YYama @ Multimodal Hate Speech Event Detection 2024: Simpler Prompts, Better Results - Enhancing Zero-shot Detection with a Large Multimodal Model

Yosuke Yamagishi
Graduate School of Medicine, The University of Tokyo, Japan
yamagishi-yosuke0115@g.ecc.u-tokyo.ac.jp

Abstract

This paper introduces a zero-shot hate detection experiment using a multimodal large model. Although the implemented model comprises an unsupervised method, results demonstrate that its performance is comparable to previous supervised methods. Furthemore, this study proposed experiments with various prompts and demonstrated that simpler prompts, as opposed to the commonly used detailed prompts in large language models, led to better performance for multimodal hate speech event detection tasks. While supervised methods offer high performance, they require significant computational resources for training, and the approach proposed here can mitigate this issue.

The code is publicly available at https://github.com/yamagishi0824/zeroshot-hate-detect.

1 Introduction

In the contemporary era marked by extensive use of social media, the forms of hate speech have diversified significantly. Hate speech embedded in images on social media, in particular, has become prevalent, rendering its detection crucial (Thapa et al., 2022; Bhandari et al., 2023). The Multimodal Hate Speech Event Detection 2024 shared task at The 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE @ EACL 2024) was a unique task focusing on detecting hateful content in text-embedded images posted on social media concerning the Russia-Ukraine conflict (Thapa et al., 2024). This task is an expanded version of the one conducted in the previous year (Thapa et al., 2023).

Prompt engineering is a method to improve the inference accuracy of a pre-trained model by adding task-specific information to the prompts that serve as inputs to the model. This approach has been extensively researched, particularly with large language models. Various studies have also been conducted on multimodal large models (Gu et al., 2023), proposing different techniques such as task instruction prompting (Efrat and Levy, 2020) and in-context learning (Brown et al., 2020).

In this multimodal hate speech event detection task, it was particularly important to acknowledge that the image was uploaded against the backdrop of the Russia-Ukraine conflict, and that the definition of hate speech was crucial for labeling. Therefore, this study examined the change in performance by using prompts that, in addition to being simple, also included contextual information explaining the task.

The main contributions of this research are as follows:

- The proposed method employs a widely accessible large multimodal model, enhancing its accessibility.
- The method operates under zero-shot conditions, eliminating the need for further model training and facilitating execution in computationally constrained environments, as long as inference is possible.
- This paper have engaged in prompt engineering to achieve improved performance under zero-shot conditions. While prompt engineering is extensively practiced for large language models, it remains limited for multimodal large models. By employing effective prompts, the performance will be improved.

2 Related Works

2.1 Multimodal Large Model

Using multimodal models enables the combination of multiple data types, including images, text, and audio, for input (Wu et al., 2023). While large language models were limited to only text data input, the ability to handle data from multiple modalities

