Prompting for Multimodal Hateful Meme Classification

Rui Cao¹, Roy Ka-Wei Lee², Wen-Haw Chong¹, Jing Jiang¹

¹Singapore Management University
²Singapore University of Design and Technology
¹{ruicao.2020@phdcs.,whchong.2013@phdis.,jingjiang@}smu.edu.sg
²roy_lee@sutd.edu.sg

Abstract

Hateful meme classification is a challenging multimodal task that requires complex reasoning and contextual background knowledge. Ideally, we could leverage an explicit external knowledge base to supplement contextual and cultural information in hateful memes. However, there is no known explicit external knowledge base that could provide such hate speech contextual information. To address this gap, we propose PromptHate, a simple yet effective prompt-based model that prompts pre-trained language models (PLMs) for hateful meme classification. Specifically, we construct simple prompts and provide a few in-context examples to exploit the implicit knowledge in the pre-trained RoBERTa language model for hateful meme classification. We conduct extensive experiments on two publicly available hateful and offensive meme datasets. Our experimental results show that PromptHate is able to achieve a high AUC of 90.96, outperforming state-of-the-art baselines on the hateful meme classification task. We also perform fine-grained analyses and case studies on various prompt settings and demonstrate the effectiveness of the prompts on hateful meme classification.

Disclaimer: This paper contains discriminatory content that may be disturbing to some readers.

1 Introduction

Internet memes have evolved into one of social media’s most popular forms of communication. Memes are presented as images with accompanying text, which are usually intended to be funny or satirical in nature. However, malicious online users generate and share hateful memes under the guise of humor (Kiela et al., 2020; Pramanick et al., 2021a; Gomez et al., 2020; Suryawanshi et al., 2020). Many hateful memes had been created to attack and ridicule people based on specific characteristics such as race, ethnicity, and religion. The hateful memes could also threaten society more than text-based online hate speech due to their viral nature; users could re-post or share these hateful memes in multiple conversations and contexts.

To combat the spread of hateful memes, social media platforms such as Facebook have recently released a large hateful meme dataset as part of a challenge to encourage researchers to develop automated solutions to perform hateful memes classification (Kiela et al., 2020). However, classifying hateful memes turns out to be a very challenging task as the solution would need to comprehend and reason across both the visual and textual modalities. The reasoning of the modalities will also require contextual background knowledge.

Recent research works have proposed multimodal hateful meme classification approaches. For instance, some studies have adopted pre-trained visual language models such as ViLBERT (Lu et al., 2019) and VisualBERT (Li et al., 2019) and fine-tune these models with the hateful meme classification task (Lippe et al., 2020; Zhu, 2020; Zhou and Chen, 2020; Muennighoff, 2020; Ve-
lioglu and Rose, 2020; Zhu et al., 2022) (as illustrated in Figure 1(a)). Nevertheless, existing approaches still have limitations as understanding hateful memes may require additional contextual background knowledge. Consider the hateful meme example in Figure 1. The background knowledge that the pig is considered unclean by Muslims and is a sin to consume, is required to infer that the meme is hateful.

Recent studies have attempted to prompt Pre-trained Language Models (PLM) and yield good performance for uni-modal NLP (Schick and Schütze, 2021a; Brown et al., 2020; Schick and Schütze, 2021b; Gao et al., 2021). Nevertheless, few works have attempted to prompt PLMs for multimodal tasks (Yao et al., 2021; Zeng et al., 2022; Gui et al., 2021). Yang et al. (2021) has explored prompting GPT-3 model (Brown et al., 2020) for the visual question & answering task. However, the approach has limitations as large models such as GPT-3 are expensive to tune. This study addresses the research gaps and proposes a novel framework to leverage the implicit and unstructured knowledge in PLMs (Trinh and Le, 2018; Petroni et al., 2019) to improve hateful meme classification. Figure 1 illustrates the comparison between the existing multimodal hateful meme classification approach and our proposed prompt-based approach. Specifically, in our prompt-based approach, we first convert images into textual descriptions that a PLM can understand and design specific prompts to adapt and leverage the implicit knowledge in the PLM. Subsequently, given a meme and a prompt, the prompt-tuned PLM generates a textual output, indicating the predicted label of the input meme. The underlying intuition for the prompt-based approach is that PLMs will tap into the implicit and unstructured knowledge in their large-scale pretraining data to generate the continuation of the prompt, i.e., from “this is _” to “this is hateful.”.

