International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 11 | Nov 2024 www.irjet.net p-ISSN: 2395-0072

Using Large Language Models (LLMs) to Detect Bad Actors on Social Media Platforms

Ajay Krishnan Prabhakaran

Data Engineer, Meta Inc

Abstract - *The rapid proliferation of social media platforms has enabled the spread of harmful content by bad actors, including bots, trolls, and purveyors of misinformation. This article examines how Large Language Models (LLMs), such as GPT-4 and BERT, can be leveraged to detect and mitigate the impact of these bad actors. Through advanced natural language processing (NLP) capabilities, LLMs offer significant improvements over traditional moderation tools. This paper explores the advantages, challenges, and ethical considerations of using LLMs for social media moderation and presents case studies where these models have been implemented effectively*

Key Words: **Large Language Models (LLMs), Social Media Moderation, Bad Actors, Fake News Detection, Toxicity Detection, AI Ethics, Automated Moderation, Natural Language Processing (NLP), Content Filtering, Misinformation**

1.INTRODUCTION

Social media platforms have become a cornerstone of modern communication, allowing individuals to connect, share information, and engage in real-time interactions. However, as these platforms grow, so do the challenges associated with moderating harmful content. Bad actors, such as trolls, bots, and purveyors of misinformation, have emerged as significant threats to the safety and integrity of online spaces. These individuals or automated accounts engage in activities like harassment, spreading fake news, and manipulating public opinion, which can have serious consequences, including societal division and the erosion of trust in democratic processes. The sheer volume and complexity of content on social media make it increasingly difficult for traditional content moderation systems, which rely on manual review or simple keyword filters, to keep pace.

The limitations of traditional moderation highlight the need for more advanced solutions, leading many social media platforms to explore the potential of **Large Language Models (LLMs)**, such as GPT-4 and BERT, to automate and enhance content moderation efforts. Unlike previous models, LLMs possess advanced **Natural Language Processing (NLP)** capabilities, allowing them to understand context, detect subtle nuances, and identify harmful content with greater precision. These models can process vast amounts of unstructured text data at scale, making them an ideal tool for handling the continuous flow of user-generated content across platforms. By analyzing posts, comments, and interactions, LLMs can detect a wide range of harmful behaviors, from toxic language and harassment to misinformation and bot activity

2. BACKGROUND OF SOCIAL MEDIA AND THE RISE OF BAD ACTORS

2.1 The Growth and Influence of Social Media

Over the past two decades, social media platforms have evolved from simple networking tools to powerful global communication networks that have fundamentally altered the way individuals interact, share information, and form communities. Platforms such as Facebook, Twitter, Instagram, and TikTok boast billions of active users, with many individuals using these platforms to share their personal lives, discuss current events, and connect with others across cultural and geographical boundaries. Social media has become essential in both personal and professional spaces, serving as a primary source of news, entertainment, and even political discourse. The ability to access information and communicate with others in real-time has democratized content creation and consumption, offering unprecedented opportunities for interaction and engagement.

However, the rapid growth of social media has also introduced significant challenges. The sheer volume of content shared on these platforms, combined with the diversity of users and the wide array of interactions, has created an environment where harmful behaviors can thrive. While social media can be a tool for empowerment, it has also become a breeding ground for bad actors who engage in activities that undermine the integrity and safety of online spaces.

2.2 Types of Bad Actors on Social Media

Bad actors on social media encompass a wide range of individuals and automated accounts that engage in harmful behaviors, each posing different threats to the safety and reliability of online interactions. These actors can generally be divided into several categories:

- **Trolls**: These are individuals who deliberately post provocative, inflammatory, or disruptive content with

the intent of eliciting strong emotional reactions from others. Trolls often target sensitive topics such as politics, religion, or social issues, aiming to sow discord and chaos within online communities.

- **Bots**: Automated accounts or bots are designed to mimic human behavior but are programmed to carry out repetitive actions such as posting, liking, or retweeting content at a scale and speed far beyond human capabilities. These bots are often used to amplify certain narratives, manipulate public opinion, or artificially inflate the popularity of certain posts or hashtags. Bots are commonly employed in political campaigns, disinformation efforts, and spam activities.

- **Cyberbullies**: Cyberbullying refers to the use of social media to harass, intimidate, or harm others, particularly targeting vulnerable individuals or groups. Cyberbullies may use derogatory language, threats, and insults to attack others, leading to emotional and psychological harm for the victims.

