International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Hybrid Moderation Framework for Social Media: Combining AI and Human Expertise

Yayati Nehe, Rushali Sarak, Rutuja Shete, Lata Verma

Yayati Nehe Department of Computer Engineering & International Institute of Information Technology
Rushali Sarak Department of Computer Engineering & International Institute of Information Technology
Rutuja Shete Department of Computer Engineering & International Institute of Information Technology
Lata Verma Department of Computer Engineering & International Institute of Information Technology

---

Abstract - Social media platforms are critical spaces for global communication, allowing users to freely share opinions, engage in discussions, and connect with communities worldwide. However, these platforms are also hosts for harmful content, including hate speech, misinformation, cyberbullying, and offensive imagery. To address these challenges, many platforms implement content moderation systems. This paper provides a comprehensive survey of existing content moderation strategies and proposes a hybrid content moderation system. The hybrid system leverages Natural Language Processing (NLP) for text analysis, Convolutional Neural Networks (CNN) for image detection, and a machine learning model to make final decisions. In cases where the model cannot make a confident decision, human-in-the-loop moderation is invoked. The human decision is final, ensuring fair and transparent content moderation. This paper also explores how each component contributes to improved accuracy and efficiency in detecting harmful content and addresses gaps in handling low-resource languages and multimodal content.

***Key Words*:** Social media moderation, hybrid system, NLP, CNN, human-in-the-loop, machine learning, social media.

1. INTRODUCTION

Social media platforms such as Facebook, Twitter, Reddit, and Instagram have revolutionized the way people communicate, share ideas, and interact online. These platforms offer an unprecedented level of global connectivity, enabling individuals to disseminate information, opinions, and personal content in real-time to vast audiences. Social media has become deeply embedded in modern life, influencing public discourse, shaping cultural trends, and impacting politics and global events. However, the very openness and accessibility that define these platforms have also created significant challenges. Harmful content, including hate speech, misinformation, abusive language, and graphic violence, can spread rapidly across networks, posing serious risks to both individuals and society at large. The spread of harmful content on social media is a complex issue. Hate speech can fuel discrimination and violence, misinformation can lead to real-world consequences such as public health crises, and abusive language can inflict psychological harm and foster hostile environments. Despite efforts by social media companies to implement policies aimed at protecting users, the sheer volume of content uploaded daily makes it difficult to manually moderate posts.

Millions of text, image, and video posts are shared each day, creating a massive stream of information that is impossible to filter entirely through human moderation alone. While much of this content is benign, a significant portion may violate platform guidelines or even laws, necessitating robust moderation strategies. One of the most pressing challenges for existing moderation systems is the multimodal nature of content on social media. Posts often include not just text but also images, videos, and even live streams, requiring systems capable of processing different types of media. Additionally, content moderation systems often struggle to detect harmful content in low-resource languages, where limited training data and complex linguistic nuances present significant barriers for automated models. Another difficulty arises from the dynamic nature of harmful content, which continuously evolves in form and language. New types of harmful content often emerge faster than the automated systems can be updated, making it difficult to consistently enforce platform policies.

Despite the increasing reliance on automated moderation, human moderators remain crucial to ensure fairness and context-based judgment. Their ability to recognize

subtleties in communication and the context in which content is shared—whether in understanding cultural references, slang, or emerging trends in harmful speech—means that humans are still needed to address content that algorithms might inaccurately flag or overlook. Nonetheless, the heavy reliance on human moderators raises concerns about the psychological toll of continuously reviewing disturbing content, as well as the sheer volume of content that exceeds human capacity to manage.

2. Body of Paper

I. RELATED WORK

A) SoK: Content Moderation in Social Media, from Guidelines to Enforcement and Research to Practice" provides a comprehensive review of content moderation practices across major social media platforms. The authors study the community guidelines and moderation policies of 14 platforms, identifying inconsistencies in how different platforms enforce these guidelines. They categorize moderation practices into two primary forms: hard moderation, which involves actions such as removing content or banning accounts, and soft moderation, which includes strategies like applying warning labels or limiting the visibility of content. This distinction highlights the varying levels of enforcement that platforms employ, often reflecting different interpretations of content safety and user autonomy.

