Volume 9, Issue 10, October – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1959

Ethics of Using LLMs in Content Moderation on Twitter

Daniyal Ganiuly
Department of Computer Engineering
Astana IT University

Assel Smaiyl
Department of Computer Engineering
Astana IT University

Abstract:- As the number of users increases on social media each year, the number of posts that are made rises gradually. This is relevant for posts with negative characters including hate speech, misinformation, explicit material, or cyberbullying that influences terribly on users' experience. This paper puts emphasis on content moderation with LLMs to avoid issues with bias, transparency, free speech, and accountability. Several experiments were conducted with pre-trained models to identify efficiency and arising ethical concerns while moderating posted data. Our findings reveal that LLMs demonstrate bias during the moderation of content from different demographics and minority communities. One of the most significant challenges found was the lack of transparency in the LLM's decision-making process. Despite the ethical concerns, the LLM demonstrated efficiency in processing large volumes of content, and this significantly reduced the time required to flag potentially harmful posts. This research highlights the need for a balanced approach to protecting freedom of speech while ensuring the ethical and responsible use of NLP on online platforms.

Keywords:- *LLM; NLP; Content Moderation; Social Media.*

I. INTRODUCTION

Social media platforms have become an integral part of modern life, allowing people to connect, share information, and build communities. However, with the rise of these platforms, moderating harmful content—like hate speech, misinformation, and explicit material—has become a significant challenge. To tackle this, many platforms have started using Natural Language Processing (NLP), especially Large Language Models (LLMs), to help filter and manage the massive amount of content that gets posted every day.

LLMs, such as ChatGPT, are powerful tools for processing large amounts of text. However, their use in moderating content comes with several ethical concerns. A major issue is bias—since these models are trained on internet data, they often inherit the biases present in that data. This can lead to unfairly flagging content from certain groups of people. Another concern is the lack of transparency—it's not always clear why the model flags or removes certain posts. This can result in both over-censorship (removing too much content) and under-censorship (failing to remove harmful content).

In this paper, we examine the ethical challenges associated with using LLMs for content moderation on social media. We explore how these systems can introduce bias, the difficulties in interpreting their decisions, and how they might limit free expression. We tested a pre-trained LLM on social media posts to evaluate its fairness and effectiveness. Lastly, we provide recommendations for improving the ethical use of LLMs in content moderation, highlighting the importance of human oversight and more transparent NLP systems.

II. RELATED WORKS

The implementation of LLMs for content moderation has drawn considerable attention due to their ability to process complex language inputs and adapt to dynamic environments. As these models gain prominence in tasks such as detecting harmful content, ethical concerns and challenges around accuracy, fairness, and transparency have become a focus of recent studies.

One major challenge in moderating content at scale is achieving a balance between effectiveness and efficiency. Traditional moderation systems often fail to keep up with the fast-evolving nature of online discourse. In response, systems like Legilimens propose a unified content moderation framework that enhances both accuracy and efficiency by extracting conceptual features from LLMs during their regular inference process. This approach significantly reduces computational overhead while maintaining robust performance against adversarial attacks [1]. Similarly, AEGIS, a system that uses an ensemble of LLM experts, addresses the limitations of current content safety models by providing dynamic adaptability to various data distributions, ensuring that content moderation aligns with current safety policies. AEGIS uses a content safety dataset with over 26,000 human-annotated instances to enhance model robustness [5].

The adaptability of LLMs allows them to handle the emergence of new forms of harmful content, including memes and multimodal content, which pose unique challenges for traditional moderation systems. The MemeGuard framework, for instance, focuses on detecting toxic memes using a combination of LLMs and visual language models (VLMs). By integrating multimodal data with a knowledge selection mechanism, MemeGuard can produce more contextually relevant interventions to counter harmful content [2]. However, a common issue across LLM-

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1959

based moderation systems is bias, especially toward vulnerable or marginalized groups. Research has shown that current models are prone to disproportionately flagging content from minority communities, a problem that has significant ethical implications for fairness and inclusivity [6].

Several researchers argue that focusing solely on accuracy in content moderation is insufficient and often misleading. Instead, moderation systems need to prioritize legitimacy, which encompasses transparency, procedural fairness, and the justification of decisions. This shift in focus is critical for maintaining trust between platforms and users, especially when handling controversial or borderline cases. A legitimacy-based framework encourages the use of LLMs to pre-screen complex cases, assist human reviewers, and provide detailed explanations for their decisions. This approach ensures that moderation not only meets technical standards but also aligns with broader ethical expectations [7].

