Design and Implementation of an AI-Powered Social Media Platform for Intelligent Content Moderation and User Engagement

Dr. Parag Thakare[1], Mr. Utkarsh Upadhyay[2], Miss. Madhavi Bansod[3], Mr. Paras Salve[4], Mr. Kartik Bhadane[5], Mr. Rahul Madavi[6] and Mr. Aditya Manwar[7]
[1]*Head of Department, Department of Computer Engineering, Jagadambha College of Engineering and Technology, Yavatmal, India*
[2,3,4,5,6,7]*B. E Student, Department of Computer Engineering, Jagadambha College of Engineering and Technology, Yavatmal, India*

Abstract—Online social media platforms have significantly changed the way people communicate, exchange ideas, and access information. However, the continuous increase in user-generated content has introduced challenges such as misinformation, abusive language, and spam activities. To address these issues, this paper presents the design and implementation of an AI-powered social media platform that automates content moderation while improving user engagement. The proposed system applies natural language processing and machine learning techniques for sentiment classification, fake news detection, and spam identification. In addition, a personalized content recommendation mechanism is incorporated to enhance user experience based on interaction patterns and preferences. Experimental evaluation shows that the system improves moderation accuracy, reduces manual effort, and increases the overall reliability of the platform.

Index Terms—Artificial Intelligence, Content Moderation, Fake News Detection, Machine Learning, Social Media Platform, User Engagement.

I. INTRODUCTION

Online social networking systems have reshaped digital communication by enabling large-scale information sharing and online interaction among users. Millions of users actively share opinions, news, and multimedia content every day, making these platforms an essential part of modern digital life. However, the continuous rise in user activity has also introduced serious concerns, including the circulation of false information, abusive language, cyberbullying, and automated spam content.

In earlier stages, content moderation relied mainly on human reviewers and predefined filtering rules. Although these methods provided basic control, they are no longer effective for handling large-scale, real-time data. Manual moderation requires significant time and resources, while static rules fail to adapt to evolving content patterns. Consequently, harmful or misleading posts often spread rapidly before corrective actions can be applied.

Recent advancements in artificial intelligence and machine learning have enabled automated analysis of large text datasets with improved accuracy and speed. Natural language processing techniques allow systems to understand contextual meaning, emotional tone, and authenticity of textual content. By learning from historical data and user behavior, AI-based models can continuously improve moderation decisions and content recommendations.

This research presents an AI-powered social media platform designed to automate content moderation and improve user engagement. The system integrates sentiment analysis, fake news detection, spam identification, and personalized content recommendation within a unified architecture. The primary goal of this work is to enhance platform safety, reduce manual moderation effort, and deliver a more trustworthy and engaging experience for users.

II. PROCEDURE FOR PAPER SUBMISSION

A. Review Stage

In the initial stage, the system requirements were analyzed by studying existing social media platforms and their limitations. A detailed review of current content moderation techniques and AI-based solutions was carried out to identify suitable approaches for sentiment analysis, fake news detection, and spam identification. Based on this analysis, the system architecture and workflow were designed to support intelligent and automated moderation.

B. Final Stage

In the final stage, the proposed system was implemented by integrating machine learning models and natural language processing techniques. User-generated content is processed through AI modules to classify sentiment, detect misinformation, and identify spam activities. The results obtained from these modules are used by the content moderation engine to make publishing decisions. The system was tested using sample datasets to evaluate accuracy and performance.

C. Figures

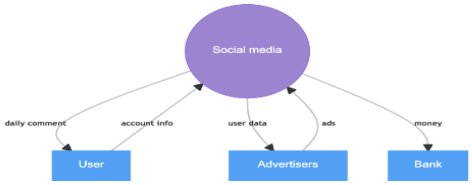

Fig. 1. Basic interaction model of the social media platform

III. METHODOLOGY

The methodology describes the systematic process followed to implement intelligent content moderation and personalized user engagement within the proposed AI-powered social media platform. The overall workflow is designed to ensure efficient handling of user-generated content through automated analysis and decision-making mechanisms.

The process begins with user registration and authentication through the platform interface. Once authenticated, users are allowed to create and submit text-based posts. These posts are forwarded to the content processing module, where preliminary text preprocessing operations are performed. This stage includes cleaning the text data by removing unnecessary symbols, normalizing words, eliminating stop words, and preparing the content for further analysis.

After preprocessing, the refined content is passed to the AI analysis module. This module applies machine learning models and natural language processing techniques to evaluate the submitted content. Sentiment analysis is performed to determine the emotional tone of the post, categorizing it as positive, negative, or neutral. Simultaneously, fake news detection algorithms assess the credibility of the information, while spam detection mechanisms identify suspicious or repetitive content based on learned patterns.

