Prompt-enhanced Network for Hateful Meme Classification

Junxi Liu[1] , **Yanyan Feng**[1] , **Jiehai Chen**[1] , **Yun Xue**[1*] and **Fenghuan Li**[2*]
[1]School of Electronics and Information Engineering, South China Normal University, Foshan 528225, China,
[2]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China
{liujunxi,fengyanyan,cjh_scnu,xueyun}@m.scnu.edu.cn, fhli20180910@gdut.edu.cn

Abstract

The dynamic expansion of social media has led to an inundation of hateful memes on media platforms, accentuating the growing need for efficient identification and removal. Acknowledging the constraints of conventional multimodal hateful meme classification, which heavily depends on external knowledge and poses the risk of including irrelevant or redundant content, we developed Pen—a prompt-enhanced network framework based on the prompt learning approach. Specifically, after constructing the sequence through the prompt method and encoding it with a language model, we performed region information global extraction on the encoded sequence for multi-view perception. By capturing global information about inference instances and demonstrations, Pen facilitates category selection by fully leveraging sequence information. This approach significantly improves model classification accuracy. Additionally, to bolster the model's reasoning capabilities in the feature space, we introduced prompt-aware contrastive learning into the framework to improve the quality of sample feature distributions. Through extensive ablation experiments on two public datasets, we evaluate the effectiveness of the Pen framework, concurrently comparing it with state-of-the-art model baselines. Our research findings highlight that Pen surpasses manual prompt methods, showcasing superior generalization and classification accuracy in hateful meme classification tasks. Our code is available at https://github.com/juszzi/Pen.

1 Introduction

With the evolution of the internet, social media has emerged as the primary mode of communication, information sharing, and expressing opinions. The rise of social media has introduced a new multimodal entity – memes, comprised of images and short texts. While this form has gained popularity on social media networks, it has also become a tool for some users to disseminate hate speech, causing serious harm to vulnerable groups [Piccoli *et al.*, 2024]. Due to the rapid dissemination of hateful memes, there is an urgent need to develop accurate classification methods. Figure 1 illustrates examples of hateful and non-hateful memes.

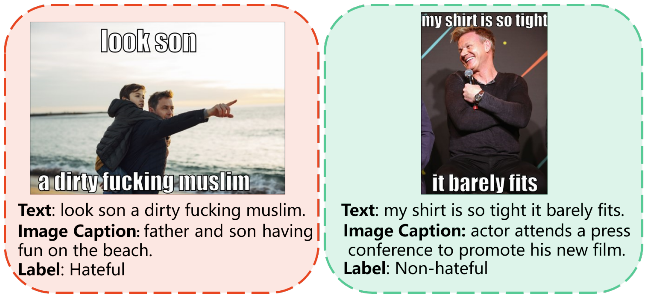
look son
a dirty fucking muslim
Text: look son a dirty fucking muslim.
Image Caption: father and son having fun on the beach.
Label: Hateful
my shirt is so tight
it barely fits
Text: my shirt is so tight it barely fits.
Image Caption: actor attends a press conference to promote his new film.
Label: Non-hateful

Figure 1: The red box indicates a sample labeled as "hateful", while the green box indicates a sample labeled as "non-hateful".

To address this issue, early efforts emphasized the alignment and fusion across modalities to recognize hateful memes in social media [Zhu, 2020; Muennighoff, 2020; Velioglu and Rose, 2020]. Recognizing the need for intricate reasoning and contextual background knowledge in determining hatred in memes, attempts were made to enhance model classification accuracy by integrating external tools [Zhou *et al.*, 2021] or incorporating additional external knowledge [Lee *et al.*, 2021] within the visual language model framework. Building upon this foundation, subsequent research began considering a modality transformation perspective. Cao *et al.* transformed meme images into image captions, employing prompting methods and introducing external knowledge to guide Pre-trained language models (PLM) in predictions [Cao *et al.*, 2022]. In the latest work, building upon their prior work [Cao *et al.*, 2022], Cao *et al.* enhanced the quality of image Captions. They employed zero-shot visual question answering (VQA) with pre-trained vision-language models (PVLMs) for generating image captions [Cao *et al.*, 2023]. This enhancement led to superior image caption quality, achieving a state-of-the-art results in the current domain.