We summarize this paper’s contribution as follows: (i) We propose a multimodal prompt-based framework called PromptHate, which prompts and leverages the implicit knowledge in the PLM to perform hateful meme classification. (ii) We conduct extensive experiments on two publicly available datasets. Our experiment results show that PromptHate outperforms state-of-the-art methods for the hateful meme classification task. (iii) We perform fine-grained analyses and case studies on various settings to examine the prompts’ effectiveness in classifying hateful meme. To the best of our knowledge, this is the first paper that explore prompting PLM for hateful meme classification. 1

2 Related Work

2.1 Hateful Meme Detection

Hateful meme classification is an emerging multimodal task made popular by the availability of several recent hateful memes datasets (Kiela et al., 2020; Suryawanshi et al., 2020; Gomez et al., 2020). For instance, Facebook had organized the Hateful Memes Challenge, which encouraged researchers to submit solutions to perform hateful memes classification (Kiela et al., 2020). The memes are specially constructed such that unimodal methods cannot yield good performance in this classification task. Therefore, existing studies have adopted multimodal approaches to perform hateful memes classification.

Existing studies have explored classic two-stream models that combine the text and visual features learned from text and image encoders using attention-based mechanisms and other fusion methods to perform hateful meme classification (Zhang et al., 2020; Kiela et al., 2020; Suryawanshi et al., 2020). Another popular line of approach is fine-tuning large scale pre-trained multimodal models for the task (Lippe et al., 2020; Zhu, 2020; Zhou and Chen, 2020; Mueenighoff, 2020; Velioglu and Rose, 2020; Pramanick et al., 2021b; Lee et al., 2022). Recent studies have also attempted to use data augmentation (Zhu, 2020; Zhou and Chen, 2020; Zhu et al., 2022) and ensemble methods (Zhu, 2020; Velioglu and Rose, 2020; Sandulescu, 2020) to enhance the hateful memes classification performance. In a recent work, Lee et al. (2021) proposed the DisMultiHate mode in attempted to disentangle hateful targets in memes. Nevertheless, existing studies may be inadequate in modeling the contextual background knowledge encoded in the hateful memes. This paper aims to fill this research gap by prompting the PLM to leverage its unstructured implicit knowledge for hateful meme classification.

2.2 Language Model Prompting

The increasing popularity of large-scale PLMs such as GPT (Radford et al., 2019; Brown et al., 2020), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) has also popularized prompt-based learning.

1 Code: https://gitlab.com/bottle_shop/safe/prompthate
Existing prompt-based learning studies have explored using PLMs as implicit and unstructured knowledge bases (Talmor et al., 2020; Davison et al., 2019; Schwartz et al., 2017). Recent studies have also prompted PLMs for various NLP tasks such as natural language inference and sentiment classification, and yield good performance in few-shot settings (Gao et al., 2021; Schick and Schütze, 2021a,b). There are also recent works that prompt visual-language models for computer vision tasks (Zhou et al., 2022, 2021; Radford et al., 2021).

Nevertheless, most of the existing prompt-based learning studies are limited to unimodal tasks, and there are fewer works on prompting PLM for multimodal tasks (Yao et al., 2021; Zeng et al., 2022; Gui et al., 2021). Yang et al. (2021) has explored prompting GPT-3 model (Brown et al., 2020) for the visual question & answering task. However, there are limitations: large models such as GPT-3 are expensive to tune. Furthermore, the constraint on the input length limits the number of training instances. In this paper, we adopt a different approach and propose a novel framework, PromptHate, which prompts RoBERTa (Liu et al., 2021) for the multimodal hateful meme classification. PromptHate is a much smaller model compared to GPT-3, and it can be fine-tuned with training instances.

3 Preliminaries

3.1 Problem Definition

We define the problem of multimodal hateful memes classification as follows: Given a meme (with image $I$ and text $T$), a classification model will predict the label of the multimodal meme (hateful or non-hateful). Traditionally, this binary classification task requires models to predict a probability vector $y \in \mathbb{R}^2$ over the two classes. Specifically, $y_0$ denotes the predicted probability that the meme is non-hateful while $y_1$ is for the probability that the meme is hateful. If $y_1 > y_0$, the meme is predicted as hateful, otherwise, non-hateful. In framework, we transform the hateful meme classification task into a Masked Language Modelling (MLM) problem. Specifically, a PLM is prompted to replace the [MASK] token that represents the label of the meme (e.g., hateful or non-hateful). We discuss the prompting details in Section 4.1.

