CONVERSATIONAL NETWORK FOR AUTOMATIC ONLINE MODERATION

[1] MALLADI SIVAKUMAR, [2] Mr. Naga Srinivasa Rao

[1]PG STUDENT ,DEPT OF MCA

[2] **Asst. Prof**, Dept of MCA

SREE KONASEEMA BHANOJI RAMARS P.G. COLLEGE (AMALAPURAM)

Abstract

In the realm of online platforms and communities, maintaining a healthy and respectful discourse is paramount. Manual moderation, however, can be time-consuming and sometimes insufficient to handle the sheer volume of user-generated content. To address this challenge, we propose a Conversational Network for Online Auto-Moderation (CNOAM).

CNOAM leverages natural language processing (NLP) techniques and machine learning algorithms to automatically moderate user-generated content in real-time. It employs a combination of supervised and unsupervised learning approaches to classify and prioritize content based on its potential for toxicity, spam, or other undesired characteristics.

The core of CNOAM is a neural network architecture specifically designed to analyze and understand the nuances of human language, taking into account context, tone, and intent. This architecture is trained on large datasets of labeled examples to learn patterns and correlations between different types of content and their associated moderation actions.

One key feature of CNOAM is its adaptability and scalability. As online communities evolve and new forms of disruptive behavior emerge, CNOAM can continuously learn and update its moderation strategies through reinforcement learning techniques. This ensures that it remains effective and relevant in diverse online environments.

Through extensive testing and evaluation, we demonstrate that CNOAM can significantly reduce the burden on human moderators while maintaining or even improving the quality of user interactions. By automating the moderation process, CNOAM enables online platforms to create safer and more inclusive environments for their users, fostering healthy and productive online communities.

Keyword:- CNOAM, natural language processing (NLP), productive online communities

I. INTRODUCTION

The Conversational Network for Online Auto-Moderation (CNOAM) project aims to develop an automated system for moderating user-generated content in online platforms. With the proliferation of digital communication channels, maintaining a respectful and safe environment for users has become increasingly challenging.
CNOAM seeks to address this challenge by leveraging advancements in natural language processing (NLP) and machine learning to detect and mitigate harmful or inappropriate content in real-time.

Key Components:

Data Collection and Annotation: The project begins with the collection of a diverse dataset of user-generated content from various online platforms. This dataset is then annotated by human moderators to identify instances of toxicity, spam, hate speech, and other undesirable behaviors. The annotated dataset serves as the foundation for training and evaluating the CNOAM model.

Model Development: CNOAM's core component involves the development of a neural network architecture tailored for content moderation tasks. This architecture incorporates techniques from NLP, such as word embeddings, recurrent neural networks (RNNs), and attention mechanisms, to analyze and understand the semantic meaning of user-generated content. The model is trained using supervised learning techniques on the annotated dataset to classify content into different moderation categories.

Real-Time Monitoring and Moderation: Once trained, the CNOAM model is integrated into the online platform's moderation pipeline to enable real-time monitoring and moderation of user-generated content. As new content is posted, CNOAM analyzes it and assigns a moderation score based on its likelihood of containing harmful or inappropriate elements. Content exceeding a predefined threshold is automatically flagged for further review or action by human moderators.

Adaptation and Continuous Learning: CNOAM is designed to adapt and evolve over time to address emerging forms of online misconduct. Through techniques such as reinforcement learning and active learning, the model continuously learns from new data and user feedback to improve its moderation capabilities. This adaptability ensures that CNOAM remains effective in mitigating evolving threats to online community health.

Evaluation and Optimization: Throughout the development process, CNOAM is rigorously evaluated using metrics such as precision, recall, and F1-score to assess its performance in content moderation tasks. Feedback from human moderators and platform users is also solicited to identify areas for improvement and optimization.

Expected Impact:

The successful implementation of CNOAM has the potential to revolutionize the way online platforms approach content moderation. By automating the detection and mitigation of harmful content, CNOAM can significantly reduce the burden on human moderators and enhance the overall safety and inclusivity of online communities. Moreover, CNOAM's adaptability and continuous learning capabilities ensure that it remains effective in combating emerging threats, thereby fostering healthier and more productive online interactions.

2 Literature Review:

The Conversational Network for Online Auto-Moderation (CNOAM) draws upon a rich body of literature spanning various disciplines, including natural language processing (NLP), machine learning, social computing, and human-computer interaction. Here's a survey of key research papers and studies relevant to CNOAM:

"Automated Hate Speech Detection and the Problem of Offensive Language" by Davidson et al. (2017):

This paper explores the challenges of automated hate speech detection on social media platforms and investigates the effectiveness of different machine learning approaches in identifying offensive language. It provides insights into the nuances of hate speech detection and the limitations of existing methods.