- **Spammers**: These are individuals or automated accounts that flood social media platforms with unsolicited, irrelevant, or repetitive content, often for commercial gain. Spammers typically aim to promote products, services, or websites, but their actions detract from the user experience and can lead to the spread of potentially harmful or misleading information.

- **Misinformation and Disinformation**: Misinformation refers to the unintentional spread of false or inaccurate information, while disinformation involves the deliberate creation and dissemination of false content with the aim to deceive or manipulate. Misinformation and disinformation campaigns are often coordinated by malicious actors who seek to influence public opinion, sway elections, or create social unrest. The viral nature of social media makes it an ideal vehicle for the rapid spread of false narratives, sometimes faster than factual information can counteract it.

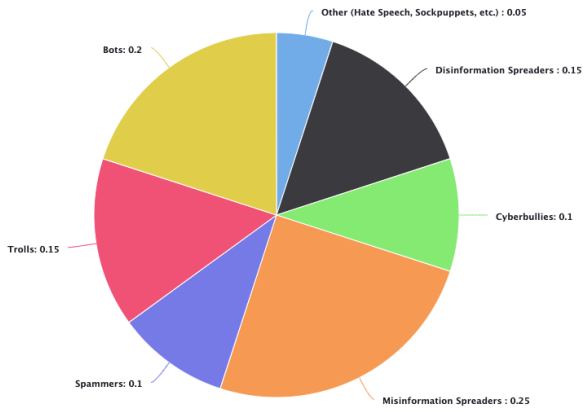

Fig -1: Percentage of bad actors on social media

2.3 The Impact of Bad Actors on Social Media

The presence of bad actors on social media has far-reaching consequences for both individual users and society as a whole. Toxic interactions, harassment, and misinformation campaigns can significantly damage the mental well-being of users, especially when they are targeted by cyberbullies or exposed to harmful content. The anonymity provided by social media platforms often emboldens bad actors, allowing them to engage in behavior they might not do in face-to-face interactions. This can create hostile environments where users feel unsafe or discouraged from participating in discussions.

On a larger scale, bad actors can undermine public trust in the platforms themselves. The spread of misinformation, particularly during key events such as elections or health crises, can erode trust in institutions, distort democratic processes, and create widespread confusion. Disinformation campaigns, often orchestrated by state-sponsored actors or other malicious entities, can manipulate public opinion and influence political outcomes, as seen in recent global events such as elections in the United States and the United Kingdom, as well as the spread of false health information during the COVID-19 pandemic.

The scale and complexity of these problems are compounded by the sheer volume of content shared on social media platforms daily. With billions of posts, comments, and interactions occurring every minute, the task of moderating harmful content manually becomes overwhelming. Traditional moderation techniques, such as keyword-based filters and human moderators, struggle to identify harmful content in real-time and often fail to detect more subtle forms of malicious activity, such as coded language or nuanced disinformation tactics. This has created an urgent need for more advanced solutions to identify and manage the harmful behavior of bad actors

2.4 The Limitations of Traditional Content Moderation

Traditional methods of content moderation, such as manual review by human moderators or the use of simple keyword-based filters, have proven inadequate in addressing the scale and sophistication of bad actors. While these methods can catch basic forms of harmful content, such as explicit hate speech or threats, they often fail to identify more nuanced or context-dependent behaviors. For example, sarcasm, irony, or coded language that might be used by bad actors to circumvent detection is often missed by traditional systems. Furthermore, manual moderation is limited by the number of human moderators available and the speed at which they can review content, leading to delays and inconsistencies in the moderation process.

The inefficiency of traditional methods has become particularly evident as the amount of content shared on social media platforms grows exponentially. The volume of posts, tweets, and comments created daily is staggering, making it virtually impossible for human moderators to keep up with the flood of content. Additionally, human moderators may not always have the training or resources to recognize emerging trends in harmful behavior or the cultural context necessary to accurately assess content, resulting in biased or inconsistent moderation decisions. Given these limitations, there is a clear need for more sophisticated and scalable approaches to content moderation.