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 60–66

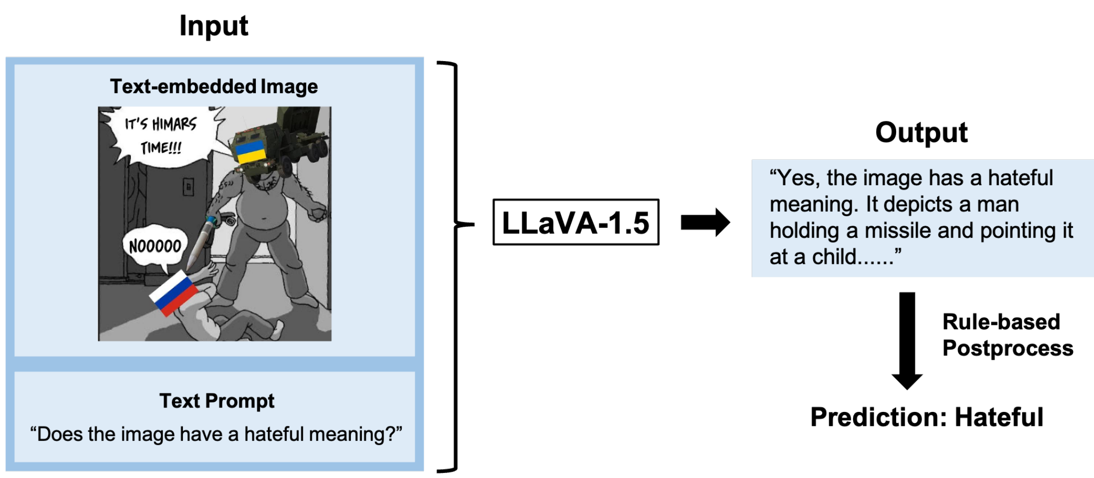

Figure 1: Flowchart of zero-shot hate detection.

expands the potential applications, making it a technology of growing interest. GPT-4 (Achiam et al., 2023) is a notable example of a large multimodal model, but its architecture details are confidential and not freely accessible. In contrast, LLaVA (Liu et al., 2023b) is an openly available model, and its updated version, LLaVA-1.5 (Liu et al., 2023a), has achieved top performance in various benchmarks and is also being used for zero-shot image classification (Islam et al., 2023).

2.2 Hate Speech Detection using Multimodal Large Model

The detection of hate speech in text-embedded images using multimodal models has been implemented for the dataset utilized in this study (Bhandari et al., 2023). In this method, multimodal models such as CLIP (Radford et al., 2021) and GroupViT (Xu et al., 2022) have been employed and fine-tuned, demonstrating superior results compared to unimodal models that use either text or images alone. Furthermore, as a method for detecting hate speech from internet memes, approaches using multimodal models with zero-shot prompting have also been experimented with (Van and Wu, 2023). In this study, by employing the LLaVA, there are cases where it surpasses the performance of past fine-tuned multimodal models. We aim to further leverage the potential of LLaVA by conducting a more detailed comparison of prompt performance.

3 Dataset & Task

3.1 Dataset

This study was conducted in line with the Multimodal Hate Speech Event Detection 2024 shared task at CASE @ EACL 2024. The dataset used was CrisisHateMM, consisting of 4,723 images collected from social media platforms such as Twitter, Facebook, and Reddit (Bhandari et al., 2023). These images are embedded with text and labeled to indicate whether they contain hateful content or not. Additionally, labels are provided to denote whether the subject is an individual, community, or organization.

3.2 Task

The shared task comprises two sub-tasks (Sub-task A & B), of which we participated solely in sub-task A.

Sub-task A is focused on hate speech detection where the objective is to examine images containing text to detect any instances of hate speech (Bhandari et al., 2023; Thapa et al., 2024). This process will utilize a dataset which has already been annotated in advance to assess the frequency of such content. For the sub-task, the dataset comprises 4,723 text-embedded images categorized into two classes: 'Hate' and 'No Hate'. Of these, 2,665 images (56.43%) are labeled as 'Hate', and 2,058 (43.57%) are labeled as 'No Hate'. Additionally, 443 of these images are designated as test data, but the breakdown of labels within the test subset is not provided.

4 Methodology

Using a pre-trained multimodal large model, we input images and text prompts, prompting the model to generate text. Then, by applying rule-based postprocessing to the generated texts, we obtained predicted values of whether the content is hateful or non-hateful. The example of flowchart is Figure 1.

4.1 Multimodal Large Model

We adopted the state-of-the-art Large Language-and-Vision Assistant model, LLaVA-1.5, as a multimodal large model (Liu et al., 2023a). LLaVA-1.5 is available to anyone under the LLAMA 2 Community License (Touvron et al., 2023). Out of the available 7 billion and 13 billion parameter versions, we chose the 13 billion parameter model for its robust capabilities. It outputs generated text when given image and text prompts as inputs. We devised multiple prompts and performed inference using images from the dataset as inputs.