B) "Taxonomy of Content Moderation," the authors propose a classification system to organize the types of content moderated, the comprehensiveness of platform guidelines, and the enforcement methods used. Their analysis spans over 200 interdisciplinary research papers and highlights significant variation in how content is moderated across platforms, particularly when comparing mainstream platforms like Facebook and Twitter to fringe platforms like Parler and Gab. Mainstream platforms tend to employ more stringent moderation techniques aimed at curbing harmful content, while fringe platforms often promote minimal restrictions and prioritize user freedom and autonomy.This disparity underscores the ongoing debate about the balance between free expression and the need to maintain safe online spaces.

C) "User-Centric Approaches to Hate Speech Detection" While much of the existing research focuses on platform-driven moderation methods, a more user-centric approach is explored in the paper "UserCentric Modeling of Online Hate Through the Lens of Psycholinguistic Patterns and Behaviors in Social Media." This study shifts the focus from content moderation to analyzing the behaviours and psycholinguistic traits of users who are likely to engage in hate speech. By examining a dataset of 5.4 million tweets related to antiAsian hate during the COVID-19 pandemic, the authors identify specific behavioral and linguistic patterns that can predict future hate speech. This perspective offers a complementary approach to traditional content-focused moderation by emphasizing the role of user characteristics in the spread of harmful content.

D) "Hate Speech Prediction Model" Using psycholinguistic and behavioural features, the authors develop a model designed to predict which users are likely to engage in hate speech in the future. This user-centric approach shifts the focus from solely monitoring content to identifying users who exhibit traits that make them more prone to harmful behaviour. By examining patterns such as emotional expression, vocabulary use, and social engagement, the model provides early detection of users who may post offensive content. This proactive strategy allows platforms to intervene before harmful content is shared, helping to mitigate the spread of hate speech. The model performs consistently well across different topics and time periods, demonstrating its adaptability and robustness in various contexts.

E) "Predicting cyberbullying on social media" in the big data era has become an essential focus for researchers due to the growing volume of online interactions.

Machine learning algorithms are increasingly being utilized to address this issue, leveraging techniques like natural language processing (NLP) to detect harmful content. The literature highlights the use of supervised learning algorithms such as SVM, Naive Bayes, and deep learning models like CNN and LSTM, which have shown promising results in identifying abusive language. However, open challenges remain, including the complexity of understanding context, handling multilingual data, and ensuring real-time detection across large datasets. Moreover, ethical concerns related to privacy and bias in algorithmic decisions call for further refinement and transparency in the models.

F) "MSCMGTB (Multimodal Social Media Content Moderation using Hybrid Graph Theory and Bio-Inspired Optimization)" introduces an innovative approach to managing harmful content across social media platforms. The method integrates multimodal data, such as text, images, and videos, to provide a more comprehensive content analysis. By employing hybrid graph theory, the system maps relationships between different types of content and user interactions, enhancing the detection of harmful behaviors like cyberbullying and hate speech. Bio-inspired optimization techniques, such as ant colony optimization or genetic algorithms, are used to refine the content moderation process, enabling efficient and scalable solutions. This approach aims to address the limitations of current moderation methods, providing faster, more accurate content filtering across large datasets.

II. EXISTING MODERATION FRAMEWORKS

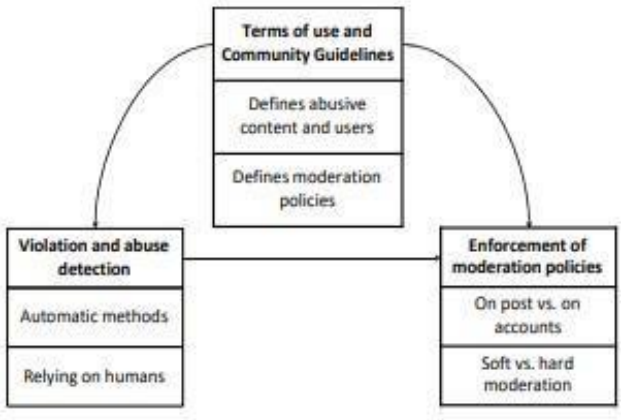

Fig 3.1 Content moderation framework

Source: [1]

Content moderation systems on social media platforms generally consist of three main components: (1) the establishment of community guidelines and terms of service, (2) the detection of harmful content, and (3) the enforcement of moderation policies.