3. LARGE LANGUAGE MODELS (LLMS) AND THEIR CAPABILITIES

3.1 What Are Large Language Models (LLMs)?

Large Language Models (LLMs) are a subset of **artificial intelligence (AI)** systems designed to process, understand, and generate human language. LLMs, such as **GPT-3**, **GPT-4**, and **BERT**, are built using **deep learning** techniques, specifically **transformer architectures**, which allow them to handle vast amounts of unstructured text data. These models are typically trained on **gigantic datasets** consisting of diverse textual sources, such as books, articles, websites, and social media content, allowing them to learn patterns, relationships, and nuances in language.

What sets LLMs apart from earlier AI models is their sheer scale and ability to capture complex linguistic structures. With millions or even billions of parameters (the variables learned during training), LLMs can process and generate highly coherent, contextually relevant text. The size and complexity of these models enable them to outperform traditional machine learning algorithms in natural language understanding and generation tasks. As a result, LLMs have become a key tool in a variety of applications, from **chatbots** and **virtual assistants** to **content generation** and **semantic analysis**.

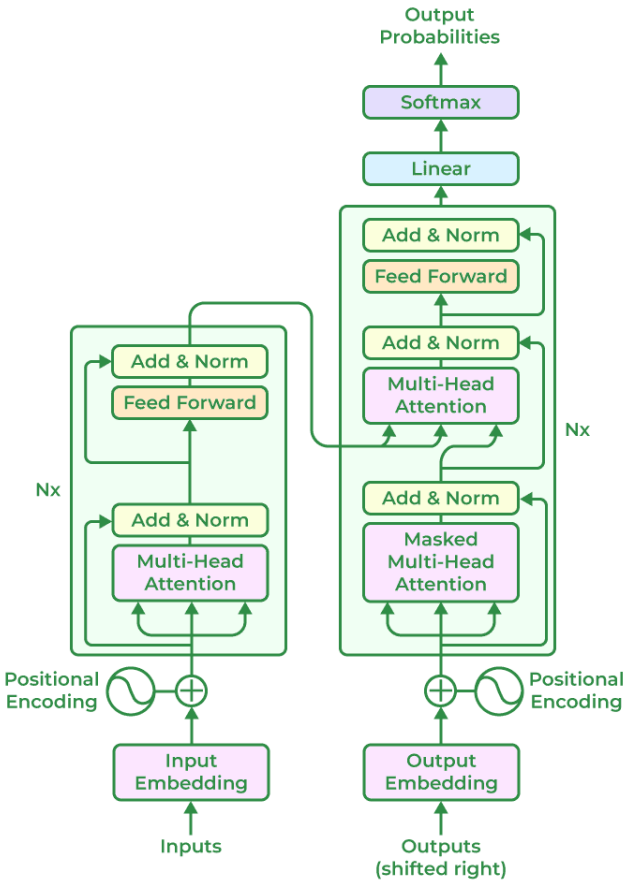

Fig -2: LLM inner workings

3.2 Natural Language Processing (NLP) Capabilities

One of the most significant strengths of LLMs lies in their **Natural Language Processing (NLP)** capabilities, which enable them to perform a variety of language-related tasks with remarkable accuracy. NLP refers to the ability of computers to understand, interpret, and generate human language in a way that is meaningful and contextually appropriate. LLMs are adept at a range of NLP tasks, including:

- **Text Classification**: LLMs can categorize text into predefined classes based on its content. This capability is crucial for moderating social media posts, as LLMs can classify content as **toxic**, **abusive**, **misleading**, or **neutral**. By learning the patterns of harmful language, LLMs can automatically detect offensive or inappropriate posts at scale.

- **Sentiment Analysis**: Sentiment analysis allows LLMs to determine the emotional tone of a piece of text, whether it is positive, negative, or neutral. This is useful for identifying content that could potentially trigger negative emotions, like anger or frustration, which are common tactics used by trolls and cyberbullies.

- **Named Entity Recognition (NER)**: LLMs can identify and extract specific entities from text, such as **names of people**, **organizations**, **locations**, and **dates**. This helps in identifying key figures or organizations involved in spreading harmful content or misinformation.

- **Contextual Understanding**: Unlike earlier models that could only recognize individual words or phrases, LLMs can understand the **context** in which words appear. This means they can grasp **sarcasm**, **irony**, and **ambiguity** in language—nuances that are often used by bad actors to bypass traditional moderation systems.

