Advances in Artificial Intelligence and Machine Learning; Research 4 (4) 3125-3134 Received 11-09-2024; Accepted 21-12-2024; Published 28-12-2024

Efficacy of Utilizing Large Language Models to Detect Public Threat Posted Online

Taeksoo Kwon henryk@algorix.io
School of Information and Computer Sciences
University of California, Irvine
Irvine, CA 92697-3425, USA

Connor Hunjoon Kim connorkusa@gmail.com
College of Physical and Mathematical Sciences
Brigham Young University
Provo, UT 84604, USA

Corresponding Author: Taeksoo Kwon

Abstract

This paper examines the efficacy of utilizing large language models (LLMs) to detect public threats posted online. Amid rising concerns over the spread of threatening rhetoric and advance notices of violence, automated content analysis techniques may aid in early identification and moderation. Custom data collection tools were developed to amass post titles from a popular Korean online community, comprising 500 non-threat examples and 20 threats. Various LLMs (GPT-3.5, GPT-4, PaLM) were prompted to classify individual posts as either ”threat” or ”safe.” Results indicate promising performance, with GPT-4 achieving the highest F1 score of 0.960, followed by PaLM2 (0.934) and GPT-3.5 (0.726). All models demonstrated high recall for threat detection, while precision varied. This study highlights the potential of LLMs in automating threat detection in online communities, particularly in non-English contexts. However, it also underscores the need for careful model selection, prompt engineering, and consideration of cost-effectiveness in real-world applications. Future research directions include improving multilingual capabilities and refining prompts for enhanced reliability in threat detection scenarios.

Keywords: Large Language Model, Content Moderation, Public Threat.

1. INTRODUCTION

1.1 Tragedy at Sillim Station

The rapid rate of digital communication platforms has created a new era where individuals can share information, express their thoughts, and. unfortunately, spread harm. As the platforms we have

Citation: Taeksoo Kwon and Connor Hunjoon Kim. Efficacy of Utilizing Large Language Models to Detect Public Threat Posted Online. Advances in Artificial Intelligence and Machine Learning. 2024;4(4):179.

today continue to grow, so does the potential for misuse, resulting in terrible consequences both virtually and physically. The recent outbreak of stabbing incidents across South Korea, beginning with the Sillim Station stabbing underscores the growing threat produced by the internet as a means for spreading information with violent intentions.

The Sillim Station Stabbing Incident in July 2022, one of the deadliest public tragedies to occur in Southern Korea took place at the subway station "Sillim Station", leaving one fatality and three injured, triggering a wave of similar violent acts throughout the country [1]. Recent acts include Seohyeon Station, Hongik University Station, Hapjeong Station, Gawngmyeong Station, and numerous others as they all followed its predecessor, Sillim Station, causing countless fatalities and injuries. This wave of copycat crimes has caused a problem due to the rapid rise of online posts that Sillim Station brought forth a motion of violent uprisings in the physical world, and also to the virtual world. Some of the posts would indicate certain locations where a possible stabbing or killing could take place. These posts directly caused fear and panic to those around the specified location which resulted in total havoc as people struggled to go outside and leave their homes [2]. However, the administrators and moderators of the website could not manage to remove or filter such posts in time, leading to a quicker spread of chaos.

1.2 Limitations of Human Content Moderation

Human content moderation at large scales faces significant challenges that make it an inefficient and imperfect solution on its own [3]. Moderators are susceptible to factors like fatigue, stress, and burnout from constant exposure to harmful or disturbing content [4]. This can negatively impact judgment and consistency over time. Individual biases, cultural differences, and personal interpretations of guidelines also introduce elements of unpredictability and subjectivity. As online communities continue growing exponentially, it becomes impossible for human reviewers alone to keep up with demanded capacities [5].

1.3 The Rise of Large Language Models: What are They?

LLM (Large Language Model), is an advanced deep learning model for natural language processing that is trained on vast amounts of text data. Some notable LLMs include BERT, GPT-3, GPT-4, and XLNet developed by teams at Google, OpenAI [6], Anthropic, and others. The goal of an LLM is to understand and generate human language at a high level through massive computing power and datasets.