4.2 Prompts

We devised various prompts and compared their performance by combining them. We prepared multiple simple prompts. One of these was selected as the base prompt, to which we added contextual information, thereby creating several variations of the prompt. The contextual information included the fact that 'the images were uploaded to social media in the context of the Russia-Ukraine conflict,' definitions of 'hateful,' and 'non-hateful.' The descriptions of these informations were cited from the expressions in the dataset paper (Bhandari et al., 2023).

Specifically, the following prompts were used as input for the model (All detailed prompts are listed in the appendixA). The abbreviations within the parentheses are used in Table 1:

- Simple Base Prompt (**Simple Base**):
  - *"Does this image have a hateful meaning?"*
    This prompt was repeatedly used as a baseline prompt, along with contextual information.
- Other Simple Prompts:
  These prompts were used to evaluate the performance of simple prompts with slight variations, such as replacing words. The detailed prompts in Tabel 2.
- Prompt with Background Information on the Russia-Ukraine Conflict (**Conflict**):
  - *"Note: the image is about the war between Russia and Ukraine that occurred in 2022, posted on social media."*
    This incorporated background information about the image being uploaded on social media related to the conflict between Russia and Ukraine.
- Prompt with Definition of 'hateful' (**Hate**):
  - *"Note: A text-embedded image is classified as hateful if it contains visual or textual hateful content such as threats, personal attacks, slander, or abuse."*
    The definition of 'hateful' was cited from the dataset paper.
- Prompt with Definition of 'non-hateful' (**Non-hate**):
  - *"Note: A text-embedded image is considered non-hateful if it reports events or objectively reports others' opinions in a non-hateful manner."*
    The definition of 'non-hateful' was cited from the dataset paper.
- Prompt with Both 'hate' and 'non-hate' Definitions (**Hate + Non-hate**):
  This prompt incorporated both the definitions of 'hateful' and 'non-hateful'.
- Prompt with All Contextual Information (**Conflict + Hate + Non-hate**):
  This prompt included all background information about the image, as well as both 'hateful' and 'non-hateful' definitions.

4.3 Postprocess

The outputs from LLaVA-1.5 were received as text, which necessitated their conversion into a binary categorization of either 'hate' or 'non-hate.' We confirmed that by inputting closed questions as prompts, without the need for detailed adjustments, the output would begin with 'Yes' if it contained hate speech. Consequently, a simple post-processing step was implemented: outputs containing "Yes" were classified as 'hate,' and those without it were classified as 'non-hate'.

5 Results

5.1 Performance of Simple Prompts

The comparison of performance using simple prompts is shown in Table1. Although there are some variations, the main evaluation metric used in this shared task, the F1 score, only varied from

Prompt	Recall	Precision	F1	Accuracy
Does this image have a hateful meaning? (**Simple Base**)	0.761	0.759	0.758	0.758
Does this picture have a hateful meaning?	0.748	0.747	0.743	0.743
Does this image have hateful content?	0.757	0.804	0.759	0.774
Does this image convey any hateful messages?	0.755	0.756	0.747	0.747
Is this image hateful?	0.757	0.755	0.755	0.756

Table 1: Comparison of different simple prompts for hate detection.

Prompt	Recall	Precision	F1 Score	Accuracy
Simple Base	**0.761**	**0.759**	**0.758**	**0.758**
Conflict	0.735	0.742	0.737	0.743
Hate	0.715	0.731	0.697	0.700
Non-hate	0.641	0.694	0.595	0.614
Hate + Non-hate	0.716	0.749	0.689	0.695
Conflict + Hate + Non-hate	0.720	0.736	0.702	0.704

Table 2: Comparison of prompts with contextual information.

0.743 to 0.759, a maximum difference of 0.016. It is reasonable to conclude that minor variations in the wording of prompts do not result in significant performance differences.

5.2 Performance of Prompts with Contextual Information

Table2 shows the results. When the contextual information was added to the simple prompts, performance decreased in all cases, with the simple baseline prompt performing the best.