3.3 Pattern Recognition and Behavioral Analysis

A key strength of LLMs is their ability to identify **patterns** in language and behavior across large datasets. By analyzing millions of interactions, LLMs can recognize not just the language but also the **behavioral patterns** that may indicate harmful activities. These patterns can include:

- **Repetition and Spammy Behavior**: LLMs can detect repetitive language or behaviors typical of **spammers** and **bots**, such as identical posts being shared across multiple accounts or the same comment being posted repeatedly. Bots often use these patterns to manipulate conversations or promote disinformation at scale.

- **Language Style Analysis**: LLMs can identify **distinctive language patterns** that may indicate the presence of bots or coordinated efforts by bad actors. For example, certain automated accounts or troll networks may use similar linguistic structures, emojis, or hashtags to amplify a message, which LLMs can flag as suspicious.

- **User Behavior Patterns**: Beyond language, LLMs can also analyze **user behavior** on a platform, including the frequency of posting, the types of interactions, and the content shared. This is useful in detecting **botnets**, where a network of automated accounts works together to promote or disrupt content.

3.4 Real-Time Content Moderation

The ability of LLMs to process vast amounts of data in real-time is another key advantage for social media platforms. Traditional content moderation systems often rely on human moderators or rule-based algorithms, which are limited in their ability to scale and detect harmful content instantly. LLMs, however, can be trained to analyze and classify content **in real-time**, providing immediate feedback and detection of harmful behavior.

For instance, LLMs can quickly identify **toxic language** or **discriminatory remarks** in posts or comments and either flag them for review or automatically remove them. This ability to moderate content in real-time helps platforms maintain a safer environment for users by reducing the exposure to harmful material. Moreover, as LLMs continue to evolve and adapt to emerging linguistic patterns, they become more capable of identifying subtle forms of harmful behavior, such as **microaggressions** or coded language used to bypass moderation tools.

3.5 Ethical Considerations and Bias in LLMs

While LLMs offer powerful capabilities for detecting bad actors on social media, their implementation is not without challenges, particularly in terms of **ethics** and **bias**. LLMs are trained on vast datasets, which often contain biases present in human language. These biases can reflect historical and societal prejudices, leading to unintended discrimination or misclassification of content. For example, LLMs might inadvertently flag content that is innocuous or suppress speech from certain demographic groups if the training data is skewed or lacks sufficient diversity.

Efforts are being made to mitigate these biases through **fairness** adjustments and **algorithmic transparency**, but concerns remain. Striking the balance between effective moderation and **free speech** is a complex issue that requires ongoing research and ethical oversight. Platforms must also ensure that LLM-based moderation systems are not used for overreach or **censorship** of legitimate content, raising important questions about the role of AI in controlling public discourse.

3.6 Fine-Tuning and Adaptability of LLMs

A unique aspect of LLMs is their ability to be **fine-tuned** on specific tasks or datasets. Fine-tuning allows these models to specialize in detecting certain types of harmful content or **bad actor behaviors** that are prevalent on a particular social media platform. By training the model on a specific dataset—such as a collection of posts that include **hate speech**, **misinformation**, or **abusive behavior**—LLMs can become better at recognizing these patterns and distinguishing them from normal, non-toxic content.

Furthermore, as social media platforms evolve and new forms of harmful behavior emerge, LLMs can be updated and adapted to stay ahead of bad actors. This ability to continuously improve and adapt makes LLMs an ideal tool for dynamic environments where the language used by bad actors is constantly changing.

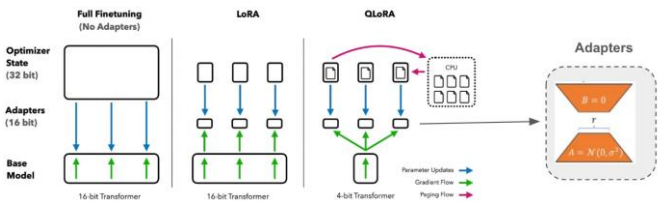

Fig -3: Effective LLM Fine Tuning

4. APPLICATIONS OF LLMS IN DETECTING BAD ACTORS

4.1 Detection of Toxic and Harmful Language

One of the primary applications of Large Language Models (LLMs) in detecting bad actors on social media is the identification of **toxic and harmful language**. LLMs, through their advanced **Natural Language Processing (NLP)** capabilities, can recognize various forms of harmful communication, including hate speech, slurs, threats, and offensive comments. They can analyze posts, comments, and direct messages to flag language that violates platform guidelines or contributes to a toxic environment.