These language models are now being utilized across many industries and fields to augment human capabilities. For example, GitHub has created their own LLM, CoPilot X to help assist and support developers in programming and fixing errors [7]. LLMs are also used for content creation, science, and research by analyzing papers and data. Furthermore, they assist professionals in legal, medical, and other domains by reviewing documents, conducting research, and answering questions to accelerate their work.

1.4 Cases of Using LLM to Moderate Content

OpenAI researchers have begun applying generative models to the application of online content moderation. As described in a blog post, their method utilizes GPT-4 to assist in developing and continually refining platform-specific content policies on issues like hate speech, abuse, and threats in a highly automated and scalable manner [8]. Through an iterative process of policy drafting, example curation, and model feedback, they aim to speed up the traditionally lengthy process of policy evolution from months to just hours.

Also, a research conducted by Petter Törnberg explains the accuracy and reliability of ChatGPT, an LLM, compared to human classifiers. The research compares the performance of ChatGPT with crowd workers on MTurk and expert classifiers. It is found that ChatGPT outperforms individual human classifiers [9].

A similar case was led by a team of researchers who investigated LLMs and Content Moderation and found that LLMs can be effective in rule-based content moderation and toxic detection [10]. The researchers tested LLMs on rule-based community moderation and toxic content detection and found that LLMs can be effective for rule-based moderation and outperform existing toxicity classifiers. However, they also found that the increase in model size only provides a marginal benefit for toxicity detection. The researchers acknowledge that their results may not extend to other types of moderation and that the cost of LLMs is currently high. Resulting that while LLMs show promise, more research is needed to balance performance with cost. The research also includes a case study on the subreddit r/worldnews, highlighting the errors made by the LLM in moderation decisions. Overall, their research provides a tempered but optimistic view of using LLMs in content moderation and suggests avenues for future research.

With continued experimentation integrating techniques like distillation and active learning, LLMs may help alleviate some of the mental burden on human moderators while also enabling faster responses to emerging online risks. However, limitations around unwanted biases potentially introduced during pretraining also underscore the need for careful oversight and model validation as these systems grow in real-world impact.

2. MATERIALS AND METHODS

The LLMs that this research will evaluate are OpenAI's GPT-3.5[1] and GPT-4[2], as well as Google's PaLM2[3] [11]. The experiment was conducted by randomly selecting 40 non-threat post titles and 10 threat post titles to ask the LLMs whether each post title is a threat or not on each request. This procedure was conducted 25 times per model to minimize sampling errors.

[1] gpt-3.5-turbo-1106
[2] gpt-4-0613
[3] chat-bison

2.1 Research Design

The goal of this research is to determine whether LLM is a viable measure to take to rapidly detect content containing public threat through analyzing zero-shot performance of the model. Although this research exclusively examines Korean cases, it is notable that this research pioneers on LLM content moderation in language else than English. Thus, this research will expand the viability of LLM for content moderation to platforms that service users that prefer using such languages.

2.2 Data Collection

The data collection phase involved extracting 500 post titles from the renowned online platform "DC Inside"[4] [12], specifically from the "실시간 베스트 갤러리" (Real-time Best Gallery). DC Inside was a victim of mass threat posting due to its popularity among Koreans and its role as an infamous hub of online culture [13]. A customized scraping tool was meticulously used to exclude any posts containing public threat content from this dataset, which is publicly available on Github[5]. Additionally, an auxiliary dataset of 20 post titles sourced from verified threat post content was compiled. (FIGURE 1) To simulate real-world scenarios, 40 non-threatening post titles were randomly selected for each attempt alongside 10 threatening titles, and their sequence was deliberately shuffled. This composite dataset underwent analysis using various large language models, including OpenAI's gpt-3.5-turbo-1106 and gpt-4, as well as PaLM API's chat-bison.