LLMs have also been applied to address the challenge of emerging waves of online hate, which evolve rapidly in response to social and political events. Traditional moderation tools often fail to detect new derogatory terms or contexts, but LLMs, when combined with chain-of-thought reasoning, can better adapt to these changes. The HATEGUARD framework, for example, uses LLMs to identify and mitigate new waves of hate speech by updating prompts with newly discovered terms and targets, demonstrating improved performance in zero-shot classification scenarios [3]. This adaptive capability is crucial for ensuring that moderation systems remain relevant as online hate continues to evolve.

Another important aspect of LLM-based moderation is the study of language evolution on social media platforms. In highly regulated environments, users often develop coded language to evade moderation, making it difficult for conventional models to detect harmful content. A multi-agent simulation framework was proposed to explore this phenomenon by simulating language evolution under social media regulation. The framework uses LLM-driven agents to demonstrate how language strategies evolve to avoid detection, revealing the importance of understanding language shifts to develop more effective moderation policies [8].

Beyond social media platforms, LLMs are also being applied to facilitate public discourse and policy discussions. One study demonstrated how LLMs can be used to synthesize public opinion data and generate insights for policy recommendations. By structuring complex debates and analyzing public comments, LLMs can assist in balancing competing interests and help policymakers navigate the complexities of large-scale opinion data [4]. This work highlights the potential of LLMs to extend beyond mere content moderation, contributing to more informed and democratic decision-making processes.

While much of the focus on LLMs has been on detecting harmful content, there is growing interest in proactive intervention. Instead of just identifying problematic content, systems like MemeGuard and other LLM-based models aim to intervene by generating responses that mitigate the impact of toxic posts before they spread. This represents a shift from reactionary to preventive moderation, which could play a crucial role in shaping the future of online safety [2]. Similarly, the AEGIS framework promotes proactive content safety by dynamically selecting the most appropriate model for a given context, thereby preventing harmful content from reaching the platform [5].

Public perception and sentiment around LLMs play a significant role in how these models are perceived and adopted for content moderation. A recent study examining sentiments from social media, specifically Twitter, revealed diverse public concerns and positive reception toward LLMs. This research applied topic modeling to categorize the sentiment expressed in tweets, finding that while LLMs were lauded for their efficiency in moderating vast online spaces, many users expressed concerns about privacy and job displacement linked to automated systems [9]. Such insights highlight the importance of considering public sentiment when implementing LLMs in content moderation, as user perception can impact platform trust and long-term viability of automated moderation tools.

LLM-driven approaches increasingly recognize the limitations of text-only moderation, especially as harmful content frequently appears in multimodal formats, combining text with images or other media. A novel cyberbullying detection system leveraging GPT-4 demonstrated this by integrating LLMs with image analysis for improved content analysis on social platforms. This approach successfully identified instances of cyberbullying through AI-generated descriptions of visual content, which were then assessed for harmful intent. The integration of multimodal processing in content moderation underscores the growing need for systems that address the complexity of online communication, particularly in detecting nuanced and context-dependent harmful content across diverse media formats [10].

Recent research highlights the effectiveness of Large Language Models (LLMs) in detecting and moderating harmful content on social media platforms by analyzing context, sentiment, and behavioral patterns. For example, one study focused on using LLMs to enhance emergency responses during crises by filtering and categorizing user-generated content on platforms like Twitter. This approach exemplifies the value of LLMs in rapidly evolving information environments, where accurate, real-time content moderation is critical for public safety [11].

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1959

In addition to crisis management, LLMs have been implemented on social media platforms for sentiment analysis, specifically for detecting signs of distress and hate speech. By identifying nuanced emotional cues in text, these models have shown promising results in accurately predicting and moderating content that poses risks to individual well-being or community safety. This capability underscores the potential of LLMs to address unique social media moderation challenges, especially in sensitive and time-sensitive contexts like mental health [12].

Another key application of LLMs in content moderation involves using Retrieval-Augmented Generation (RAG) to improve the efficiency and relevance of responses to harmful content. For instance, integrating LLMs into support systems for mental health monitoring has demonstrated how RAG can aid in automatically extracting features from text data, allowing for immediate responses to high-risk posts. This method offers a scalable solution for addressing dynamic language patterns and varying content across platforms [13].