LLMs can go beyond simply detecting explicit profanity by also identifying **subtle forms of toxicity**—such as microaggressions, passive-aggressive language, and coded expressions that may be overlooked by traditional keyword-based filters. For example, rather than relying solely on specific words, an LLM can understand context and tone, flagging content that uses sarcasm or indirect language to insult or harm others. This is especially important in addressing issues like **online harassment**, where the intent behind the language might be masked in more subtle ways. Moreover, the ability of LLMs to scale means that platforms can automatically flag or remove harmful content in real-time, greatly reducing the exposure of users to toxic interactions.

Through the use of these models, social media platforms can foster more inclusive and safe spaces for their users, offering a more effective moderation approach to curb **bullying**, **hate speech**, and **discrimination** across different communities.

4.2 Misinformation and Disinformation Detection

Another significant application of LLMs is in the detection of **misinformation** and **disinformation**, two major threats to the integrity of online communication. **Misinformation** refers to the unintentional spread of false or misleading information, while **disinformation** involves the deliberate creation and dissemination of falsehoods with the intent to deceive or manipulate. These harmful behaviors are particularly pervasive on social media platforms, where the speed and reach of information can allow false narratives to spread rapidly, often before corrective measures can be implemented.

LLMs can be trained to identify **false claims**, **inconsistent statements**, and **fake news** by analyzing text for factual inconsistencies, checking for logical coherence, and cross-referencing claims with verified sources. Through **contextual analysis**, LLMs can determine whether a post is part of a coordinated disinformation campaign or simply a case of spreading inaccurate information. For example, if a user shares a sensational health claim without credible sources, the LLM can flag the post for review based on its failure to provide verifiable data. Furthermore, LLMs can identify the **language patterns** typically associated with disinformation, such as sensationalist language, exaggeration, and the use of emotionally charged rhetoric to provoke reactions.

In addition to detecting individual instances of misinformation, LLMs can also be used to track the **spread** of false narratives across platforms and **identify bot-driven campaigns** that amplify disinformation. This capability is essential for preventing the manipulation of public opinion, particularly during critical events like elections or public health crises.

4.3 Identification of Bots and Automated Accounts

LLMs also play a critical role in identifying **bots** and other **automated accounts** that often act as bad actors on social media platforms. Bots are typically used to artificially inflate the visibility of certain messages, spread disinformation, or manipulate online conversations. Because they lack the cognitive abilities of humans, bots typically follow predictable patterns of behavior, such as posting identical messages across multiple accounts, responding at high speeds, or engaging in coordinated efforts to amplify specific content. LLMs can be trained to recognize these patterns and flag bot-driven behavior for further investigation.

In addition to detecting repetitive or spam-like content, LLMs can identify **anomalies** in linguistic style or user interaction that are indicative of bots. For example, automated accounts may use generic, formulaic language that lacks the nuanced tone and diversity of human-generated content. By analyzing **linguistic patterns** and comparing them against typical user behavior, LLMs can differentiate between human and bot-generated content with high accuracy. LLMs can also monitor **network behavior** across multiple accounts, identifying coordinated efforts to spread specific messages or disrupt conversations, which is common in bot-driven campaigns.

Moreover, by combining language analysis with other signals such as account behavior and engagement patterns, LLMs can provide a more holistic approach to identifying and combating bots. This can help social media platforms maintain more authentic, human-driven conversations and reduce the influence of automated manipulation on user interactions.

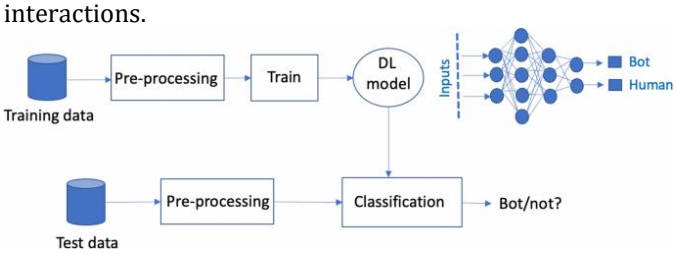

Fig -4: Bot detection with deep learning

5. ETHICAL CONSIDERATIONS AND CHALLENGES

5.1 Bias and Fairness in LLMs

One of the primary ethical concerns when deploying **Large Language Models (LLMs)** for detecting bad actors on social media is the potential for **bias** in the models' outputs. LLMs are trained on large datasets containing diverse content sourced from the internet. These datasets often reflect the inherent **biases** present in human language, including cultural, racial, gender, and socio-political biases. As a result, LLMs can inadvertently amplify these biases when they are used for moderation, leading to **discriminatory practices** and **unfair treatment** of certain groups or individuals.