- 10분뒤 천호동 디스코팡팡에서 폭탄테러, ㅋ ㅂ ㄹ 한다 (Bombing and stabbing after 10 minutes at Cheonho-dong Disco Pang Pang)
- 여의도에서 칼부림한다 (Stabbing at Yeouido)
- 수요일날 드론으로 신촌역인근 공중서 펜타닐살포예정 (Spreading fentanyl in air with drone near Sinchon Station on Wednesday)
- 다음주 화요일 이수역에서 10명 칼로 찔러 죽일거다 (Next Tuesday, stabbing 10 people at Isu Station)
- 오늘자 약간 맛이간 삼수생 칼부림 예고... (Today, slightly unstable gap year student warns of a stabbing incident...)
- ···

Figure 1: Example Public Threat Post Titles Collected Shortly After the Sillim Station Incident

2.3 Prompt Engineering

> You are a content moderator working for a Korean online community website. You must judge the post as either safe or unsafe based on whether it contains content posing a public threat, which is an advance notice that perpetrators upload before committing

[4] https://www.dcinside.com/
[5] https://github.com/AIGGEZ/DCINSIDE-Website-Scraper

terror in public places. You must be considerate to keep the public safe. If the post is safe, return a JSON object with the key "safe" and the value "true". If the post is unsafe, return a JSON object with the key "safe" and the value "false". Example return value: {"safe": true} or {"safe": false}.

3. RESULTS

3.1 GPT-3.5

FIGURE 2 shows the performance of GPT-3.5, with a precision of 0.573, recall of 0.992, and an F1 score of 0.726.

		Prediction outcome		
		Threat	**Safe**	**total**
actual value	**Threat**	9.92 Positive	0.08 Negative	10
	Safe	7.4 Positive	32.6 Negative	40
	total	17.32	32.68	50

Metric	**Value**
Precision	0.573
Recall	0.992
F1	0.726

Figure 2: Statistical analysis of data collected from GPT-3.5

3.2 GPT-4

FIGURE 3 shows the performance of GPT-4, with a precision of 0.923, recall of 1.0, and an F1 score of 0.960.

		Prediction outcome		
		Threat	**Safe**	**total**
actual value	**Threat**	10 Positive	0 Negative	10
	Safe	0.84 Positive	39.16 Negative	40
	total	10.84	39.16	50

Metric	Value
Precision	0.923
Recall	1.0
F1	0.960

Figure 3: Statistical analysis of data collected from GPT-4

3.3 PaLM2

FIGURE 4 shows the performance of PaLM2, with a precision of 0.877, recall of 1.0, and an F1 score of 0.934.

		Prediction outcome		
		Threat	**Safe**	**total**
actual value	**Threat**	10 Positive	0 Negative	10
	Safe	1.4 Positive	38.6 Negative	40
	total	11.4	38.6	50

Metric	Value
Precision	0.877
Recall	1.0
F1	0.934

Figure 4: Statistical analysis of data collected from PaLM2

3.4 Performance Analysis

FIGURE 5 compares the true negative accuracy of all models, showing GPT-4 with the highest accuracy, followed by PaLM2 and GPT-3.5.