Based on the outcomes of these analyses, the content moderation engine makes an automated decision. Content that meets platform guidelines is approved and stored in the database for public visibility. Content identified as harmful or suspicious is either restricted or flagged for further review. Flagged posts are forwarded to the administrative interface, where moderators can examine the content and take appropriate action.

Approved content is further processed by the recommendation module, which prioritizes posts according to user interests, interaction history, and engagement behavior. All moderation actions and system decisions are recorded in moderation logs, enabling performance evaluation and future system improvements. This structured methodology ensures scalability, accuracy, and real-time responsiveness in managing social media content.

IV. SYSTEM PARAMETERS

The system parameters describe the hardware and software requirements used for the development and implementation of the proposed AI-powered social media platform. Since the system is developed using the MERN stack and also supports a mobile application version, the parameters are defined to ensure compatibility across both web and mobile platforms.

A. Hardware Requirements for Development and Deployment:

- Processor: Intel Core i5 or higher

- RAM: Minimum 8 GB (Recommended: 16 GB)
- Storage: Minimum 256 GB SSD
- Input Devices: Keyboard and Mouse
- Output Device: Monitor

For Mobile Application Testing:
- Android Smartphone
- Processor: Octa-core or higher
- RAM: Minimum 4 GB
- Operating System: Android 9.0 or above

B. Software Requirements Web Application:
- Operating System: Windows 10 / Linux / macOS
- Frontend: React.js
- Backend: Node.js with Express.js
- Database: MongoDB
- Programming Languages: JavaScript, HTML, CSS
- API Communication: RESTful APIs

Mobile Application:
- Mobile Framework: React Native
- Development Environment: Android Studio
- Mobile Operating System: Android

AI and Machine Learning:
- Programming Language: Python
- Libraries: Scikit-learn, TensorFlow
- NLP Tools: NLTK

Development Tools:
- Code Editor: Visual Studio Code
- Version Control: Git

C. Dataset Parameters
- Type of Data: Text-based social media posts
- Dataset Size: Approximately 5,000–10,000 records
- Data Format: JSON and CSV
- Data Labels: Positive, Negative, Neutral, Fake, Genuine, Spam

D. Model Parameters
- Machine Learning Algorithms: Naïve Bayes, Support Vector Machine
- Feature Extraction Technique: TF-IDF
- Training and Testing Ratio: 80:20
- Evaluation Metrics: Accuracy, Precision, Recall, F1-Score

V. HELPFUL HINTS

A. System Architecture Description

The architecture of the proposed AI-powered social media platform is designed to support automated content moderation and personalized content delivery in a scalable and efficient manner. The system is organized into independent yet interconnected modules, each responsible for a specific functionality within the platform.

The User Interface module provides facilities for user registration, authentication, content creation, and interaction. All user-generated posts are forwarded to the Content Processing module, where initial text preparation and formatting are performed to ensure compatibility with analytical models.

Processed content is then transmitted to the AI Analysis module, which applies machine learning and natural language processing techniques to examine the content. This module performs sentiment classification, credibility assessment for fake news detection, and identification of spam or malicious activity. The results generated by this analysis are passed to the Content Moderation Engine.

The Content Moderation Engine evaluates the analysis output and determines whether the content should be approved, restricted, or flagged for review. Approved posts are stored in the database and made available to users through the Recommendation System. Content marked as suspicious is redirected to the Admin Panel, where moderators can monitor activity and take appropriate action.

This modular architecture enables efficient data handling, improves system scalability, and allows future enhancements such as multimedia content analysis and multilingual support. Fig. 1 illustrates the basic interaction model of the proposed social media platform.

B. Data Flow Diagram Description

The Data Flow Diagram illustrates how information moves through the AI-powered social media platform. At the highest level, the system interacts with two primary external entities: users and administrators. Users provide input in the form of registration details, login credentials, and post content, while administrators receive moderation reports and flagged content notifications.

At the detailed level, user input first passes through the authentication process before content submission. Submitted content flows to the preprocessing and AI analysis modules, where it is evaluated for sentiment, authenticity, and spam characteristics. Based on the moderation decision, approved content is forwarded to the recommendation process, while flagged content is stored in moderation logs and sent to the administrative interface.

The structured flow of data ensures that moderation decisions are performed efficiently and that administrators maintain control over platform activity. This clear separation of processes enhances transparency, system reliability, and ease of maintenance.