Community Guidelines and Terms of Service: Each social media platform sets out its own community guidelines and terms of service that define what types of content are allowed and what behaviours are prohibited. These rules cover a wide range of harmful content, including hate speech, harassment, misinformation, and explicit imagery. However, these guidelines can vary significantly between platforms, leading to inconsistencies in enforcement. Some platforms, such as Facebook, enforce strict guidelines, while others, such as Parler, have more lenient policies.

The existing systems for content moderation across social media platforms, as outlined in the paper, rely on a combination of human moderators and automated algorithms (including machine learning models). These systems function at different levels of moderation, such as:

Human-based moderation: human moderators, including paid workers or volunteers, are responsible for identifying and removing abusive content. Platforms like facebook and twitter employ large groups of human moderators, often freelancers, to enforce moderation policies. Some platforms like reddit also rely on volunteers, who are active users of the community, to create and enforce local rules. Human moderation is time-consuming, labor-intensive, and emotionally taxing for the moderators, especially when dealing with sensitive or disturbing content.

Algorithmic moderation: to reduce the emotional and financial costs of human moderation, social media platforms deploy automated algorithms. These include rule based systems and machine learning models. For instance, youtube uses automated flagging systems to detect offensive comments, while facebook employs pre-trained language models like bert and roberta to detect hate speech. Perceptual hashing techniques are also used to detect inappropriate media such as pornography. Automated systems can process large volumes of content efficiently, but they often struggle with context and may produce false positives or overlook nuanced violations.

Human-in-the-loop moderation: some platforms combine both human and automated moderation through a human-in-the-loop system, where human feedback is

used to improve the performance of ai models. This approach addresses the limitations of ai models, especially in cases where context or cultural sensitivities are important. Additionally, regulations like the eu gdpr mandate that certain decisions, if challenged, must undergo human review. Enforcement mechanisms: enforcement of moderation policies can take the form of hard or soft moderation.

Hard moderation involves the removal of content or accounts, such as banning users or taking down posts. Platforms like facebook, instagram, and twitter have strict policies for hard moderation, including account suspensions or permanent bans for serious violations. **Soft moderation** uses more subtle methods, such as applying warning labels, fact-checking posts, or quarantining certain content. For example, twitter adds warning labels to posts with unverified content, and reddit quarantines communities that violate policies.

ADDRESSING IDENTIFIED GAPS :

Multimodal Content Detection: Current models often struggle with detecting harmful content that combines text and images, such as memes. By integrating CNNs with NLP models, the hybrid system ensures that multimodal content is accurately analyzed. This addresses the gap in detecting offensive memes or complex visual-text combinations.

Cross-Lingual Capability for Low-Resource Languages: Most existing content moderation models are designed for high-resource languages like English. Our system leverages transfer learning techniques to adapt models for lowresource languages, ensuring that harmful content in languages like Pashto can be detected effectively.

Human-in-the-Loop Oversight: While AI models are efficient at processing large volumes of content, they may miss nuanced cases. The human-in-the-loop component ensures that complex or borderline content is reviewed by human moderators, reducing the risk of false positives or negatives. This improves the overall fairness and transparency of the system