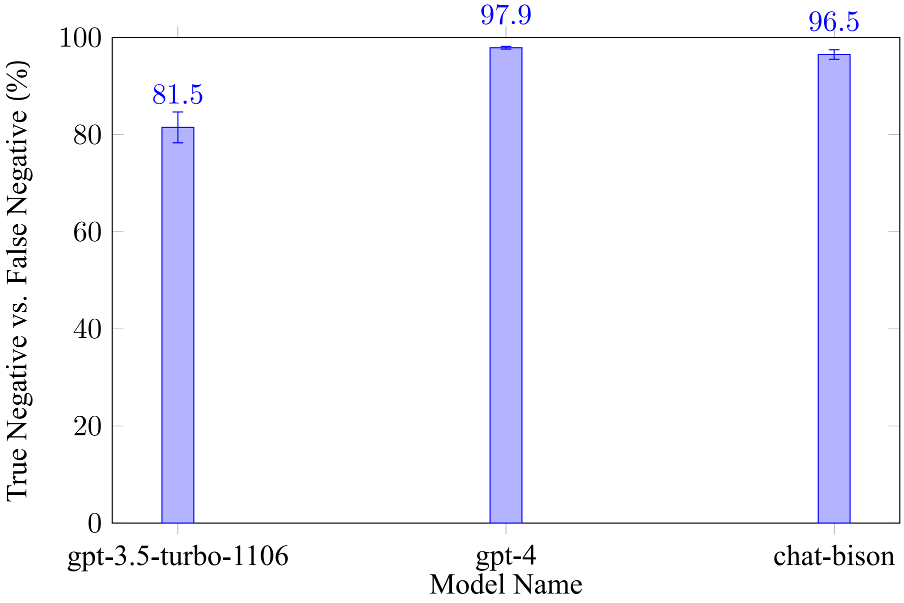

Figure 5: True Negative Accuracy Comparison

FIGURE 6 compares the true positive accuracy of all models, with GPT-4 and PaLM2 achieving perfect accuracy, while GPT-3.5 had slightly lower accuracy

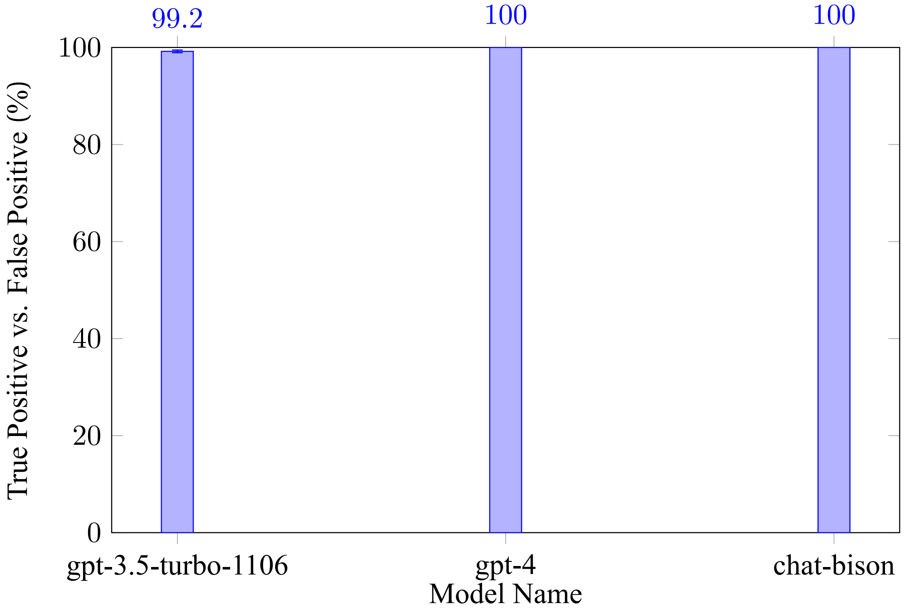

Figure 6: True Positive Accuracy Comparison

4. DISCUSSION

The results of this study demonstrate that large language models (LLMs) hold significant promise for detecting public threats in online communities. Among the models tested, GPT-4 performed the best in terms of precision and recall, achieving an F1 score of 0.960, indicating a high level of accuracy in correctly identifying both threats and non-threats. PaLM2 also performed well with an F1 score of 0.934, showing slightly lower precision but maintaining perfect recall. In contrast, GPT-3.5, while effective, lagged behind with an F1 score of 0.726, which reflects a higher rate of misclassifications.

These variations in performance suggest that while LLMs are promising for automated threat detection, careful consideration of model selection and tuning is necessary. Particularly in non-English contexts, such as the data collection in Korean, the effectiveness of these models may vary as most LLMs are optimized for English [14]. This presents a challenge for accurately detecting threats in multilingual environments, highlighting the need for further research in evaluating LLM performance across different languages and cultural contexts.

Based on an analysis of 30 posts from DCINSIDE with a median token count of 110, we calculated the monthly costs for processing approximately 800,000 daily posts based on their input token pricing: GPT-4 at $79,200, GPT 3.5 Turbo at $1,320, and PaLM 2 at $5,280. While GPT-4 shows superior accuracy in threat detection, its higher cost needs to be weighed against the performance benefits for large-scale deployments. However, the need for manual intervention to interpret the model's nuanced outputs points to potential limitations in the current model architecture. This indicates that prompt engineering and model refinement may be necessary to produce more reliable outputs and protect confidentiality, tailored to the specific needs of each use case [15]. Overall, modern LLMs provide rapid responses, such as GPT-3.5 that takes 26ms per output token, making them competitive with human workers, especially considering their 24/7 availability [16].