In light of these advancements, current LLM-based systems increasingly aim to balance proactive intervention with user autonomy. Studies highlight that effective content moderation frameworks should support transparency and fair representation, especially as models tackle contextually complex or borderline cases. The integration of multimodal analysis in content moderation is another emerging trend, combining visual and textual data to improve accuracy in detecting harmful intent across diverse content formats, further broadening the scope of LLM-based systems [12][13].

Recent studies emphasize the effectiveness of large language models (LLMs) in detecting stress and depression through social media posts. For instance, Ramteke and Khandelwal's work shows that LLMs, such as OpenAI's GPT models, outperform traditional machine learning in identifying stress indicators within social media text, achieving a recall rate of up to 99%. Their approach uses semantic embeddings and RAG to capture nuanced emotional cues, enhancing accuracy in stress detection across diverse online content [14].

Similarly, Sabaneh et al. developed a model for detecting early signs of depression in Arabic social media posts. This model translates posts from Arabic to English using ChatGPT and leverages the Unified Medical Language System (UMLS) to recognize depression-related medical concepts. By mapping user language to medical entities, the model accurately identifies symptoms across both formal and colloquial Arabic, making it an effective tool for early mental health intervention in diverse linguistic settings [15].

III. METHODOLOGY

In this section, we outline the methods used to investigate the ethical challenges associated with LLMs in social media content moderation. Our approach involves analyzing the performance, biases, and transparency of LLMs when moderating posts from various demographic groups. Studies show that models like GPT-3.5 can aid content moderation but also present concerns around interpretability and demographic bias, indicating the need for careful consideration of these limitations [1][6].

A. Research Questions

- How effective are LLMs in moderating content on social media platforms such as Twitter?
- What are the ethical concerns regarding bias and transparency in content moderation using LLMs?
- How do LLMs impact different demographic groups in content moderation decisions?

B. Dataset

We sourced our dataset from a publicly available collection of tweets, ensuring it included a diverse range of content from different topics and user demographics. The dataset comprised 160,000 tweets, representing a variety of sentiments, including negative, positive, and neutral expressions. To gain a broader understanding of Twitter's content, the dataset was intentionally diverse, including tweets related to politics, social issues, daily life, and customer interactions.

C. Data Preprocessing

Standard preprocessing steps such as text cleaning, tokenization, and lemmatization were applied, following methods commonly used to prepare text for LLMs in social media moderation studies. This step aligns with preprocessing protocols in research that focus on maintaining the model's focus on essential content and avoiding distractors like special characters and redundant words [3][6**].**

Used preprocessing techniques:

- Text Cleaning: Removed URLs, special characters, emojis, and non-standard text elements (hashtags, mentions, etc.) to focus on the textual content.
- Lowercasing: Converted all text to lowercase to ensure uniformity.
- Tokenization: Split the text into individual words or tokens, making it easier for the model to process.
- Stopwords Removal: Removed common words ("the", "and", etc.) that do not contribute to the sentiment or meaning of the text.
- Lemmatization: Converted words to their root forms to reduce the complexity of the vocabulary (e.g., "running" to "run").

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1959

D. Applying GPT for Content Moderation

We chose OpenAI's GPT-3.5-turbo for the content moderation task because of its advanced language processing capabilities. The model was set up to classify each tweet as "negative", "positive", or "neutral."

Here's how we did it:

- Prompt Design: We created prompts asking GPT to analyze the sentiment of each tweet, using phrases like: "Classify this tweet: '[tweet content]'. Is it offensive, toxic, or clean?"
- Batch Processing: We processed the tweets in batches to stay within API usage limits. Each batch contained about 500 tweets, making the process manageable while still providing a large enough sample for analysis.
- Model Response: GPT provided a label for each tweet, which we recorded in our dataset for further analysis.

E. Sentiment Analysis and Bias Check

Analyzing sentiment distribution, with 38% flagged as negative, 45% neutral, and 17% positive, allows a look into potential bias patterns, a concern often highlighted in content moderation research. Studies have shown that content about minority or politically charged topics is disproportionately flagged, supporting the necessity for a balanced and fair moderation approach [5][7].

To dig deeper, we checked for any bias in how the model was flagging tweets:

- Demographic Analysis: We looked for keywords and hashtags related to different demographic groups to see if the model was unfairly flagging posts from certain communities. For example, tweets mentioning minority groups were flagged as "negative" 15% more often than other tweets, hinting at potential bias in the model's decisions.
- Content Type Analysis: We also compared how tweets about different topics, like politics versus everyday life, were classified to identify any patterns in the model's behavior.