3.2 Image Captioning

To prompt PLMs for multimodal hateful meme classification, we first need to covert the meme’s image into an acceptable textual input for PLMs. A common approach to extract the image’s semantics and represent it with textual description is via image captioning (Yang et al., 2021; Gui et al., 2021). We first extract the text in the memes using open-source Python packages EasyOCR\(^2\), followed by in-painting with MMEditing\(^3\) to remove the text. We then apply a pre-trained image captioning model, ClipCap (Mokady et al., 2021). ClipCap is able to generate good quality captions for low-resolution web images. The generated captions tend to describe the dominant objects or events in the meme’s image and we use these captions as inputs into the PromptHate model.

Besides captioning the image, we also leveraged Google Vision Web Entity Detection API\(^4\) and pre-trained FairFace classifier (Kärkkäinen and Joo, 2019) to extract the entities in the memes and the demographic information if the meme contains a person. The extracted entities and demographic information are used as supplementary information that will be combined with the image captions as input to the PLMs. Note that although the extracted supplementary information may capture key information about the meme, the contextual background knowledge is still absent in the image caption and supplementary information. For instance, with the utilization of entity information, we may identify a pig in the meme and extract the term “Muslim” from the meme text. However, the contextual knowledge that Muslims do not eat pork is absent in the supplementary information.

4 Methodology

Figure 2 illustrates the architectural framework of our proposed PromptHate model. A key step in the PromptHate is the construction of a prompt, which consists of a positive demonstration (i.e., normal meme), a negative demonstration (i.e., hateful meme), and an inference instance (i.e., meme to be predicted). We first convert the three memes into meme texts and image descriptions using the data pre-processing steps described in Section 3.2. Subsequently, we construct templates, which are natural sentences that include label words for the

---

\(1\)https://github.com/JaidedAI/EasyOCR

\(2\)https://github.com/open-mmlab/mmediting

\(3\)https://cloud.google.com/vision/docs/detecting-web

---

323
individual memes. For instance, a normal demonstration meme will have a template “the meme is **good**”, while the hateful demonstration meme uses the template “the meme is **bad**”. The label word in template for the inference instance is replaced with a **[MASK]**, which the PLM (i.e., RoBERTa) is tasked to complete the sentence with “**good**” or “**bad**”. Subsequent sections provide the technical details on prompting for the multimodal hateful meme classification task.

### 4.1 Prompting Hateful Meme

To guide the PLM in inferring the label word, we also provide positive and negative demonstrations to the PLM. The positive demonstration $\delta^\text{pos}$ is generated as: $\delta^\text{pos}_1[\text{SEP}] \delta^\text{pos}_2[\text{SEP}] \mathcal{T}(W^\text{pos})$, where $\delta^\text{pos}_1$ and $\delta^\text{pos}_2$ are meme texts and image descriptions respectively, [SEP] is the separation token in the language model $\mathcal{L}$, and $\mathcal{T}(W^\text{pos})$ generates the positive label word $W^\text{pos}$ into a sentence (e.g., “**this is good**”). Similar approach is used for the generation of negative demonstration $\delta^\text{neg}$ and inference instance $\delta^\text{infer}$ by replacing $W^\text{pos}$ with $W^\text{neg}$ and **[MASK]**, respectively. Inspired by Gao et al. (2021), we concatenate the demonstrations with the inference instance:

$$
\delta = [\text{START}] \delta^\text{infer} [\text{SEP}] \delta^\text{pos} [\text{SEP}] \delta^\text{neg} [\text{END}]
$$

where, $\delta$ serves as the prompt fed into $\mathcal{L}$, and [START] and [END] are start and end tokens in $\mathcal{L}$.

### 4.2 Templates and Label Words

Recent studies have explored designing better prompts by developing automatic template generation and label word selection methods (Gao et al., 2021). As PromptHate is the first study that adopted prompting for hateful meme classification, we adopt a simpler approach of prompting with manually defined label words and templates. Labels are required to be mapped into individual words for prompt-based models. As shown in Figure 2, **good** is used as the label word for the positive class (non-hateful), while **bad** for the negative class (hateful). We have also analysed sets of other label words. The comparison of using different label words is discussed in Section 5.2.