However, recent strategies for classifying hateful memes tend to emphasize improving model performance through the incorporation of additional external knowledge, potentially neglecting issues related to irrelevant or redundant content

[*] Corresponding author

within such knowledge[Lee *et al.*, 2021; Blaier *et al.*, 2021; Cao *et al.*, 2022; Fang *et al.*, 2022]. For instance, incorporating image entity recognition information [Lee *et al.*, 2021] may introduce entities that are unrelated or redundant to hateful memes, thereby adding irrelevant details that could interfere with the model's classification judgment. While some studies utilize prompting methods to guide PLM in leveraging external knowledge [Cao *et al.*, 2022], this approach predominantly focuses on the data processing stage. It enhances the contextual learning capabilities of language models for classification by introducing prompt template tokens and demonstrations of different categories to the original sequence. However, it does not comprehensively address the training conditions of the sequence in the feature space.

Hence, our focus lies in extracting valuable information through a simple and effective network mechanism, enabling the PLM to adaptively select pertinent information for hateful meme classification. Existing prompt method guides PLM in classification by providing demonstrations corresponding to each label. Given the demonstrations for each label, there should be specific feature-level connections in the feature space between the contextual information of inference instances and the contextual information corresponding to the demonstrations of their correct labels. Building upon this concept, we extend prompt method into the feature space, introducing a novel framework called the **P**rompt-**en**hanced network for hateful meme classification (**Pen**). In this framework, we initially process the sequences of the input PLM with prompts, followed by region segmentation. We extract global information features from both inference instance and demonstration regions, incorporating the prompt-enhanced multi-view perception module. This module perceives the global information features of inference instances and demonstrations from multiple views to make hate emotion judgments, enhancing the model's classification accuracy by effectively utilizing contextual information in input sequences. To better capture the relationships between hate and non-hate in the feature space, we introduce contrastive learning and adapt it to our framework, forming prompt-aware contrastive learning. This adaptation enhances the quality of the feature distribution for samples. In summary, the primary contributions of this paper are as follows:

- We propose a model framework named Pen, which extends the prompt method into the feature space. By incorporating multi-view perception of inference instances and demonstrations in the feature space, Pen enhances hate classification accuracy, thereby improving the utilization of sequences.

- We propose a contrasting learning method compatible with manual prompting to align and differentiate sample features used for hate judgment. This method sharpens the features of samples from different categories, thereby improving the accuracy of the classifier.

- Through extensive ablation experiments conducted on two publicly available datasets, we validate the effectiveness of the prompt-enhanced framework, demonstrating its superiority over state-of-the-art baselines.

2 Related Work

2.1 Multimodal Hateful Meme Classification

Multimodal hateful meme classification aims to detect hateful implications in both text and images within memes. This task was initially introduced by the Hateful Memes Challenge (HMC) competition [Kiela *et al.*, 2020]. Researchers, including Kiela *et al.*, conducted a series of experiments on this dataset, involving both unimodal and multimodal models, with superior performance observed in multimodal approaches. Subsequent studies [Suryawanshi *et al.*, 2020; Muennighoff, 2020; Zhu, 2020; Velioglu and Rose, 2020; Pramanick *et al.*, 2021b] delved into exploring enhanced modality fusion methods, incorporating features learned from text and visual encoders using attention mechanisms and other fusion techniques.

Given the complexity of inferring hate in memes and the need for contextual background knowledge, recent research has started exploring approaches that integrate external knowledge to assist in hateful meme classification. Attempts have been made to augment the model's input with relevant external knowledge [Zhou *et al.*, 2021; Lee *et al.*, 2021] to enhance the classification and interpretability of hateful content. Cao *et al.* transformed meme images into image captions, introducing external knowledge and using prompting methods to guide PLM in prediction [Cao *et al.*, 2022]. In the latest approach, Cao *et al.* improved the quality of image Captions based on modality transformation, employing zero-shot VQA with PVLMs for image. This enhancement achieved SOTA results in the current research landscape. However, recent solutions, while supplementing external knowledge for assisting model judgments, still overlook irrelevant or redundant content within this knowledge. Thus, the effective and adaptive utilization of such external knowledge remains an urgent issue.