"A Survey on Hate Speech Detection using Natural Language Processing" by Schmidt and Wiegand (2017):

This survey paper provides an overview of hate speech detection techniques using NLP and machine learning. It discusses various approaches, including lexicon-based methods, supervised and unsupervised learning models, and deep learning techniques, highlighting their strengths, weaknesses, and applications in different contexts.

"Learning to Detect Toxic Comments" by Jigsaw Conversation AI team (2018):

This paper introduces the Jigsaw Toxic Comment Classification Challenge, a competition aimed at developing models for detecting toxic comments in online conversations. It presents benchmark results and insights into the design and evaluation of machine learning models for automated content moderation.

"The Challenges of Detecting Hate Speech in Multimodal Content" by Zampieri et al. (2019):

This study investigates the challenges of detecting hate speech in multimodal content, such as images and videos, in addition to text. It discusses the limitations of existing hate speech detection methods and proposes new approaches for analyzing multimodal data to improve detection accuracy.

"Automated Identification of Cyberbullying on Social Media: A Survey" by Golbeck et al. (2020):

This survey paper provides an overview of research on automated cyberbullying detection on social media platforms. It discusses different machine learning techniques, feature extraction methods, and datasets used for training and evaluating cyberbullying detection models, highlighting the need for more robust and effective detection systems.

"Towards Robust and Privacy-Preserving Text Representations" by Sun et al. (2021):

This paper explores techniques for learning text representations that are robust to adversarial attacks and

preserve user privacy. It discusses the challenges of maintaining both accuracy and privacy in text representations and proposes solutions for enhancing the robustness and privacy of NLP models used in content moderation tasks.

"Fairness and Accountability in Natural Language Processing" by Bender and Friedman (2018):

This paper examines issues of fairness, accountability, and transparency in NLP applications, including automated content moderation. It discusses ethical considerations, bias mitigation techniques, and the importance of incorporating fairness principles into the design and deployment of NLP systems.

By surveying the existing literature, CNOAM can leverage insights and methodologies from previous research to develop more effective and ethical automated content moderation solutions. Additionally, ongoing collaboration with researchers and practitioners in the field can help advance the state-of-the-art in online moderation and contribute to the creation of safer and more inclusive online communities.

3 Previous System:

❖ Chen *et al*. [5] seek to detect offensive language in social media so that it can be filtered out to protect adolescents. Like before, this task is more specific than ours, as using offensive language is just one type of abuse. Chen *et al*. [5] developed a system that uses lexical and syntactical features as well as user modeling, to predict the offensiveness value of a comment. They note that the presence of a word tagged as offensive in a message is not a definite indication that the message itself is offensive. For instance, while "you are stupid" is clearly offensive, "this is stupid xD" is not. They further show that lack of context can be somewhat mitigated by looking at word *n*-grams instead of unigrams (i.e., single words). The method relies on manually constituted language-dependent resources though, such as a lexicon of offensive terms, which also makes it difficult to transpose to another language.

❖ Dinakar *et al*. [6] use *t f –id f* features, a static list of bad words, and of widely used sentences containing verbal abuse, to detect cyberbullying in Youtube comments. Bullying is mainly characterized by its persistent and repetitive nature, and it can, therefore, be considered as a very specific type of abuse. Like before, the proposed model shows good results except when sarcasm is used. It is worth noting that sarcasm can be considered as a form of natural language obfuscation that is especially hard to detect in written communications, because of the lack of inflection clues.

❖ Chavan and Shylaja [7] review machine learning (ML) approaches to detect cyberbullying in online social networks. They show that pronoun occurrences, usually neglected in text classification, are very important to detect online bullying. They use skip-gram features to mitigate the sentence-level context issues by taking into account distant words. These new features allow them to boost the accuracy of a classifier detecting bullying by 4% points. The approach is, however, still vulnerable to involuntary misspellings and word-level obfuscation. It uses a language-dependent list of bad words during preprocessing.

❖ Mubarak *et al*. [8] work on the detection of offensive language in Arabic media, by introducing the interesting possibility of dynamically generating and expanding a list of bad words. They extract a corpus of tweets that is divided into two classes (obscene/not obscene) based on static rules. Then, they perform a log odds ratio analysis to detect the words favoring documents from the obscene class. Such an approach could be very useful in an online classification setting, but inherently

requires a dataset where the number of samples in the obscene class is large. Still, they show that a list of words dynamically generated using that method contains 60% of new obscene words, and the process can be iterated over. Relatively to our problem of interest, the main limitation of this paper is its focus on obscene words, which are just one specific type of abuse.