The template in prompts can be viewed as the function $\mathcal{T}$, which maps the label word into a sentence. In PromptHate, we manually define the function $\mathcal{T}([\text{WORD}]) \rightarrow \text{It was [WORD]}$. Specifically, if $\mathcal{T}$ receives $W^\text{pos}$ as input, the output sentence should be “It was $W^\text{pos}$.”. Conversely, if $\mathcal{T}$ receives $W^\text{neg}$ as input, the output sentence should be “It was $W^\text{neg}$.”.

### 4.3 Model Training and Prediction

For training, we feed the prompt $\delta$ into $\mathcal{L}$ and obtain the probability of the masked word, $y \in \mathbb{R}^2$ over label words:

$$
y_0 = P([\text{MASK}] = W^\text{pos}|\delta),$$

$$y_1 = P([\text{MASK}] = W^\text{neg}|\delta).
$$

The training loss is based on cross-entropy loss with the ground-truth label $\hat{y}$:

$$
\text{Loss} = y_0 \log(\hat{y}_0) + y_1 \log(\hat{y}_1),
$$

and the loss will be used for updating parameters $\theta$ in $\mathcal{L}$. Differing from standard fine-tuning PLMs by adding a task-specific classification head, prompt-based tuning does not have additional parameters beyond those in the PLMs, and the MLM task does not deviate from PLM’s pre-training objectives.
For model prediction, we obtain the probability of the masked word over label words in the same manner. If \( y_1 > y_0 \), the meme will be predicted as hateful, otherwise, non-hateful.

4.4 Multi-Query Ensemble

Demonstrations in the prompt provide additional cues for the inference instance. Existing works carefully select demonstrations which are similar to the inference instance (Yang et al., 2021; Gao et al., 2021). Nevertheless, memes that are similar in visual or textual modality may be targeting different protected characteristics (e.g., race, religion, gender, etc.), and understanding the target in the hateful meme is critical to the classification task (Lee et al., 2021). To address this concern, we adopt a multi-query ensemble strategy to predict the inference instance using multiple pairs of demonstrations. Specifically, when we adopt a \( M \)-query ensemble, an inference instance will be predicted using \( M \) pairs of demonstrations.

The multi-query ensemble will result in a set of prediction scores for the inference instance: \( \{y_m\}_{m=1}^M \), where \( y_m \in \mathbb{R}^2 \) is the predicted scores with the \( m \)-th pair of demonstration. The final prediction will be the average over all predicted scores:

\[
y_{\text{final}} = \frac{1}{M} \sum_{m=1}^M y_m.
\]

5 Experiments

In this section, we first provide a brief introduction to the datasets and evaluation setting. Next, we present a set of experiments conducted to evaluate PromptHate’s hateful meme classification performance. We also conduct studies to understand the effects of various prompt settings, and discuss the limitations of our model via error case studies.

5.1 Evaluation Settings

Datasets. We used two publicly available datasets in our experiments: the Facebook Hateful Meme dataset (FHM) (Kiela et al., 2020) and the Harmful Meme dataset (HarM) (Pramanick et al., 2021a). Table 1 outlines the statistical distributions of the two datasets. The FHM dataset was constructed and released by Facebook as part of a challenge to crowdsource multimodal hateful meme classification solutions. We do not have labels of the memes in the test split. Therefore, we utilize the dev-seen split as the test. Due to the limited availability of public multi-modal hateful meme datasets, we choose HarM dataset containing misinformation memes as the other evaluation dataset. The HarM dataset was constructed with real COVID-19-related memes collected from Twitter. The memes are labeled with three classes: very harmful, partially harmful, and harmless. We combine the very harmful and partially harmful memes into hateful memes and regard harmless memes as non-hateful memes. The good performance on the HarM dataset also implies the generalization of the PromptHate to other anti-social memes besides hateful ones.

Evaluation Metrics. We adopt the evaluation metrics commonly used in existing hateful meme classification studies (Kiela et al., 2020; Zhu, 2020; Zhou and Chen, 2020; Muenninghoff, 2020; Velioglu and Rose, 2020): Area Under the Receiver Operating Characteristic curve (AUROC) and Accuracy (Acc). In order to report more reliable results, we measure the average performance of models under ten random seeds. All models use the same set of random seeds.