2.2 Prompt For Hateful Meme Classification

The natural approach to creating manual prompts involves using prompts that include task-specific descriptions and textual demonstration in a natural language manner as inputs for the model. For instance, in the case of sentiment classification for the movie review "The film offers an intriguing what-if premise", a prompt template, such as "the sentiment of this review is [mask]", can be added during data processing. Positive and negative examples are then appended to the sequence after the augmented prompt, allowing the language model to classify the [mask] token. A successful classification result should indicate a positive sentiment. Cao *et al.* pioneered the use of prompt methods in multimodal hateful meme classification to guide PLM in hate reasoning. In a study by He *et al.*, the combination of large language models and prompt learning was explored to address toxic content detection, demonstrating the effectiveness of prompt methods in hate detection tasks [He *et al.*, 2023]. Despite manual prompts being proven to solve various tasks with considerable accuracy [Liu *et al.*, 2023], manual prompt methods only process the input sequence to guide the natural language reasoning ability of PLM. The uncertainty remains about whether models can effectively assimilate sequences enhanced through prompt methods. Therefore, enhancing the model's utilization of sequence information remains a pressing issue.

2.3 Contrastive Learning

In the field of natural language processing, contrastive learning has gained significant traction in various studies[Zhang *et al.*, 2022; Jian *et al.*, 2022; Liang *et al.*, 2022; Qu *et al.*, 2023]. Zhang *et al.* introduced contrastive learning in multitask pretraining, leveraging unlabeled data clustering to obtain self-supervised signals and achieving optimal results in intent detection tasks. Jian *et al.* combined contrastive loss from prompt-based few-shot learners with standard Masked Language Modeling (MLM) loss. Liang *et al.* applied target-aware prototype-based contrastive learning in zero-shot stance detection tasks.Recently, Qu *et al.* explored the application of contrastive learning for hate classification, utilizing a multimodal contrastive learning model to unsupervisedly identify the primary groups associated with potential hateful memes. Currently, contrastive learning has demonstrated significant effectiveness across various tasks, improving model performance in classification tasks. However, the current approaches primarily rely on simple label-oriented sample feature clustering or sample-driven self-supervised contrastive learning. The goals of contrastive learning applications are relatively narrow. Exploring additional forms of contrastive learning methods to enhance sample features is expected to enable models to learn diverse information, thereby improving model performance.

3 Methodology

Problem Definition We define the hateful meme classification task as a series of binary tuples represented by the entity $M = \{T, I\}$, where text T and image I are interrelated. Our objective is to train a model to assess these tuples and output either "hateful" or "non-hateful". Following the framework proposed by Cao *et al.*, we transform the multimodal setting into an unimodal one. Utilizing the image-to-text tool ClipCap[Mokady *et al.*, 2021], we convert images into image captions. After concatenating the text with the image caption, we introduce a prompt template "it was [mask]", along with demonstrations and external knowledge, incorporating them into a PLM. The language model then evaluates the [mask] token, selecting the appropriate label as the output.

In this study, our core idea is to extend the prompt method, applying the concept of prompt methods in the feature space to strengthen the connection between inference instances and demonstrations. By incorporating information from the entire sequence, we aim to improve the classification effectiveness of the language model. In this section, we will provide a detailed overview of our approach to handling the hateful meme classification task. Figure 2 illustrates the structure of our proposed Pen framework, comprising Regional Information Global Extraction (Section 3.1), Prompt-enhanced Multi-view Perception (Section 3.2), and Prompt-aware Contrastive Learning (Section 3.3).