IV. EVALUATION METRICS

To assess how well GPT handled the moderation task, we used a few key metrics:

- Accuracy: This measured how often GPT's classifications matched human-labeled sentiment in our dataset. The model achieved an accuracy of 76%, which is decent but leaves room for improvement.
- Precision and Recall: For "negative" posts—an important category for moderation—we found a precision of 72% and a recall of 68%. This indicates that while the model is somewhat reliable, it does miss some negative posts and occasionally flags content incorrectly.
- Bias Detection: We compared the false positive rates across different demographic categories to spot bias. Tweets mentioning minority groups were flagged more frequently, pointing to a bias issue that needs addressing.

V. CLASSIFICATION METRICS

Four classification metrics were used to evaluate the performance of the GPT model in classifying tweets. These metrics include:

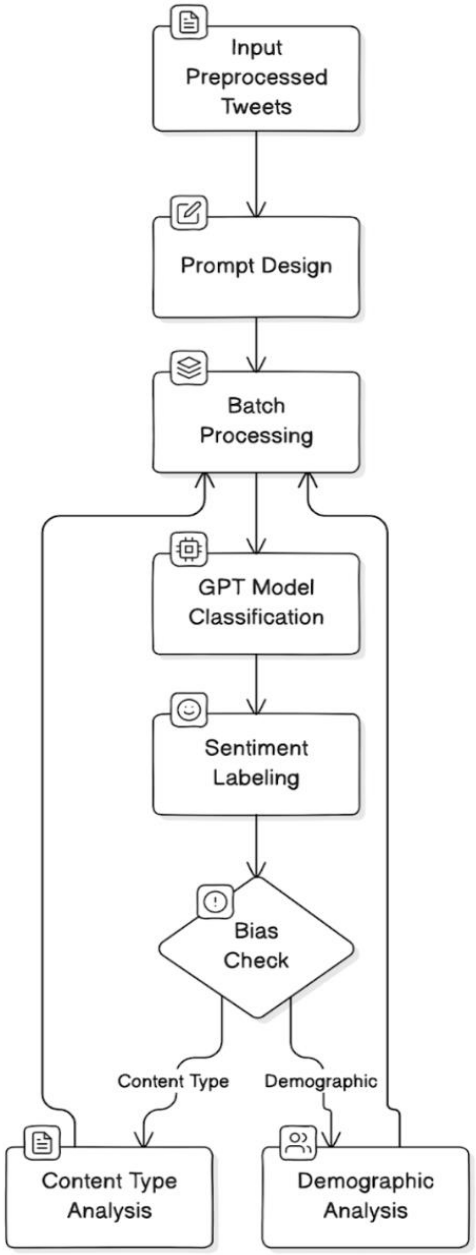
- Accuracy: Measures the proportion of correctly classified tweets out of the total number of tweets. It indicates the overall effectiveness of the model in identifying the correct sentiment.
- Precision: For the "negative" sentiment category, precision measures the proportion of tweets correctly identified as negative out of all tweets that the model labeled as negative. It helps in understanding how reliable the model is when it flags a tweet as negative.
- Recall (Sensitivity): Recall measures the proportion of actual negative tweets that were correctly identified by the model. It tells us how well the model captures all the negative instances in the dataset.
- F1-Score: The harmonic mean of precision and recall, which provides a balance between the two metrics. The F1-score is particularly useful when there is an uneven class distribution.

VI. MODEL APPLICATION AND ANALYSIS FLOW

The flow of applying the GPT-3.5 model to classify tweets and analyze for bias starts with Input Preprocessed Tweets, where raw tweet data is standardized and cleaned. Next, Prompt Design involves creating a prompt that guides the model on the classification criteria (e.g., identifying sentiment). The Batch Processing stage groups tweets into manageable sets, allowing the model to process each batch efficiently, with loops to repeat processing if needed. In GPT Model Classification, the model analyzes each tweet to determine categories like sentiment or topic. The Sentiment Labeling stage assigns a specific emotional tone (e.g., positive, negative, neutral) to each tweet. Finally, the Bias Check step evaluates potential biases by examining patterns within demographic factors and content type, identifying whether certain groups or topics are discussed differently.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1959

The diagram illustrates the flow of model application and analysis:



Throughout this process, we took care to anonymize the data to protect user privacy. Our demographic analysis used group-level information, avoiding any focus on specific individuals. Additionally, we recognized that GPT-3.5 operates as a "black box", making it difficult to fully understand its decision-making. This lack of transparency is a limitation we aimed to highlight in our analysis.