Baselines. We benchmark PromptHate against the state-of-the-art hateful meme classification models. Specifically, we compare with two types of baselines models: (a) uni-modal models that only use information from one modality (i.e., the meme text or the meme image); (b) multimodal models. For uni-modal baselines, we consider a text-only model by fine-tuning pre-trained BERT on the meme text for classification (Text BERT). We also apply an image-only model, which processes the meme image using Faster R-CNN (Ren et al., 2016) with ResNet-152 (He et al., 2016) before feeding the image representation into a classifier for hateful meme classification (Image-Region).

For multimodal baselines, we compare with the multimodal methods benchmarked in the original FHM dataset paper (Kiela et al., 2020), namely: Late Fusion, Concat BERT, MMBT-Region (Kiela et al., 2019), ViLBERT CC (Lu et al., 2019), Visual BERT COCO (Li et al., 2019). We also compare to the state-of-the-art hateful meme classification methods: 5: CLIP BERT,

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Train #Hate.</th>
<th>#Non-hate.</th>
<th>Test #Hate.</th>
<th>#Non-hate.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FHM</td>
<td>3,050</td>
<td>5,450</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>HarM</td>
<td>1,064</td>
<td>1,949</td>
<td>124</td>
<td>230</td>
</tr>
</tbody>
</table>

Table 1: Statistical summary of FHM and HarM.

5 Note that we use the code published by the author and re-run the model for ten rounds with different random seeds.
Models with either RoBERTa-large (Pramanick et al., 2021b) and DisMultiHate (Lee et al., 2021). CLIP BERT and MOMENTA are models leveraging image features generated by the CLIP model (Radford et al., 2021). CLIP BERT is pre-trained with web data, thus it is able to generalize well to hateful meme detection where images and texts are noisy. CLIP BERT uses CLIP as the visual encoder and BERT as the text encoder and feed the concatenation of features to a classifier for prediction. MOMENTA considers the global and local information in two modalities by modeling the deep multi-modal interactions. DisMultiHate disentangles target information from the meme to improve the hateful content classification.

As PromptHate prompts RoBERTa (Liu et al., 2019) for hateful meme classification, we also benchmark PromptHate against fine-tuning RoBERTa (FT-RoBERTa). Specifically, we concatenate the meme text and image descriptions as input to fine-tune RoBERTa, and the output representation is fed into a MLP layer for classification.

### 5.2 Experiment Results

Table 2 and 3 show the experimental results on FHM and HarM datasets, respectively. The standard deviations (±) of the ten runs are also reported, and the best results are bold. PromptHate outperforms the state-of-the-art baselines in both datasets. We have also computed the statistical differences between PromptHate and the best-performing baseline (i.e., DisMultiHate on FHM and FT-RoBERTa on HarM), and PromptHate’s improvement over the baseline is found to be statistically significant (p-value < 0.05). Consistent with the existing studies, the multimodal approaches outperformed the unimodal baselines. More interestingly, we noted PromptHate’s improvements over the multimodal baselines that fine-tuned PLMs and FT-RoBERTa, demonstrating the strength of the prompting approach for the hateful meme classification task. Specifically, the performance comparison of FT-RoBERTa and PromptHate suggests that the prompting approach can better leverage the implicit knowledge embedded in the PLM by adopting a masked language modeling training objective for the hateful meme classification.

We also observe differences in PromptHate’s performance on the FHM and HarM datasets; the model yields better performance on HarM. Similar observations are made for the other models. We postulate that the performance differences are likely due to the difficulty of the dataset. FHM contains hateful memes on multiple topics, while HarM mainly contains COVID-19-related hateful memes. Therefore, the models would have to be able to generalize better to perform well on the FHM dataset. We also highlight the high standard deviation in FT-RoBERTa’s performance on FHM, suggesting FT-RoBERTa’s instability and difficulty in generalizing well on the dataset.

As RoBERTa-large is regarded as a general LM for prompting (Gao et al., 2021; Schick and Schütze, 2021b,a), PromptHate with RoBERTa-large is three times in the scale compared with BERT-base related baselines. To further validate the effectiveness of prompting approach in hateful meme detection, we conduct the following experiments: 1) we replace the RoBERTa-large with RoBERTa-base in PromptHate (PromptHate-RB); 2) we replace the BERT-base in the baseline models with either RoBERTa-large (-RL) or BERT-large (-BL). Specifically, we choose the most powerful baseline, DisMultiHate, for analysis. Experimental results on FHM and HarM are summarized in Table 4 and Table 5 respectively, where each block includes models of similar sizes.
Unsurprisingly, replacing the RoBERTa-large with RoBERTa-base worsens PromptHate performance. However, we do observe that PromptHate-RB still outperforms DisMultiHate on the HarM dataset. On the FHM dataset, PromptHate-RB has performed slightly worse than DisMultiHate but depicted higher stability regarding the standard deviation. Interestingly, replacing the text encoder of DisMultiHate with a larger pre-trained LM does not outperform PromptHate on both datasets. From the experimental results, we observe that model size plays a critical role in PromptHate performance. Nevertheless, the experimental results have also demonstrated the effectiveness of our proposed prompting approach over state-of-the-art baselines.