The decrease in performance was particularly notable when the definition of 'Non-hate' was added, with a reduction of 0.120 in the F1 score. In the case of the simple prompts, the number predicted as 'hateful' was 220 (49.7%), whereas with the 'Non-hate' prompt, it dropped to 105 (23.7%), less than half.

In prompts with added contextual information, the 'Conflict' prompt performed the best. However, even then, there was a decrease in performance in terms of precision, recall, and F1 score compared to any of the other simple baselines. The performance was also the lowest in terms of accuracy, matching the lowest score among them.

5.3 Comparison with Previous Baselines

Compare with the baseline performance shown in the dataset paper (Bhandari et al., 2023). In the baselines, fine-tuning and prediction were performed for models with only text, only image, and multimodal of text and image. Table3 displays the performance of each along with the F1 score and accuracy by our simple base prompt. Our proposed method demonstrated superior performance compared to the image model, yet it showed inferior results when compared to the text and multimodal models. The difference in the F1 score relative to the text model was 0.011.

Method	F1	Accuracy
Textual	0.769	0.779
Visual	0.739	0.741
Multimodal	0.786	0.798
Ours	0.758	0.758

Table 3: Comparison with previous baselines.

5.4 Output Characteristics for Development Data

The labels for the test data have not been published, therefore, we conducted error analysis using the development data using the simple base prompt.

The performance on the development data was a recall of 0.794, a precision of 0.794, an F1 score of 0.772, and an accuracy of 0.774.

Of the outputs generated by LLaVA-1.5, the initial sentences included phrases like *"Yes, the image has a hateful meaning"* or *"Yes, the image contains a hateful meaning,"* comprising 243 instances (54.9%). There were 186 instances (42.0%) that clearly predicted no-hate, containing either *"No, the image does not have a hateful meaning"* or *"The image does not have a hateful meaning."* The remaining 14 instances (3.2%) were either merely descriptions of the image content or avoided explicitly stating whether the content was hate or no-hate.

5.5 Qualitative Error Analysis

LLaVA-1.5 not only predicts but also outputs the reasoning behind its predictions. This was utilized for a qualitative error analysis.

The figure 2 represents an example where the label is 'no hate', but it was predicted as 'hate'. This image depicts the Lithuanian independence revolution, during which Ukraine supported Lithuania, and now Lithuania is supporting Ukraine, making it a 'no hate' content.

The model interpreted it completely oppositely as *"It shows a protest sign with a message that is anti-Ukrainian, which is offensive and promotes discrimination"*, although no OCR results of the sign or text were provided (the full output is in the appendixA).

It is presumed that an understanding of historical context and accurate OCR are necessary for prediction, but these seem to have failed in this case.

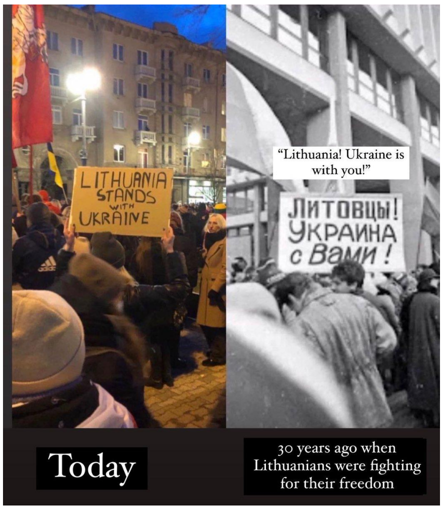

Figure 2: An example where it was predicted as hate despite being labeled as no hate.

6 Discussion

In this study, we found that prompts containing background information performed worse than the base simple prompts. While it is generally expected that performance improves with the use of instruction prompting, it is intriguing that performance declined when task-specific information, such as the definition of hate speech, was provided. Particularly, adding the definition of no-hate to the prompt seemed to decrease performance. This can be attributed to the bias introduced in the inference due to the information included in the prompt, resulting in an increased prediction of no-hate.

On the other hand, simply providing simple prompts surpassed the performance of past fine-tuned image models and closely matched text models. This result demonstrates the potential of pre-trained multimodal large models to be utilized for hate speech detection even under zero-shot conditions.