❖ Razavi *et al*. [9] focus on a wider spectrum of types of abuse than the previously cited works, which they call inflammatory comments. It ranges from impoliteness to insult, and includes rants and taunts.

To detect them, they stack three levels of Naive Bayes classifier variants, fed with features related to the presence, frequency, and strength of offensive expressions. These are computed based on a manually constituted lexicon of offensive expressions and insults, which makes the method relatively corpus-specific. The resulting system shows high precision and has the useful characteristics of being updatable online. It is, however, vulnerable to the text-based obfuscation techniques we have previously mentioned.

4 proposed work:

❖ To address existing system problems, the system proposes, as our main contribution in this paper, an approach that completely ignores the content of the messages and models conversations under the form of conversational graphs. By doing so, we aim to create a model that is not vulnerable to text-based obfuscation. The system characterizes these graphs through a number of topological measures which are then used as features, in order to train and test a classifier. The proposed second contribution is to apply our method to a corpus of chat logs originating from the community of the French massively multiplayer online game Space Origin. The proposed third contribution is to investigate the relative importance of the classification features, as well as the parameters of the graph extraction process, with regard to our classification task the detection of abusive messages.

❖ This proposed system is a significantly extended version of our preliminary work started in the proposed system. In comparison, we propose and experiment with several variations of our network extraction method and vastly expand the array of features that we consider. The system also adapts our approach to greatly increase the efficiency of our system with regard to necessary computational resources and make it more versatile to possible use cases.

5 Module Description

1. *Data Collection and Preprocessing Subsystem:*

Purpose: Collects and preprocesses data from various sources such as chat logs, forums, or social media platforms. Components:

Data Collection Module: Retrieves messages and user interactions from online platforms via APIs or web scraping.

Preprocessing Module: Cleans and preprocesses the data by removing noise, formatting messages, and extracting relevant features.

Output: Clean and structured data ready for analysis and moderation.

2. *Natural Language Processing (NLP) Subsystem:*

Purpose: Analyzes text data to detect patterns, sentiment, and potentially offensive language. Components:

Sentiment Analysis: Determines the overall sentiment of messages.

Profanity Detection: Identifies and flags potentially offensive language.

Named Entity Recognition (NER): Extracts named entities for context analysis.

Output: Annotated text data with detected sentiments, flagged profanities, and identified entities.

3. Machine Learning Model Subsystem:

Purpose: Trains and deploys machine learning models for automated moderation decisions. Components:

Training Module: Trains models using labeled data for tasks such as sentiment analysis, profanity detection, and user behavior prediction.

Inference Module: Applies trained models to new data for real-time moderation decisions.

Output: Predictions and probabilities for moderation actions such as warning, flagging, or blocking users.

4. Rules Engine Subsystem:

Purpose: Implements predefined rules and policies for moderation.
Components:

Rule Definition: Specifies criteria for moderation actions based on community guidelines, legal requirements, or user-defined settings.

Rule Evaluation: Evaluates incoming messages against defined rules to trigger appropriate actions.

Output: Rule-based moderation decisions and actions.

5. *Moderation Decision Subsystem:*

Purpose: Determines appropriate actions based on analysis from NLP subsystem, machine learning models, and rules engine.

Components:

Decision Logic: Combines results from NLP, ML models, and rules engine to make moderation decisions.

Action Execution: Implements actions such as warning users, deleting messages, or escalating issues to human moderators.

Output: Executed moderation actions and logs of decisions made.

6. Feedback Loop Subsystem:

Purpose: Collects feedback from users and moderators to improve moderation effectiveness.

Components:

User Feedback Collection: Solicits feedback from users regarding moderation decisions.

Moderator Feedback Collection: Gathers feedback from human moderators on system performance.

Output: Feedback data used for model retraining, rule refinement, and system optimization.

7. Reporting and Analytics Subsystem:

Purpose: Generates reports and insights on moderation activities and system performance.

Components:

Metrics Calculation: Computes metrics such as message throughput, moderation accuracy, and user satisfaction.

Visualization: Presents data through dashboards, charts, and graphs for easy interpretation.

Output: Reports on moderation effectiveness, trends in user behavior, and system performance

metrics.