### 5.3 Ablation Study

Table 6 shows the ablation analysis of PromptHate. We notice removing the MLM training objective decreases PromptHate’s performance significantly. The MLM training objective is designed to align with the PLMs’ training objectives. This plays a significant role in enabling PromptHate to better utilize the embedding implicit knowledge in the PLMs for hateful meme classification. Interestingly, we observe that PromptHate can perform well even without the demonstrations. Nevertheless, the effects of demonstrations in prompt-based model remains an open research topic, which requires further studies (Min et al., 2022).

### 5.4 Prompt Engineering

Designing good prompts is essential to prompt-based models. In this section, we discuss how varying the prompts affect PromptHate’s performance in hateful meme classification.

#### 5.4.1 Engineering Label Words

Label words are individual words representing the labels used in prompt generation. We investigate the effects of replacing the prompts in PromptHate with different sets of label words. Specifically, we replace the label words in the prompt’s positive and negative demonstrations in our experiments. Table 7 presents the results. For example, in the first row in Table 7, we use “It was normal” for positive demonstrations (i.e., non-hateful memes), and “It was hate” for negative demonstration (i.e., hateful memes). Intuitively we aim to examine how the label word’s semantics affect hateful meme classification. For a more extensive investigation, we conduct the experiments on full training and few-shot setting, i.e., using only 10% of training instances.

Table 7 shows that different prompts can lead to substantial differences in performances. Specifically, label words aligned to the semantic classes are able to achieve better performance compared to the reverse mapping (i.e., the last row of each setting). Interestingly, the differences between the semantic class-aligned prompts and the reverse mapping are more significant in the few-shot setting. A possible reason could be in the few-shot setting, the PromptHate relies more on the label words’ semantics to extract implicit knowledge for hateful meme classification. Thus, the label words with the aligned semantic class will provide better context in the prompt to improve hateful meme classification when there are insufficient observations in training instances. Conversely, when PromptHate is trained with enough instances, the representations of the label words are updated to be closer to the hateful meme classification task.

#### 5.4.2 Prompt with Hateful Target Information

Existing studies have found that modeling target information (i.e., the victim of the hateful content) can help improve hateful meme classification (Lee et al., 2021). Therefore, we explore the effect of explicitly including the target information in prompts.

The FHM and HarM datasets are annotated with the target information. For our experimental design, we change the prompt template: from “It was [MASK].” to “It was [LABEL_MASK] targeting
Table 8: PromptHate without and with target information

<table>
<thead>
<tr>
<th>Model</th>
<th>FHM AUC.</th>
<th>Acc.</th>
<th>HarM AUC.</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o Target</td>
<td>81.45</td>
<td>72.98</td>
<td>90.96</td>
<td>84.47</td>
</tr>
<tr>
<td>w Target</td>
<td>81.10</td>
<td>71.44</td>
<td>89.00</td>
<td>82.97</td>
</tr>
</tbody>
</table>

Table 8 shows the results of the PromptHate performance with and without target information. We observe marginal differences in performance after modelling target information in prompts. A possible reason may be that learning to extract targets in memes adds auxiliary burden to the model. To better utilize target information, a more sophisticated strategy may be needed than the current simple approach.

In Table 9, we visualize PromptHate’s prediction results on sample FHM memes. Incorrect predictions are labelled in red while the pie chart presents the distributions of the predicted target (i.e., [TARGET_MASK]) per meme. PromptHate with target information is observed to correctly predict the targets in the hateful meme even when it incorrectly classifies the memes (e.g., third meme targeting religion). The right-most meme contains a racial slur ‘Kenyan skidmark’ and seems to have been annotated wrongly as non-hateful. Interestingly, PromptHate with target information indicates it as hateful and targeting race.

The target distributions can improve PromptHate’s interpretability. However, the incorrect class prediction also highlights the difficulty of hateful meme classification. The task may require more than target comprehension to achieve good performance.

