Semantic Conversational Content Moderation

Ravinarayana B[1], Ansar Ul Haq[2], Anush Shetty[3], Ankush Shenoy[4], Deepansh[5]

Computer Science & Engineering, MITE, Moodabidre, India[1]

Student, Computer Science & Engineering, MITE, Moodabidre, India[2]

Student, Computer Science & Engineering, MITE, Moodabidre, India[3]

Student, Computer Science & Engineering, MITE, Moodabidre, India[4]

Student, Computer Science & Engineering, MITE, Moodabidre, India[5]

Abstract - In today's digital environment, social media platforms have become an integral part of our daily lives, facilitating global communication, knowledge sharing and community building. However, these platforms are increasingly vulnerable to the spread of offensive and toxic content, including misinformation, harassment and hate speech. Such content poses a significant threat to the safety and well-being of Internet users. In response to this immediate problem, we set out to develop an AI-based conversation moderation service aimed at effectively detecting and removing offensive or semantically toxic information in real time. Our solutions strive to make the internet safer and more user-friendly, thereby promoting a more positive and inclusive online environment.

***Key Words*:** Natural Language Processing(NLP), Moderation, Semantic Analysis, Mistral.

1. INTRODUCTION

In today's connected world where social media platforms play a central role in communication, ensuring a safe and inclusive online environment is of utmost importance. This project is an innovative AI-driven conversation moderation service carefully crafted to quickly identify and filter out semantically toxic or offensive content in real time. Utilizing state-of-the-art natural language processing (NLP) and machine learning techniques, the service offers a robust solution to combat online toxicity, promoting a safe and welcoming digital space for users. It works seamlessly across multiple social media platforms and provides continuous monitoring and analytics that leverage advanced NLP models capable of understanding linguistic nuances, context and user intent. Thanks to its ability to distinguish harmless conversations from malicious content, the system effectively flags, warns or removes offensive messages and images, thereby stopping the spread of malicious content and minimizing its emotional impact on users. In the ongoing fight against online hate and toxic behavior, this AI-based moderation service is a beacon of promise, constantly improving the user experience, promoting inclusivity and protecting against harm. This introduction sets the stage to delve into the architecture, training methodologies, and ethical considerations underlying this transformative solution, offering valuable insights into its implementation and potential societal impact. As social media platforms strive to cultivate safer and more inclusive digital communities, the adoption of this AI moderation service represents a key step towards achieving this goal.

2. LITERATURE REVIEW

[1] Heng Sun & Wan Ni (2022): This article proposed an AI-based TCM system, which was developed on an AI cloud service platform, and the system could automatically and intelligently analyze and detect the text content input by users from the web-end and app-end by calling built-in algorithm models on the cloud-end. The model developed in this paper can dynamically balance workload according to use cases, hence making the system more efficient.

[2] Parikshit S (2023): This research paper, they presented a comprehensive study on the advancements in OCR through the development and application of a novel deep learning algorithm for enhanced text recognition. Their algorithm harnesses the power of Convolution Neural Networks (CNNs) to achieve significant improvements in OCR accuracy, thereby overcoming several limitations of traditional OCR methods.

[3] Jiang A Q et al (2023): This paper presents the pre-trained model we are using in our application, Mistral 7B demonstrates that language models may compress knowledge more than what was previously thought. This opens up interesting perspectives: the field has so far put

the emphasis on scaling laws in 2 dimensions (directly associating model capabilities to training cost, as in other models); the problem is rather 3 dimensional (model capabilities, training cost, inference cost), and much remains to be explored to obtain the best performance with the smallest possible model.

3. SCOPE AND METHODOLOGY

3.1 Aim of the project

The artificial intelligence-based moderation service offers a comprehensive solution to increase the security of users of digital platforms. By perfectly identifying and filtering semantically toxic or offensive content, the service effectively reduces online harassment, bullying and hate speech and promotes a safer digital space for individuals to express their views without fear. This not only reduces the psychological impact of encountering toxic content, but also contributes to improving the overall user experience. With a reduced offensive presence, users are more likely to engage in meaningful discussions, act respectfully, and share their thoughts, increasing both user engagement and loyalty to the platform. The real-time response service features further ensure rapid detection and action against offensive content, preventing it from spreading widely across user devices.

3.2 Scope of the project

The scope of this project involves using advanced natural language processing (NLP) and machine learning (NLU) techniques, along with machine learning (ML) and artificial intelligence (AI), to perform in-depth content analysis to detect and moderate various forms. harmful content such as hate speech, bullying and misinformation, while understanding context, sarcasm, idiomatic expressions and cultural nuances in text across different languages and dialects. Data and privacy considerations are paramount, ensuring strict adherence to data protection laws and transparent policies around data collection and use. Integration and scalability are addressed through seamless integration with existing platforms and infrastructures coupled with high scalability to manage different volumes of data and user interactions. User interface and experience are prioritized through the development of intuitive interfaces and real-time notification systems with user feedback mechanisms to increase trust and transparency. Ethical and social implications are carefully considered to maintain a balance between freedom of expression and content moderation while ensuring compliance with regulatory requirements and industry standards. A comprehensive cost analysis is performed to assess return on investment in terms of improved platform security, user satisfaction and compliance.