III. PROPOSED HYBRID MODERATION SYSTEM

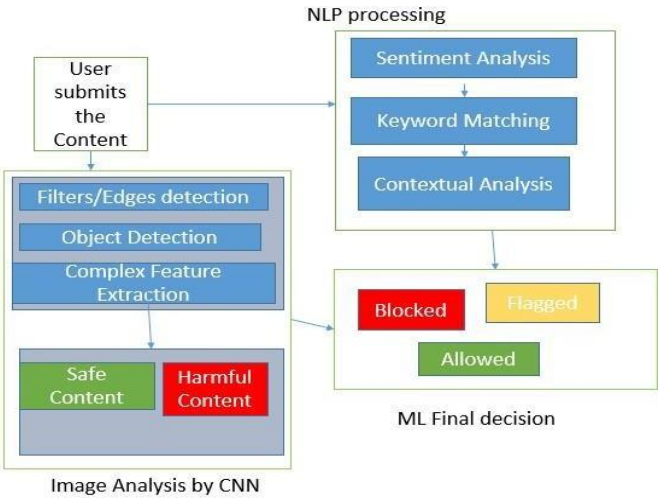

Fig 4.1 Hybrid Content Moderation Workflow Using CNN and NLP

To overcome the limitations of current moderation systems, we propose a hybrid content moderation system that integrates text analysis (NLP), image detection (CNNs), machine learning for decision-making, and human-in-the-loop oversight. Our project, based on the system architecture presented, offers a framework for integrating image analysis using Convolutional Neural Networks (CNN) with Natural Language Processing (NLP) to address multimodal content moderation. As illustrated in the diagram, user-submitted content undergoes image analysis through processes like filters/edges detection, object detection, and complex feature extraction to classify content as harmful or safe. In parallel, text-based content is processed through NLP techniques such as sentiment analysis, keyword matching, and contextual analysis to assess its risk. Both streams—image and text—converge in a final decision-making phase where the content is classified as blocked, flagged, or allowed based on the system's predefined moderation rules. This architecture enables scalable, realtime moderation while ensuring that both visual and textual cues are considered in the moderation process.

Text Analysis Using NLP

NLP models will be employed to detect harmful content in text posts, comments, and messages. Models like BERT and RoBERTa are pre-trained on large corpora and fine-tuned for detecting specific types of harmful language. These models can classify text based on semantics and context, enabling them to identify hate speech, misinformation, and abusive language with high accuracy.

To address the challenge of moderating lowresource languages, we will leverage transfer learning techniques. Pre-trained models will be fine-tuned using smaller datasets in languages like Pashto, ensuring that the system can effectively moderate multilingual content.

Image Detection Using CNNs

For content containing images, CNNs will be used to detect harmful visual elements. CNNs can recognize offensive imagery, symbols, and text embedded within images (e.g., memes). The model will be trained on datasets containing examples of harmful images, allowing it to classify whether the content violates platform guidelines.

Multimodal content (e.g., memes) requires both text and image analysis, and the hybrid system will integrate CNNs with NLP models to analyze both the visual and textual elements of a post. This ensures that offensive content, even when it appears in a combined format, is accurately detected.

Machine Learning for Decision-Making

Once the NLP and CNN models have analyzed the content, the outputs are combined into a machine learning decision model. This model takes the results of the analysis (e.g., whether the text is harmful, whether the image contains offensive elements) and makes a final decision on how the content should be moderated. The model is trained on large datasets of labeled content, allowing it to make decisions based on learned patterns across different types of content.

Human-in-the-Loop Moderation

The final component of the hybrid system is human-in-the-loop moderation. When the machine learning model cannot confidently classify content, the system escalates the case to human moderators. Human moderators bring contextual understanding and the ability to interpret nuanced content that may be difficult for AI models to process, such as sarcasm, cultural references, or evolving trends.

The human-in-the-loop approach also serves as a fail-safe, ensuring that content is not incorrectly flagged or removed due to limitations in the AI models. Human moderators' decisions are final, ensuring fairness and transparency in the moderation process

3. CONCLUSION

In this paper, we proposed a hybrid content moderation system that integrates advanced technologies such as natural language processing (NLP), convolutional neural networks (CNNs), and machine learning with human oversight to create a robust and scalable solution for moderating harmful content on social media platforms. By addressing the challenges of multimodal content, cross-lingual detection, and human interpretation, the system enhances the accuracy and effectiveness of content moderation. This ensures harmful content is identified and managed appropriately while maintaining fairness and transparency.

Our hybrid approach helps mitigate some of the limitations of fully automated systems, which may misclassify or overlook nuanced content, such as sarcasm or cultural differences. Additionally, the involvement of human moderators ensures that decisions—especially for borderline or incorrectly flagged content—are final and are made with a deeper understanding of context, ensuring fairness in the moderation process.

While the proposed system presents significant improvements, there are areas that warrant further exploration. Future work could focus on enhancing the system's AI models, improving their ability to understand complex contexts, including sarcasm, cultural subtleties, and evolving social media trends. Furthermore, expanding the system's cross-lingual detection capabilities and integrating community-driven moderation could offer additional layers of adaptability and fairness.