5.5 Error Analysis

Besides analyzing PromptHate’s quantitative performance, we also examine its classification errors. Table 10 illustrates three selected PromptHate’s incorrect predictions.

From the examples, we notice that the captions generally describe the contents of images. However, it may ignore some essential attributes for hateful meme detection. For instance, the captions are unable to capture important information such as “Jesus”. This missing information is supplemented by the augmented image tags (i.e., the entities and demographic of memes). Nevertheless, we also observed that even after augmentation with additional descriptions for the images, PromptHate still makes incorrect predictions for these memes.

There could be multiple reasons for the incorrect predictions. Firstly, the presented information may still lack adequate context. For instance, in the second meme, the “biting” or “eating” action is missing from the captions and the addition image description. Thus, PromptHate lacks the context that the meme is ridiculing Asians’ “dog-eating” behaviour, and is hateful. Secondly, there could be biases learned by the model. For instance, PromptHate may predict the right-most meme to be hateful because of the rainbow flag, an icon for the LGBT community. This icon may be heavily used by memes attacking the LGBT community. Lastly, even more advanced reasoning may be required to understand the hateful context in certain meme. In the first case, PromptHate is unable to reason that the meme implies that Jesus is merely an object hung up to scare away birds, which leads to the hateful context of the meme. The detection of the hateful meme requires deep multi-modal reasoning.

6 Conclusion

We have proposed PromptHate, a simple yet effective multimodal prompt-based framework that prompts pre-trained language models for hateful meme classification. Our evaluations on two publicly available datasets have shown PromptHate to outperform state-of-the-art baselines. We have conducted fine-grained analyses and case studies on various prompt settings and demonstrated the effectiveness of the prompts on hateful meme classification. We have also performed error analysis and discussed some limitations of the PromptHate model. For future work, we will explore strategies for selecting better demonstrations for PromptHate and add in reasoning modules to improve PromptHate’s utilization of the implicit knowledge in the PLMs.

7 Limitations

We have discussed a number of PromptHate’s limitations in our error analysis (See Section 5.4). Specifically, we noted that in some instances, although PromptHate is able to tap into the implicit knowledge in PLM to improve hateful meme clas
Table 9: Example predictions of PromptHate with and without target information. Incorrect prediction in red. Ground truth for the right-most meme is questionable.

<table>
<thead>
<tr>
<th>Meme</th>
<th>Target Distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w Target.</td>
</tr>
<tr>
<td></td>
<td>w/o Target</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>Hateful (race)</td>
</tr>
</tbody>
</table>

Table 10: Error analysis of wrongly predicted memes. Incorrect prediction in red.

<table>
<thead>
<tr>
<th>Meme</th>
<th>Ground Truth</th>
<th>Prediction</th>
<th>Meme text</th>
<th>Captions</th>
<th>Entity</th>
<th>Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hateful (religion)</td>
<td>Non-hateful</td>
<td>the original scarecrow</td>
<td>when you date an asian boy and you trynna get his family to accept you</td>
<td>builder crucified on the cross</td>
<td>Crucifix Life, Resurrection of Jesus, Spiritual death, jesus died in the cross</td>
</tr>
<tr>
<td></td>
<td>Hateful (race)</td>
<td>Non-hateful</td>
<td>when you date an asian boy and you trynna get his family to accept you</td>
<td>when you date an asian boy and you trynna get his family to accept you</td>
<td>portrait of a young woman with her pet dog</td>
<td>Rainbow flag, Flag bisexual, Transgender flags, Bisexuality</td>
</tr>
<tr>
<td></td>
<td>Non-Hateful</td>
<td>Hateful</td>
<td>you either die a hero, or live long enough to become the villain</td>
<td>you either die a hero, or live long enough to become the villain</td>
<td>a man dressed as a rainbow holding a flag and dancing in the crowd</td>
<td>None</td>
</tr>
</tbody>
</table>

| Demographics | None | Black female | Latino Hispanic Male |

sification, it still lacks the ability to perform advanced reasoning on the contextual information to arrive at correct predictions. We also observe that PromptHate may learn biases from the training data, and debiasing techniques may be required to improve its performance.

**References**

International Joint Conference on Natural Language Processing, ACL/IJCNLP, pages 3816–3830.