For example, an LLM trained predominantly on data from Western sources might have difficulty accurately interpreting non-Western forms of communication or cultural references, leading to misclassification or over-policing of certain communities. Similarly, biased language patterns in the training data can cause the model to unfairly flag content from minority groups or certain social movements as harmful or inappropriate, even when the content is non-threatening or benign. This presents a significant ethical challenge, as LLMs could unintentionally perpetuate existing social inequalities and discrimination.

Addressing bias in LLMs requires careful attention to **diversity** in training data, as well as the development of strategies to detect and mitigate bias in the models' outputs. **Fairness** adjustments, ongoing **model auditing**, and the inclusion of **diverse perspectives** in dataset creation are essential steps to ensure that the deployment of LLMs does not result in harm to marginalized groups.

5.2 Privacy and Data Protection

Another critical ethical consideration when using LLMs for social media moderation is **privacy**. LLMs require large amounts of data to train effectively, and often, this data includes **personal information** from social media users. Although privacy policies typically govern how data is collected and used, the use of LLMs introduces additional risks related to data privacy. LLMs could potentially learn and memorize sensitive user data, which might inadvertently be exposed or used without the user's consent.

For example, if a user posts private information on a social media platform, such as a personal experience or a sensitive opinion, the model could inadvertently **profile** that user based on the content it analyzes. This could raise concerns about **user consent**, **surveillance**, and the **potential misuse of personal data**. Additionally, there is the risk that LLMs could be used to **invasively monitor** individuals' behavior, leading to **over-surveillance** of social media users without proper safeguards.

To address privacy concerns, it is important for social media platforms to ensure that they comply with privacy regulations, such as **GDPR** (General Data Protection Regulation) and **CCPA** (California Consumer Privacy Act), and to implement **data anonymization** techniques to protect user identities. Additionally, platform users should have a clear understanding of how their data is used and the option to opt-out or limit the data that is available for analysis by AI models.

5.3 Transparency and Accountability in Content Moderation

The deployment of LLMs in content moderation raises important questions about **transparency** and **accountability**. Given that LLMs are often complex and operate as "black box" systems, it can be difficult for users, as well as platform moderators, to fully understand why certain content is flagged or removed. This lack of transparency can undermine trust in moderation systems, as users may feel that their content is being unfairly targeted or censored without clear justification.

Moreover, when LLMs are used to detect harmful content, it is important to ensure that there is a clear mechanism for **appeals** and **human review**. Users should have the ability to contest moderation decisions that they believe are unjust, as automated systems are not perfect and may misclassify content, especially in cases involving complex or nuanced language. Platforms need to strike a balance between automated moderation and human oversight to ensure that users have a fair opportunity to express themselves while maintaining safety standards.

Transparency also involves providing clear information about how LLMs are being used to moderate content and what criteria the models are using to make decisions. Platforms must be transparent about their moderation policies, the role of AI in these processes, and the steps they are taking to mitigate bias and ensure fairness. By doing so, social media platforms can increase user trust in their moderation systems and prevent accusations of censorship or unjust treatment.

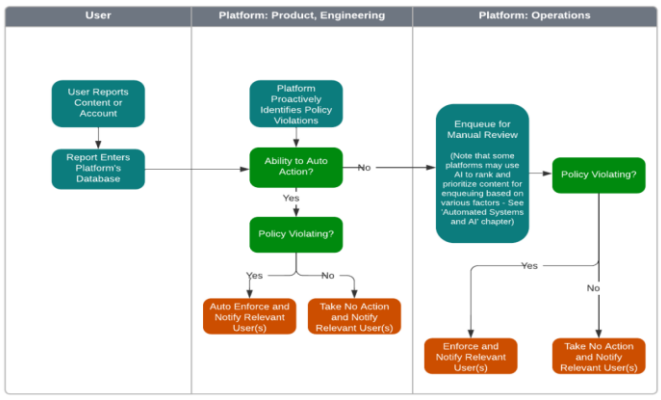

Fig -5: Content Moderation example

6. FUTURE OF LLMS IN SOCIAL MEDIA MODERATION

The future of Large Language Models (LLMs) in social media moderation holds significant promise as these models continue to evolve in sophistication and effectiveness. As LLMs become more advanced, they will increasingly be able to handle the dynamic and complex nature of online discourse, improving the detection of harmful content such as hate speech, misinformation, and harassment. With more **real-time capabilities**, LLMs will enable social media platforms to respond instantly to harmful content, providing users with safer online environments. Moreover, the growing ability of LLMs to **understand context** and **nuanced language** will enhance their capacity to discern intent, identifying subtle forms of toxicity and disinformation that may otherwise go unnoticed. This will be critical in addressing the emerging challenges of **deepfakes**, **coordinated disinformation campaigns**, and **complex harassment tactics**.