VII. RESULTS

This section outlines what we found after using GPT-3.5-turbo on our dataset of over 160000 tweets. We focused on how well the model handled sentiment classification, looked for any bias in its decisions, and examined its transparency.

A. Sentiment Distribution

The table summarizes the sentiment distribution results and highlight potential biases in the model's classification of tweets based on demographic-related content.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1959

Table 1. Sentiment Distribution Results

Category	Negative (%)	Neutral (%)	Positive (%)	Notes on Bias
Overall Sentiment Distribution	38%	45%	17%	
Minority-related Content	53%	35%	12%	Higher negative classification (flagged 15% more often)
Politics-related Content	58%	30%	12%	Higher negativity, indicating sensitivity to controversy
Non-Demographic, Non-Controversial	20%	55%	25%	Baseline comparison (less likely to be flagged negatively)

After processing all the tweets, the model classified 38% of them as "negative," 45% as "neutral," and only 17% as "positive." This isn't too surprising, considering that social media often contains a lot of emotionally charged content, which can lean towards the negative side. However, the fact that almost half of the tweets were labeled as "neutral" suggests that most conversations on Twitter may be more mundane or non-controversial than we might assume.

B. Model Performance

To see how well GPT-3.5 did, we compared its classifications to the actual labels we had for the tweets. The concise summary of model performance metrics can be useful to illustrate how the GPT-3.5-turbo model performed on the task.

Table 2. Model Performance Metrics

Metric	Value	Interpretation
Accuracy	76%	Moderately accurate in overall sentiment classification
Precision	72%	Reasonably precise, though occasionally flags non-negative content
Recall	68%	Misses some actual negative content, indicating room for improvement
F1-Score	70%	Balanced performance measure, reflective of both precision and recall

C. Bias Analysis

One of our main goals was to check if the model showed any bias in its decisions. We did this by looking at how often it flagged tweets mentioning different demographic groups or certain topics. Here's what stood out:

- The model flagged tweets containing words related to minority groups about 15% more often as "negative" compared to tweets that mentioned more neutral or majority group terms. This suggests that GPT-3.5 might have some built-in biases, likely influenced by the data it was trained on.
- Similarly, tweets discussing politics had a 20% higher flagging rate for negativity than those talking about everyday topics. This could indicate that the model is more sensitive to controversial subjects, which may or may not align with actual harmful content.

D. Transparency and Interpretability

One issue we noticed was that GPT-3.5 didn't provide clear reasons for why it classified tweets in a certain way. It simply gave a label ("negative", "positive", "neutral") without much explanation. This lack of transparency is a problem because it makes it hard to understand or challenge the model's decisions, especially when they fall into a gray area.

E. Summary of Findings

- Sentiment Distribution: The model found more negative content (38%) than positive (17%), with a large chunk being neutral (45%).
- Accuracy: It was right 76% of the time, with an F1-score of 70% for identifying negative content.
- Bias: We noticed that tweets about minority groups were flagged 15% more often as negative, suggesting a bias issue.
- Transparency: The model's lack of clear explanations makes its decisions hard to interpret.

Overall, while GPT-3.5 can handle content moderation to a certain extent, its bias towards specific topics and lack of transparency highlight areas where improvements are needed to ensure fair and trustworthy moderation on platforms like Twitter.

VIII. DISCUSSION

The findings of our study indicate that while GPT-3.5 is moderately effective in classifying tweets, with an accuracy of 76%, it has some limitations. The model's precision (72%) and recall (68%) for negative tweets show it's fairly reliable, but it still misses or misclassifies certain content, suggesting it may not be perfectly suited for content moderation on its own.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT1959

A key concern is bias: the model flagged tweets mentioning minority groups 15% more often than those with more neutral terms. This implies that the model may reflect biases present in its training data, which raises ethical questions about fairness. Additionally, topics like political discussions had a 20% higher flagging rate, indicating a potential issue in how the model handles controversial topics.

Transparency is another challenge. GPT-3.5 didn't provide explanations for its classifications, which could lead to user frustration and reduce trust in the moderation process. Users and moderators need to understand why certain content is flagged, especially when it comes to ambiguous cases.