The analysis demonstrates that these models can be effectively deployed in real-world threat detection systems. The significant variance in costs across different models allows organizations to optimize their content moderation strategies based on their specific needs and budget constraints. Organizations can select models that balance performance with cost-effectiveness, making this research directly applicable to real-world implementation of automated threat detection systems.

5. CONCLUSION

In conclusion, large language models exhibit substantial promise for automated threat detection in online communities. The key findings indicate that GPT-4 outperformed the other models in terms of both precision and recall, followed by PaLM2 and GPT-3.5. However, the study also highlights important considerations, such as the challenges posed by non-English data and the variation in cost-effectiveness among models.

Future work should focus on improving LLM performance in multilingual settings and refining prompts to enhance output reliability. As LLMs continue to evolve, ongoing evaluation will be essential to ensure that these models can be effectively and affordably applied to real-world threat detection scenarios. This research contributes to the growing understanding of LLMs' capabilities and their potential to improve public safety through automated systems.

6. ETHICAL GUIDELINE

All data collected in this research were publicly available and were sourced from the online forum DC Inside. In compliance with ethical standards, we strictly adhered to the site's robots.txt policy and privacy policy while scraping post titles. This ensured that no sensitive or private user information was accessed or collected during the process. Only public-facing content, such as post titles, was gathered, and no personal or identifiable information from users was included in the dataset.

7. ACKNOWLEDGEMENT

We acknowledge the indispensable contributions of the online community, DC Inside, for their assistance in collecting safe post titles data, and the editors of Namuwiki for their efforts in gathering threat posts, ultimately benefiting our study and the wider community.

Furthermore, special thanks to Dr. Hojin Moon, a luminary in the field of Statistics and a revered professor at California State University Long Beach for his invaluable support and meticulous review of our statistical analysis method.

References

[1] https://www.koreaherald.com/article/3316266

[2] https://www.koreatimes.co.kr/www/nation/2023/08/251_356392.html

[3] https://www.unitary.ai/articles/is-ai-content-moderation-better-than-human-content-moderation

[4] https://medium.com/@heatherrasley/can-ai-replace-one-of-the-most-traumatic-jobs-on-the-internet-adc41ef03d95

[5] https://www.magellan-solutions.com/blog/ai-vs-human-content-moderation/

[6] https://openai.com/api/pricing/

[7] https://github.blog/news-insights/product-news/github-copilot-x-the-ai-powered-developer-experience/

[8] https://openai.com/index/using-gpt-4-for-content-moderation/

[9] Törnberg P. Chatgpt-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages With Zero-Shot Learning. 2023. ArXiv preprint: https://arxiv.org/pdf/2304.06588

[10] Kumar D, AbuHashem Y, Durumeric Z. Watch Your Language: Investigating Content Moderation With Large Language Models. 2024. ArXiv preprint: https://arxiv.org/pdf/2309.14517v2

[11] https://livechatai.com/palm-2-pricing-calculator

[12] DCINSIDE. "실시간 베스트 갤러리 - 커뮤니티 포털 디시인사이드. 2021. Avaialable at: https://www.dcinside.com/

[13] Yang S. "Loser" Aesthetics: Korean Internet Freaks and Gender Politics. Fem Media Stud. 2018;19:858-872.

[14] Rust P, Pfeiffer J, Vulić I, Ruder S, Gurevych, I. How Good Is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics. 2021;1:3118-3135.

[15] Kleinig O, Gao C, Kovoor JG, Gupta AK, Bacchi S, et. al. How to Use Large Language Models in Ophthalmology: From Prompt Engineering to Protecting Confidentiality. Eye. 2024 Mar;38:649-653.

[16] https:// www.prompthub.us/ blog/ comparing-latencies-get-faster-responses-from-openai-azure-and-anthropic