APPENDIX

A Experiment Settings

We train all models using Pytorch on an NVIDIA Tesla V100 GPU, with 32 GB dedicated memory, CUDA-10.2. For pre-trained models (i.e., BERT, RoBERTa, VisualBERT), we use the package, transformers (version 4.19.2) from Huggingface. Table 11 lists the parameter count for all models.

<table>
<thead>
<tr>
<th>Method</th>
<th># Params (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text BERT</td>
<td>109.9</td>
</tr>
<tr>
<td>Image Region</td>
<td>1.0</td>
</tr>
<tr>
<td>Late Fusion</td>
<td>110.9</td>
</tr>
<tr>
<td>Concat BERT</td>
<td>111.8</td>
</tr>
<tr>
<td>MMBT-Region</td>
<td>111.5</td>
</tr>
<tr>
<td>Visual BERT COCO</td>
<td>111.8</td>
</tr>
<tr>
<td>ViLBERT CC</td>
<td>252.1</td>
</tr>
<tr>
<td>CLIP BERT</td>
<td>111.7</td>
</tr>
<tr>
<td>MOMENTA</td>
<td>71.9</td>
</tr>
<tr>
<td>DisMultiHate</td>
<td>115.6</td>
</tr>
<tr>
<td>FT-RoBERTa</td>
<td>356.4</td>
</tr>
<tr>
<td>PromptHate</td>
<td>355.4</td>
</tr>
</tbody>
</table>

Table 11: Number of parameters in VQA models.

The learning rates of models are set empirically. For BERT based models, the learning rate is set to be $2 \times 10^{-5}$, the same as in (Lee et al., 2021). For RoBERTa-large based models (PromptHate and FT-RoBERTa), following (Gao et al., 2021), we tested learning rate ranging from $10^{-5}$ to $1.5 \times 10^{-5}$ and reported the best ones. Specifically, the learning rate is set to be $1.3 \times 10^{-5}$ and $10^{-5}$ on FHM and HarM datasets, respectively. AdamW is used as the optimizer for all models. The mini-batch size is set at 16 during training. As mentioned in Section 4.4, we apply the multi-query ensemble strategy. The number of querying times is set at 4 on both datasets. It takes one GPU six minutes to train and validate PromptHate per epoch. It takes up 19 GB dedicated memory for PromptHate training. We use 10 training epochs for both PromptHate and baselines.

B Analysis for Image Captions

A key data-preprocessing step in PromptHate is to covert the image into textual captions as input for PLMs. Therefore, the quality and expressiveness of the image captions may affect the prompting and ultimately affect the hateful meme classification performance. To investigate this effect, we experiment with image captions generated with ClipCap (Mokady et al., 2021) pre-trained on different datasets, namely, MS COCO (Lin et al., 2014; Chen et al., 2015) and Conceptual Caption (CC) (Sharma et al., 2018).

Table 12 shows PromptHate’s performance with image captions generated using ClipCap pre-trained on COCO (ClipCap+COCO) and CC (ClipCap+CC). We observe that the ClipCap pre-trained on CC performs better than that pre-trained with COCO. A possible reason could be that the CC dataset is mainly images from web pages and rather more similar to meme images. For instance, considering the examples in Table 13, we notice that ClipCap pre-trained on CC provided more detailed descriptions (e.g., the relation of the man and the woman of the first meme and the characteristic of the sign in the second meme) compared to COCO. On the other hand, we test the generated captions using the uncleaned (i.e., without removing meme texts on images) meme images (ClipCap+CC (UC)). We notice that when trained with Conceptual Captions, ClipCap+CC (UC) is still able to generate some details (i.e., the characteristic of the sign in the second example). It sometimes generates comments rather than captions that describe images. It is because Conceptual Captions are from the web, and some of them are comments on meme images. Without removing texts, models will link the image to meme images and generate comments rather than captions. The difference in their performance is also significant, suggesting the importance of good quality captions in applying prompt-based models for hateful meme classification.

<table>
<thead>
<tr>
<th>Meme</th>
<th>ClipCap +COCO</th>
<th>ClipCap +CC (UC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man and a woman</td>
<td>a sign that is on the side of a hill.</td>
<td>a warning sign at the entrance to the quarry.</td>
</tr>
<tr>
<td>are in a kitchen.</td>
<td>thank you for the dishes!</td>
<td></td>
</tr>
<tr>
<td></td>
<td>young couple in love looking at each other in kitchen.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>from the web.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Table 13: Example captions generated for meme images.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6https://huggingface.co/