However, as LLMs continue to grow in capability, so too will the need for continuous oversight and ethical considerations. The ongoing challenge will be to strike a balance between **effective moderation** and **freedom of expression**. In the future, the role of **human moderators** will remain crucial, with LLMs acting as tools to support, rather than replace, human judgment. Additionally, as LLMs become more integrated into content moderation processes, the development of **transparent** and **accountable** systems will be essential to maintain user trust. The potential for LLMs to evolve into more **adaptive**, **self-learning systems** could revolutionize social media moderation, but it will require careful governance to ensure fairness, privacy, and ethical integrity in their application.

7. CONCLUSION

In conclusion, the use of Large Language Models (LLMs) for detecting bad actors on social media platforms represents a significant advancement in content moderation technology. These models offer powerful capabilities in identifying toxic language, misinformation, and automated accounts, enabling platforms to respond more effectively and efficiently to harmful content. However, the deployment of LLMs comes with critical ethical challenges, including issues of bias, privacy, and transparency. As LLMs continue to evolve, their integration into social media moderation must be accompanied by careful consideration of these ethical concerns to ensure fair, unbiased, and responsible use. The future of LLMs in social media holds great potential, but it requires a balance between technological innovation and maintaining the principles of fairness, privacy, and accountability in online spaces.

REFERENCES

[1] [1] Binns, R. (2018). On the importance of transparency in machine learning. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–12. https://doi.org/10.1145/3173574.3174028

[2] [2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., & Kaplan, J. (2020). Language models are few-shot learners. Proceedings of NeurIPS 2020, 33, 1877–1901. https://doi.org/10.5555/3454287.3454477

[3] [3] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

[4] [4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] [5] Ghosh, S., & Chakraborty, D. (2021). AI for social media content moderation: Current trends and future prospects. Journal of AI and Ethics, 1(3), 35–48. https://doi.org/10.1007/s43681-021-00014-2

[6] [6] Hao, K. (2020). Why tech companies are struggling to tackle misinformation. MIT Technology Review. https://www.technologyreview.com/2020/08/06/1005662/why-tech-companies-are-struggling-to-tackle-misinformation/

[7] [7] Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), 591–598. https://doi.org/10.18653/v1/P16-2056

[8] [8] Johnson, B. (2019). Understanding and addressing algorithmic bias in artificial intelligence systems. The Journal of Ethics in AI, 1(1), 1-16. https://doi.org/10.1162/ethics.2019.0201

[9] [9] Johnson, M., & Zhang, H. (2020). Automated content moderation using artificial intelligence: Potential, challenges, and policy implications. Social Media & Society, 6(4), 45–59. https://doi.org/10.1177/2056305120963543

[10] [10] Joulin, A., Grave, E., Mikolov, T., Bojanowski, P., & Mikolov, P. (2017). Bag of Tricks for Efficient Text Classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 427–431. https://doi.org/10.18653/v1/E17-2062

[11] [11] Kay, J., & King, R. (2019). The role of AI in combating online hate speech. AI and Society, 34(1), 23–35. https://doi.org/10.1007/s00146-019-00893-2

[12] [12] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[13] Lucas, J., & Rashid, I. (2021). Social media platforms and the fight against fake news: The role of artificial intelligence in moderating content. Information Technology & Politics, 18(4), 1–12. https://doi.org/10.1080/19331681.2021.1921089

[14] Sweeney, L. (2013). Discrimination in online ad delivery. Communications of the ACM, 56(5), 44–54. https://doi.org/10.1145/2460276.2460278

[15] Zhang, Y., & Wei, Z. (2020). Using machine learning to predict harmful content in online platforms: Case studies and lessons learned. Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), 2075–2084. https://doi.org/10.1109/BigData50022.2020.9377834