In conclusion, this hybrid system is a step toward more effective and transparent content moderation on social media platforms, providing a balanced approach that combines the speed of AI with the nuanced judgment of human moderators. As the system evolves and incorporates more advanced techniques, it will continue to provide a scalable and efficient solution for managing harmful content in an increasingly digital world.

ACKNOWLEDGEMENT

We extend our heartfelt gratitude to all individuals and institutions who supported and contributed to the development of this research on the **"Hybrid Moderation Framework for Social Media: Combining AI and Human Expertise."**
We are profoundly thankful to our academic institution, International Institute Of Information Technology, for

providing the resources and a conducive environment to carry out this research. We acknowledge the guidance of our mentors and professors, whose insights and expertise have been invaluable throughout this journey.

Special thanks are due to the human moderators and AI practitioners who provided us with practical insights and feedback, helping shape a balanced perspective for this framework. We also thank our peers and collaborators for their constructive discussions and support in refining our approach.

Finally, we express our gratitude to our families and friends for their constant encouragement and understanding, which motivated us to complete this work.

REFERENCES

[1] SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice.

[2] User-Centric Modeling of Online Hate Through the Lens of Psycholinguistic Patterns and Behaviors in Social Media.

[3] Social Media Forensics: An Adaptive Cyberbullying-Related Hate Speech Detection Approach Based on Neural Networks With UncertaintyBlockchain in Humanitarian Operations Management: A Review of Research and Practice. (2022).

[4] Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges

[5] MSCMGTB: A Novel Approach for Multimodal Social Media Content Moderation Using Hybrid Graph Theory and Bio-Inspired Optimization

[6] Comparing the Impact of Social Media Regulations on News Consumption

[7] Marsh v. Alabama, 1946.

[8] 47 U.S. Code §230 - Protection for Private Blocking and Screening of Offensive Material. https://www.law.cornell.edu/uscode/text/4 7/230, 1996.

[9] Zeran v. America Online, Inc., 1997.

[10] Contest, Sweepstakes, and Giveaway Guidelines. https://www.tumblr.com/policy/en/contestguidelines, 2012.

[11] The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. https://www.wired.com/2014/10/contentmoderation/, 2014.

[12] General Data Protection Regulation (GDPR). https://gdpr-info.eu/art-22-gdpr/, 2016.

[13] Recital 71 - Profiling. https://gdprinfo.eu/recitals/no-71/, 2016.

[14] Automation Rules. https://help.twitter.com/en/rules-andpolicies/twitter-automation, 2017.

[15] Facebook, Microsoft, Twitter, and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism. https://perma.cc/6QYS-ML76, 2017.

[16] 2018 Was the Year We (Sort Of) Cleaned Up the Internet. https://mashable.com/article/deplatforming - alex-jones-2018, 2018.

[17] Deplatforming Works. https://www.vice.com/en/article/bjbp9d/do -social-media-bans-work, 2018.

[18] Gab, the Social Media Site for the AltRight, Gets Deplatformed. https://nymag.com/intelligencer/2018/10/g ab-the-alt-right-social-media-site-getsbanned.html, 2018.

[19] Inside Facebook's Fast-Growing Content-Moderation Effort. https://www.theatlantic.com/technology/ar chive/2018/02/what-facebook-toldinsiders-about-how-it-moderatesposts/552632/, 2018.

[20] Facebook Bans White Nationalism and White Separatism. https://www.vice.com/en/article/nexpbx/fa cebook-bans-white-nationalism-and-whiteseparatism, 2019.

[21] Facebook While Black: Users Call It Getting 'Zucked,' Say Talking About Racism Is Censored as Hate Speech. https://www.usatoday.com/story/news/201 9/04/24/facebook-while-black-zuckedusers-say-they-get-blocked-racismdiscussion/2859593002/, 2019. [

22] GIFCT. https://perma.cc/44V5-554U, 2019.

[23] Here's What Happens When News Comes With a Nutrition Label. https://www.wired.com/story/gallup-pollfake-news-ratings/, 2019