3.1 Regional Information Global Extraction

During the data processing stage, we started by randomly selecting demonstrations of both hateful and non-hateful instances from the train set. To enhance the PLM understanding of the content in the inference instance and facilitate a more effective perception between the inference instance and demonstrations for category determination, we needed to extract global information from the input inference instance and demonstrations. Due to the variable sequence lengths caused by the indeterminate nature of past sequence concatenation methods, it became necessary to perform region segmentation on the input model's sequence.

Figure 3 illustrates the composition of the input sequence. The blue region, denoted as s^{infer}, encompasses information related to inference instances, including the text and image captions of the meme requiring inference, as well as external knowledge about the meme. The red region s^{neg} and the green region s^{pos} correspond to the information selected for hateful and non-hateful demonstrations, respectively. They share the same information structure as s^{infer}. The orange region (p^{infer}, p^{neg} and p^{pos}) corresponds to the prompt template. Each region has a fixed maximum length. If the length falls short, padding is applied to reach the maximum length, and if it exceeds the maximum length, truncation is performed, ensuring fixed positioning of each region. This facilitates the extraction of global information from each region and strengthens the PLM's understanding of the overall sequence. Recognizing the significance of information in the inference instances during the prediction phase, we appropriately extended the length of the inference instance region while relatively shortening the demonstration regions. This ensures that the inference instance region contains sufficient information. The sequence composition is as follows:

$$L = [Start][s^{infer}, p^{infer}][S][s^{neg}, p^{neg}][S][s^{pos}, p^{pos}][S] \quad (1)$$

Here, L represents the processed sequence through the prompt method, and it is fed into the PLM. $[Start]$ is the starting token in L, and $[S]$ serves as the separator in L.

Next, we feed L into a PLM. Specifically, we employ the Roberta-large model[Liu *et al.*, 2019] to obtain the overall embedding features $E \in \mathbf{R}^{n \times d}$, where d represents the dimension of the hidden layers in the PLM, and n denotes the length of the entire sequence. The process is illustrated as follows:

$$\begin{aligned} E &= LanguageModel(L) \\ &= [Start][e^{infer}, e_p^{infer}][S][e_p^{neg}, p^{neg}][S][e^{pos}, e_p^{pos}][S] \end{aligned} \quad (2)$$

Next, we employed Long Short-Term Memory (LSTM) networks to extract global information from the encoded representations of the three regions (e^{infer}, e^{neg}, and e^{pos}), resulting in global information for inference instances and demonstrations: t^{infer}, t^{neg}, and t^{pos}.

3.2 Prompt-enhanced Multi-view Perception

Due to the fact that the label token in the prompt template corresponding to the demonstration within sequence L already indicates the category, we contemplate incorporating features of special tokens in the prompt template (as highlighted in bold in the origin region of Figure 3) to enhance the hateful-related features in both the global information of the inference instance and the global information of the demonstration.