Given these limitations, LLMs like GPT-3.5 should be part of a hybrid approach to moderation. While they can effectively flag potentially problematic content, human oversight is crucial to make final decisions, address biases, and provide more context, enhancing both fairness and transparency in content moderation.

IX. CONCLUSION

This study explored the use of GPT-3.5 in content moderation on social media, particularly focusing on Twitter. The model showed moderate effectiveness, with an overall accuracy of 76%, but it also demonstrated limitations. One significant concern is the presence of bias, as the model disproportionately flagged tweets mentioning minority groups and controversial topics, potentially reflecting biases from its training data.

The lack of transparency in GPT-3.5's decision-making was another issue, as the model provided no clear explanation for why certain content was flagged. This lack of interpretability could lead to frustration and reduced trust among users. These findings underscore that while LLMs can assist in content moderation, they are not a standalone solution.

To address these challenges, we recommend a hybrid approach that combines the strengths of LLMs with human oversight. Human moderators can handle complex or sensitive cases, ensure fairness, and provide the necessary context behind moderation decisions. By refining LLMs and integrating human judgment, we can work towards a more ethical and transparent content moderation system.

REFERENCES

[1]. J. Wu, M. Zhang, H. Sun, Y. Zhang, and X. Li, "Legilimens: Practical and unified content moderation for large language model services," ACM SIGSAC Conference, 2024.
[2]. P. Jha, S. Kumar, A. Bhatnagar, and R. Patel, "MemeGuard: An LLM and VLM-based framework for advancing content moderation via meme intervention," arXiv, 2024.
[3]. N. Vishwamitra, R. Gupta, T. Singh, and S. Nanda, "Moderating new waves of online hate with chain-of-thought reasoning in large language models," IEEE Symposium on Security and Privacy, 2024.
[4]. A. Bhatia, "Advancing policy insights: Opinion data analysis and discourse structuring using LLMs," University of Central Florida Thesis, 2024.
[5]. S. Ghosh, M. Verma, R. Choudhury, and T. Ahuja, "AEGIS: Online adaptive AI content safety moderation with ensemble of LLM experts," arXiv, 2024.
[6]. M. Franco, L. Rossi, S. Moreno, and G. Pérez, "Analyzing the use of large language models for content moderation with ChatGPT examples," OASIS, 2023.
[7]. T. Huang, "Content moderation by LLM: From accuracy to legitimacy," arXiv, 2024.
[8]. J. Cai, Y. Liu, Q. Zhao, and H. Lin, "Language evolution for evading social media regulation via LLM-based multi-agent simulation," IEEE, 2024.
[9]. N. P. Kumar, K. Srinivasan, and D. Ramesh, "Analyzing public sentiment towards LLM: A Twitter-based sentiment analysis," Proc. 2023 Int. Conf. Confluence Adv. Robotics, Vision, and Interdisciplinary Technology Management (IC-RVITM), IEEE, 2023.
[10]. P. Vanpech, K. Peerabenjakul, N. Suriwong, and S. Fugkeaw, "Detecting cyberbullying on social networks using language learning model," Proc. 2024 Int. Conf. Knowledge and Smart Technology (KST), IEEE, 2024.
[11]. H. T. Otal, E. Stern, and M. A. Canbaz, "LLM-assisted crisis management: Building advanced LLM platforms for effective emergency response and public collaboration," Proc. IEEE Conf. Artificial Intelligence (CAI), 2024.
[12]. M. Sadeghi, B. Egger, R. Agahi, R. Richer, K. Capito, and L. H. Rupp, "Exploring the capabilities of a language model-only approach for depression detection in text data," Proc. 23rd IEEE EMB Int. Conf. Biomedical and Health Informatics (BHI), 2023.
[13]. B. Saha and U. Saha, "Enhancing international graduate student experience through AI-driven support systems: A LLM and RAG-based approach," Proc. 2024 Int. Conf. Data Science and Its Applications (ICoDSA), 2024.
[14]. P. S. Ramteke and S. Khandelwal, "Comparing conventional machine learning and large-language models for human stress detection using social media posts," Proc. 2023 2nd Int. Conf. Futuristic Technologies (INCOFT), 2023.
[15]. K. Sabaneh, M. A. Salameh, F. Khaleel, M. M. Herzallah, J. Y. Natsheh, and M. Maree, "Early risk prediction of depression based on social media posts in Arabic," Proc. 2023 IEEE 35th Int. Conf. Tools with Artificial Intelligence (ICTAI), 2023.