This study was exclusively focused on using LLaVA-1.5, and exploring other large multimodal models might produce different results. Given that LLaVA-1.5 is a top-performing, freely available model, the emergence of new models may necessitate additional validation. The research was specifically aimed at detecting hate speech in images containing text on social media, a critical but narrowly focused task. Applying more complex prompts in varied tasks could enhance performance. The significance of identifying hate speech in such images is heightened by the extensive use of social media today. As datasets grow, continued research in this field will be increasingly valuable.

Due to the emergence of large language and multimodal models, zero-shot detection is expected to be increasingly used for sensitive tasks. It's essential to balance the freedom of social media posting with avoiding excessive censorship. Hence, enhanced performance and proper management in zero-shot hate detection are imperative as future tasks.

7 Acknowledgements

We would like to express my sincere gratitude to the hosts of the shared task for providing us with the opportunity to conduct our research. We state that this research was carried out independently without any financial support.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aashish Bhandari, Siddhant Bikram Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.

Ashhadul Islam, Md Rafiul Biswas, Wajdi Zaghouani, Samir Brahim Belhaouari, and Zubair Shah. 2023. Pushing boundaries: Exploring zero shot object classification with large multimodal models. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–5. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Surendrabikram Thapa, Farhan Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka-Wei Lee, and Usman Naseem. 2023. Multimodal hate speech event detection - shared task 4, CASE 2023. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 151–159, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Hari Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during russia-ukraine crisis - shared task at case 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*.

Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. 2022. A multimodal dataset for hate speech detection on social media: Case-study of russia-Ukraine conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 1–6, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Minh-Hao Van and Xintao Wu. 2023. Detecting and correcting hate speech in multimodal memes with large visual language model. *arXiv preprint arXiv:2311.06737*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144.

A Appendix

A.1 Prompts

These are the entire texts of the input prompts:

- Simple Base Prompt
  - *"Does this image have a hateful meaning?"*
- Other Simple Prompts:
  - *"Does this picture have a hateful meaning?"*
  - *"Does this image have hateful content?"*
  - *"Does this image convey any hateful messages?"*
  - *"Is this image hateful?"*
- Prompt with Background Information on the Russia-Ukraine Conflict:
  - *"Does this image have a hateful meaning?\nNote: the image is about the war between Russia and Ukraine that occurred in 2022, posted on social media."*

- Prompt with Definition of 'hateful':
  - *"Does this image have a hateful meaning?\nNote: A text-embedded image is classified as hateful if it contains visual or textual hateful content such as threats, personal attacks, slander, or abuse."*
- Prompt with Definition of 'non-hateful':
  - *"Does this image have a hateful meaning?\nNote: A text-embedded image is considered non-hateful if it reports events or objectively reports others' opinions in a non-hateful manner."*
- Prompt with Both 'hate' and 'non-hate' Definitions:
  - *"Does this image have a hateful meaning?\nNote: A text-embedded image is classified as hateful if it contains visual or textual hateful content such as threats, personal attacks, slander, or abuse. A text-embedded image is considered non-hateful if it reports events or objectively reports others' opinions in a non-hateful manner."*
- Prompt with All Background Information:
  - *"Does this image have a hateful meaning?\nNote: the image is about the war between Russia and Ukraine that occurred in 2022, posted on social media. A text-embedded image is classified as hateful if it contains visual or textual hateful content such as threats, personal attacks, slander, or abuse. A text-embedded image is considered non-hateful if it reports events or objectively reports others' opinions in a non-hateful manner."*

A.2 Incorrect Output

This is the entire output that contains incorrect predictions for Figure 2.

"Yes, the image has a hateful meaning. It shows a protest sign with a message that is anti-Ukrainian, which is offensive and promotes discrimination. The sign is being held by a crowd of people, indicating that the message is being supported by others. Such messages should not be tolerated, as they can lead to further division and harm among people. It is essential to promote respect, understanding, and inclusivity in society."