3.3 Methodology:

The methodology involves deploying the Mistral 8x7B model to analyze real-time messages within social media groups. Once integrated into selected groups, the model continuously monitors messages and uses its semantic content analysis capabilities to identify offensive or malicious content. Detection triggers appropriate moderation actions, such as warnings or deletion of messages, aimed at promoting a safer online environment. The system may include ongoing updates and optimizations to increase its effectiveness over time.

4. IMPLEMENTATION

4.1 System Architecture

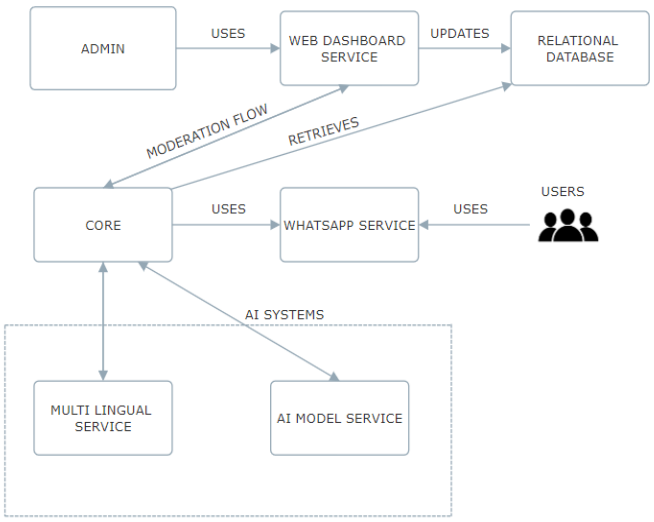


Fig-1: System Architecture

The implementation of Semantic Content Moderation for social media includes a systematic methodology to ensure effective monitoring and moderation of group messages. Initially, users log into a dedicated website associated with the moderation service, where they add a social media group to be moderated and configure the service settings to their specific requirements. Subsequently, the moderation service is integrated into the social media group, with users granting it administrative privileges to access and monitor messages. When messages are sent within a group, they are automatically routed to the server of the moderation service, where they undergo analysis using an AI model specialized in the analysis of semantic content. After the analysis is complete, the AI model sends a response notification back to the server indicating the potential offensiveness or corruption of the message. Based on this response, the moderation service server determines

the appropriate moderation action to take, which may include notifying the sender, removing the message, other actions according to settings configured by the user. In addition, the moderation system may include a feedback mechanism that allows users to report false positives or provide feedback on moderation decisions, thereby increasing the accuracy and effectiveness of the system over time. This methodology ensures a proactive approach to maintaining a safe and respectful online environment in social media groups and encourages positive interactions.

4.2 Sequence Diagram

Sequence diagrams describe interactions among classes in terms of an exchange of messages over time. They're also called event diagrams. A sequence diagram is a good way to visualize and validate various runtime scenarios. These can help to predict how a system will behave and discover responsibilities a class may need to have in the process of modeling a new system.

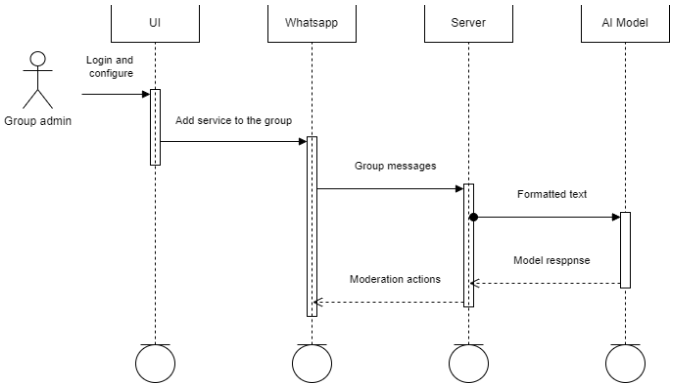
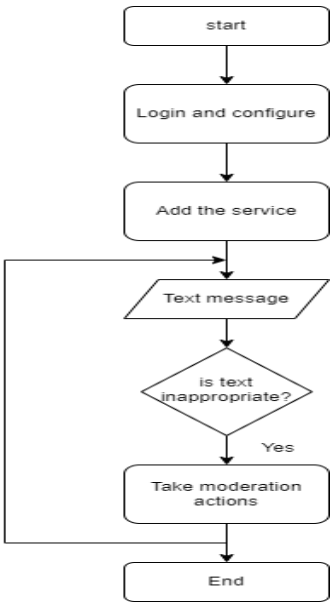


Fig- 2: Sequence Diagram

4.3 Activity Diagram



The activity diagram is another important diagram in UML to describe the dynamic aspects of the system activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. Activity is a particular operation of the system. Activity diagrams are not only used for visualizing the dynamic nature of a system but are also used to construct the executable system by using forward and reverse engineering techniques.

