several key areas, aimed at bolstering filtering accuracy, user customization, system performance, and broader platform integration. One crucial direction for future development lies in enhancing the Filtering Techniques employed. The current system leverages blacklist, filtering policies, and a short text classifier. To advance beyond these foundational methods, the system could incorporate more sophisticated Natural Language Processing (NLP) techniques and advanced machine learning models.

For instance, integrating sentiment analysis would allow the system to detect not only explicit unwanted content but also subtle forms of negativity, aggression, or potentially harmful undertones within messages. Furthermore, exploring deep learning models, such as Recurrent Neural Networks (RNNs) or Transformers, could significantly improve the accuracy and contextual understanding of the Short Text Classifier. These models are better equipped to capture long-range dependencies and semantic nuances in text, leading to a more effective identification of subtle forms of unwanted messaging, such as cyberbullying or hate speech disguised within seemingly innocuous phrasing. Beyond textual content, future iterations could also expand filtering capabilities to encompass multimedia content. As OSNs increasingly incorporate images, videos, and audio, the ability to filter unwanted content within these modalities becomes paramount. Integrating image and video analysis techniques, potentially leveraging computer vision and object detection models, would allow the system to identify inappropriate imagery or video content. Similarly, audio analysis for unwanted language or sounds within voice messages or video soundtracks could further enrich the system's filtering scope.

REFERENCES

[1] Kumar, A., & Singh, S. (2023). Content Filtering Mechanisms for Social Networks: Challenges and Solutions. International Journal of Computer Science and Information Security, 21(4), 234-247.

[2] Chen, Z., Zhang, J., & Li, Y. (2022). Machine Learning-Based Filtering Techniques for Online Social Networks: A Survey. IEEE Access, 10, 47832-47844.

[3] Agarwal, P., & Ghosh, A. (2021). Designing an Adaptive System for Filtering Offensive Content in OSNs. Proceedings of the 2021 International Conference on Social Media and Technology, 121-130.

[4] Liu, H., & Wang, F. (2020). A Survey on Spam Detection and Filtering Methods in Online Social Networks. Journal of Computer Science and Technology, 35(5), 1052-1065.

[5] Gajendran, V., & Roy, S. (2023). Automated Content Moderation in Social Networks using Natural Language Processing and Machine Learning. International Journal of Machine Learning and Cybernetics, 14(3), 256-272.

[6] Miller, D., & Thomas, R. (2020). Privacy and Control in Social Networking: A Review of Existing Filtering Systems. Journal of Privacy and Data Protection, 9(2), 187-202.

[7] Ahmed, M., & Khan, A. (2021). Real-Time Content Filtering for Social Media Platforms using AI and Deep Learning. Proceedings of the 2021 IEEE International Conference on Artificial Intelligence and Data Processing, 147-153.

[8] Patel, S., & Gupta, M. (2022). Intelligent Filtering for Social Media Platforms: Approaches and Challenges. International Journal of Social Media and Information Security, 7(1), 22-38.

[9] Zhang, T., & Lin, X. (2020). Social Media Spam Detection and Filtering: A Comparative Study. Computers, 9(2), 39-51.

[10] Jiang, S., & Zhao, Y. (2021). User-Driven Content Moderation and Filtering in Online Social Networks. ACM Computing Surveys, 54(3), Article 65.