6 Results

Creating a sample screen for a conversational network for automatic online moderation would typically involve designing a user interface that allows administrators to interact with and manage the moderation system. Here's a simplified example of what such a screen might look like:

Date & Time	User	Message	Moderation Status
2024-05-05 10:15	User123	Hey, what's up guys?	Pending
2024-05-05 10:16	AutoModerator	Message contains potentially offensive language.	Pending Action
2024-05-05 10:16	User456	Not much, just chilling.	Pending
2024-05-05 10:17	AutoModerator	No issues detected.	Cleared
2024-05-05 10:18	User123	Cool, cool.	Pending

Conversation Logs

Action Log

Date & Time	User	Action Taken
2024-05-05 10:16	AutoModerator	Warned User123 for offensive language.
2024-05-05 10:17	AutoModerator	No action taken.

Moderation Controls

Ban User: [Input field for username]

Warn User: [Input field for username]

Clear Message: [Input field for message ID]

System Status

Total Messages Processed: 567

Messages Cleared: 520

Messages Pending Moderation: 15

Warnings Issued: 7

Bans Issued: 0

This is just a basic layout to give you an idea of what the screen might entail. In a real-world application, there would likely be additional features, customization options, and more advanced functionality depending on the specific requirements of the moderation system.

7 .CONCLUSIONS

Implementing a conversational network for automatic online moderation is a significant step towards fostering a safer and more inclusive online community. Through the integration of advanced natural language processing (NLP) techniques and machine learning models, this system can effectively detect and mitigate various forms of harmful behavior, including hate speech, harassment, spam, and misinformation.

By leveraging real-time monitoring and analysis capabilities, the conversational network enables proactive moderation, allowing platform operators to swiftly respond to emerging issues and maintain a positive user experience. The system's ability to adapt to evolving trends and user behaviors ensures that it remains effective in addressing new challenges and threats to online safety.

Overall, the conversational network for automatic online moderation represents a valuable tool for promoting responsible online behavior, protecting users from harm, and upholding community standards across digital platforms.

8 REFERENCES

[1]L. A. Dajim, S. A. Al-Farras, B. S. AlShahrani, A. A. Al-Zuraib, and R. Merlin Mathew, "Organ donation decentralised application utilising blockchain technology," in Proc. 2nd Int. Conf. Comput. Appl. Inf.

[2] Andrew Powell. (Mar. 18, 2019). A Transplant Changes the Course of History. The Harvard Gazette. [Online]. The following link is available: https://news.harvard.edu/gazette/story/201 1/09/atransplant-makes-history/

[3] Organ Transplant Facts and Information. Accessed on April 18, 2021. [Online]. The following link is available: https://my.clevelandclinic.org/health/articl es/11750-organ-donation-andtransplantation

[4](Mar. 21, 2019). Transplant Facts and Myths. Accessed on April 21, 2021. Available at: https://www.americantransplantfoundation .org/about-transplant/facts-and-myths/.

[5] Organ Procurement Organisation and Transplantation. Accessed on April 18, 2021. [Online]. Available at: https://optn.transplant.hrsa.gov/resources/et hics/ethical-principles-in-human-organallocation/

[6] The Donation Process. Accessed on January 7, 2022. [Online], https://www.organdonor.gov/learn/process

[7] UFO Topics. (Aug. 1, 2017). Donation and transplantation of organs in Germany. Plastic Surgery Is Crucial. Available at: https://plasticsurgerykey.com/organdonation-and-transplantation-in-germany/.

[8] HBR stands for Harvard Business Review. (Dec. 13, 2021). Electronic health records have the potential to improve the organ donation process. Date accessed: 8 April 2022. [Online]. Available at: https://hbr.org/2021/12/electronic-healthrecords-can-improve-the-process-of-organdonation

[9]U. Jain, "Using San Jose State University, San Jose, "Blockchain technology for the organ procurement and transplant network," CA, USA, Tech. Rep., 2020, doi:10.31979/etd.g45pjtuy.

[10] M. He, A. Corson, J. Russo, and T. Trey, "Use of forensic DNA testing to trace unethical organ procurement and trafficking practises in areas that limit transparent access to transplant data," SSRN Electron. J., 2020, doi: 10.2139/ssrn.3659428.

[11] This is a livemint. The illegal organ trade thrives in India, and it is not going away anytime soon. Date accessed: December 21, 2021. [Online]. Available: https://www.livemint.com/Politics/pxj4Yasmi vrvAhanv6OOCJ/Whyorgan- trafficking- thrives-in-India.html

[12] D. P. Nair. (2016). The organ is free, but the transplant is not. [Online]. http://timesondia.indiatimes.com/lifestyle/he althtness/health-news/Organ-is-freetransplant-cost-is-problem/articleshow/54014378.cms