5. RESULT

The Semantic moderation of conversational content offers promising results in maintaining a favorable online environment using advanced natural language processing (NLP) techniques. By analyzing the semantic meaning of user-generated content, these systems can effectively identify and mitigate potentially harmful or inappropriate material and promote respectful and constructive interactions within digital platforms. Through constant refinement and adaptation, these moderation mechanisms demonstrate the ability to evolve with changing language patterns and cultural nuances, increasing their effectiveness in promoting positive online discourse and community engagement.

Fig- 4 User Dashboard

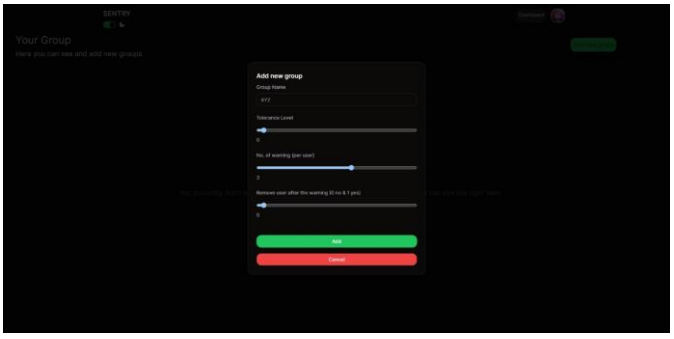


Fig- 5 Group configuration

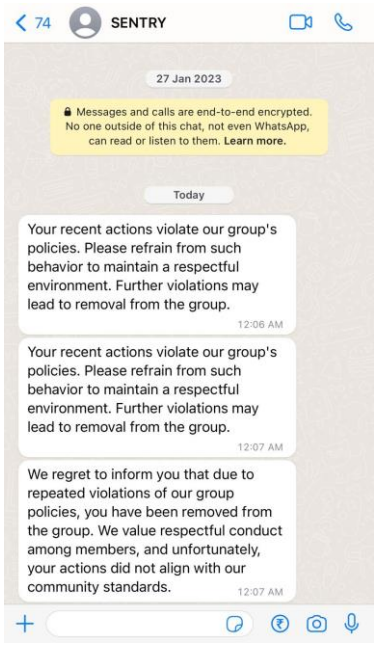


Fig- 6 Warning messages

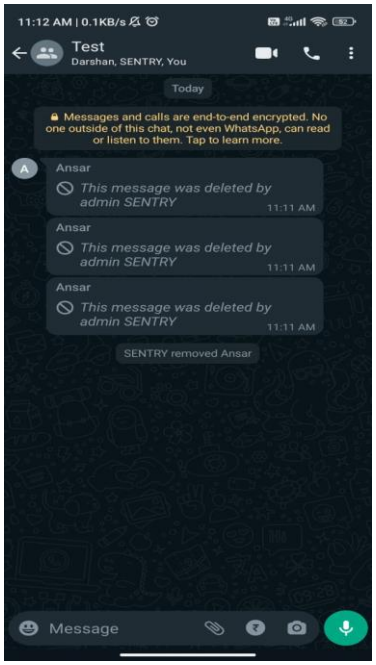


Fig- 7 Moderation actions

CONCLUSION

As we complete our project on a semantic content moderation system, we are thinking about our journey through the intricacies of natural language understanding, machine learning, and the ethical dimensions of content moderation. This effort has equipped us with a robust system capable of dynamically moderating online content to ensure a safer digital space. Designed for scalability and adaptability, our architecture effectively balances advanced algorithmic intervention with human judgment, increasing accuracy and efficiency in identifying and managing malicious content. The project highlighted the importance of continuous improvement and adaptation, with the feedback loop and analytics playing a key role in refining our approaches and understanding new trends. Going forward, we remain committed to evolving our system in response to the ever-changing digital landscape to foster an online environment where expression is free and respectful, thereby contributing to a healthier and more inclusive digital community.

REFERENCES

1. Sun, Heng & Ni, Wan. (2022). Design and Application of an AI-Based Text Content Moderation System. Scientific Programming. 2022. 1-9. 10.1155/2022/2576535.
2. Sharma, Parikshit. (2023). Advancements in OCR: A Deep Learning Algorithm for Enhanced Text Recognition. International Journal of Inventive Engineering andSciences. 10. 1-7. 10.35940/ijies.F4263.0810823.
3. Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, & William El Sayed. (2023). Mistral 7B.