C. Design Considerations

While designing the proposed system, several important considerations were taken into account. The system is designed to be scalable to handle a large number of users and posts. Security measures are incorporated to protect user data and ensure privacy. The AI models are selected to balance accuracy and computational efficiency. The modular design allows easy updates and future enhancements, such as support for multimedia content moderation and multilingual analysis.

VI. PUBLICATION PRINCIPLES

A. System Performance

The performance of the proposed AI-powered social media platform was evaluated based on its ability to analyze and moderate user-generated content efficiently. The system processes textual posts in real time using machine learning and natural language processing techniques. The modular design enables smooth interaction between content preprocessing, AI analysis, moderation, and recommendation components, resulting in stable and scalable system operation even under increased user activity.

B. Results

The experimental evaluation demonstrates that the proposed system effectively performs sentiment classification, fake news identification, and spam detection. The sentiment analysis module accurately categorized posts into positive, negative, and neutral classes. The fake news detection component successfully distinguished between credible and misleading information, while the spam detection module identified repetitive and automated content with high reliability. Additionally, the recommendation mechanism improved content relevance, leading to increased user interaction and engagement.

C. Discussion

The obtained results indicate that AI-based content moderation provides significant advantages over traditional manual and rule-based approaches. Automated analysis reduces response time and minimizes human effort, allowing harmful or misleading content to be addressed more promptly. The use of machine learning models enables the system to adapt to evolving content patterns and user behavior, improving moderation accuracy over time. The integrated recommendation system further enhances the overall user experience by delivering personalized content.

D. Observations

During testing, the system consistently demonstrated reliable moderation performance across different content types. An increase in user activity did not significantly impact processing speed or accuracy. The administrative interface allowed efficient monitoring and review of flagged content. These observations highlight the effectiveness of the system design and its suitability for large-scale social media environments.

E. Comparison

When compared with conventional content moderation techniques, the proposed platform offers improved efficiency, consistency, and scalability. Manual moderation methods are resource-intensive and susceptible to human bias, whereas the AI-driven approach provides automated and objective decision-making. Unlike static rule-based systems, the proposed solution continuously learns from data, making it more adaptable to modern social media challenges.

VII. CONCLUSION

This paper presented the design and implementation of an AI-powered social media platform focused on intelligent content moderation and enhanced user engagement. The system addresses critical challenges

such as misinformation, toxic language, and spam by integrating sentiment analysis, fake news detection, and spam identification within a unified framework.

The experimental results confirm that the proposed approach improves moderation accuracy, reduces reliance on manual intervention, and enhances platform reliability. The inclusion of a personalized recommendation mechanism further contributes to a positive user experience. Overall, the proposed system demonstrates the potential of artificial intelligence in creating safer, more trustworthy, and engaging social media platforms. Future work may extend this approach to support multimedia content analysis and multilingual moderation.

VIII. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Computer Engineering, Jagadambha College of Engineering and Technology, Yavatmal, for providing the necessary facilities and support to carry out this research work. We are thankful to our project guide and faculty members for their continuous guidance, encouragement, and valuable suggestions throughout the development of this research paper.

We also extend our appreciation to all those who directly or indirectly contributed to the successful completion of this work. The support and cooperation received during the course of this project are gratefully acknowledged.

REFERENCES

[1] Sharma A, Verma R. Artificial intelligence techniques for social media content moderation. Int J Computer Sci Eng. 2020;12(4):215–222.

[2] Kumar S, Patel D. Machine learning approaches for fake news detection on social media platforms. J Inf Secure Appl. 2021; 58:102720.

[3] Brown J, Gupta S. Sentiment analysis of social media data using natural language processing. IEEE Access. 2020; 8:12345–12354.

[4] Verma P, Singh R. Spam and bot detection techniques in online social networks. Int J Eng Res Technol. 2019;8(6):410–415.

[5] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.

[6] Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media: A data mining perspective. SIGKDD Explorer. 2017;19(1):22–36.

[7] Russell S, Norvig P. Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River (NJ): Pearson Education; 2010.

[8] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge (MA): MIT Press; 2016.

[9] Hasan M, Orgun MA, Schwitter R. Real-time detection of fake news on social media platforms. Proc IEEE Adv Computer Conf. 2018;1–6.

[10] Wang Y, Ma F, Jin Z. EANN: Event adversarial neural networks for multi-modal fake news detection. Proc ACM SIGKDD. 2018;849–857.

[11] Zhang Y, Zhao J. Content recommendation techniques in social media platforms. Int J Adv Computer Sci Appl. 2019;10(5):321–327.

[12] Ahmed A, Rahman M. Privacy and security challenges in social networking sites. J Network Compuert Appl. 2018; 112:84–96.