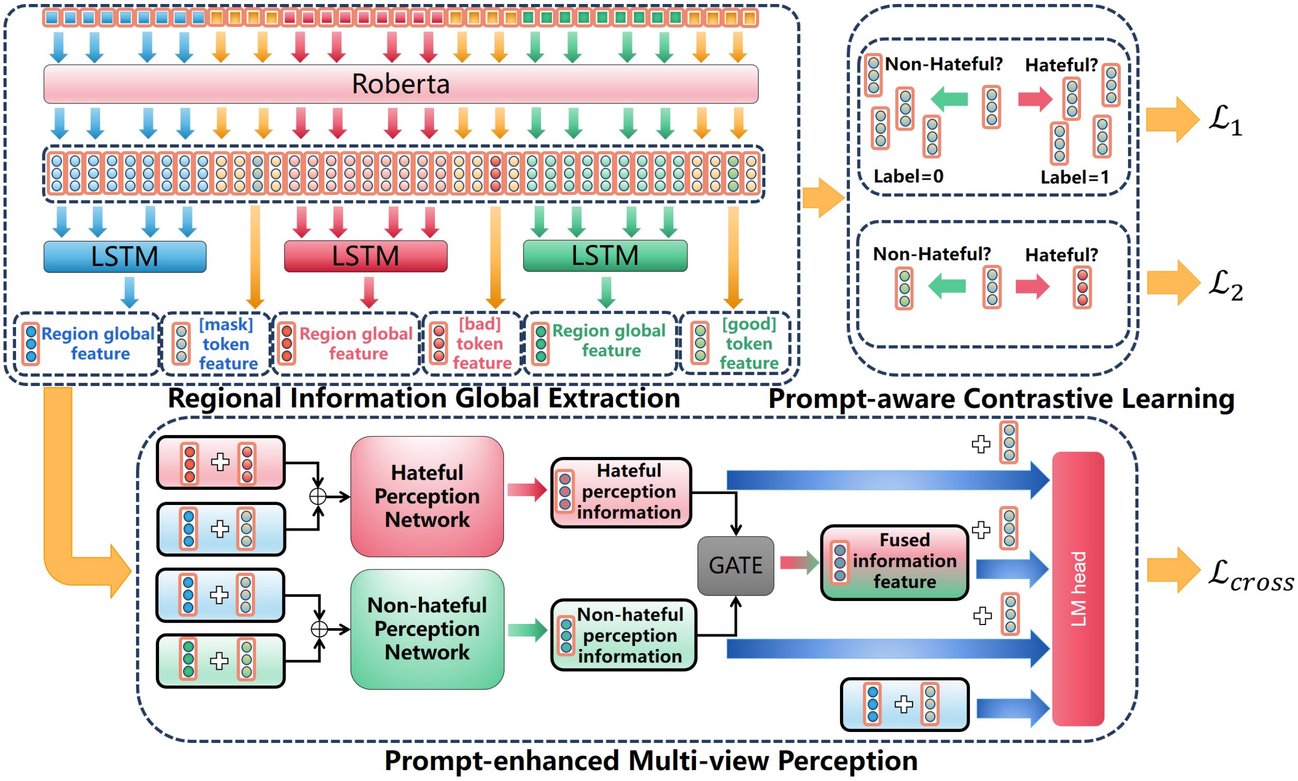

Figure 2: Overview of Pen Framework

Given the fixed length of each region, the position of special tokens remains constant in the sequence, allowing for the direct extraction of feature vectors $t^{infer}_{specical}$, $t^{neg}_{specical}$, and $t^{pos}_{specical}$, extracted from the orange regions e^{infer}_{p}, e^{neg}_{p}, and e^{pos}_{p}, respectively. Subsequently, the feature vectors of global information for the inference instance and demonstration are fused with their corresponding special token feature vectors through a simple merging process. These paired vectors are then input into the hateful perception network and non-hateful perception network, facilitating the learning of relationships between the inference instance and both hateful and non-hateful demonstrations. Ultimately, the obtained hateful perception information I^{mix}_{0} and non-hateful perception information I^{mix}_{1} are fed into a soft gating mechanism to derive the ultimate fused information feature $\hat{I}^{mix}$. The specific fusion operations are illustrated in the following formulas:

$$I^{mix}_{0} = HPN((t^{infer} + t^{infer}_{special}) \oplus (t^{neg} + t^{neg}_{special})) \quad (3)$$

$$I^{mix}_{1} = NHPN((t^{infer} + t^{infer}_{special}) \oplus (t^{pos} + t^{pos}_{special})) \quad (4)$$

$$\hat{I}^{mix} = GATE(I^{mix}_{0}, I^{mix}_{1}) \quad (5)$$

The symbol $\oplus$ represents concatenation. HPN and $NHPN$ stand for Hateful Perception Network and Non-Hateful Perception Network, respectively, comprising fully connected layers with trainable parameters. $GATE$ represents a soft gating mechanism constructed by fully connected layers with trainable parameters, designed to control the fusion of I^{mix}_{0} and I^{mix}_{1}.

The aforementioned information fusion process thoroughly learns the feature information between inference instance and hateful and non-hateful demonstrations. However, to more accurately assess whether the inference instance contains hateful elements, we not only employ the fused information feature $\hat{I}^{mix}$ for classification but also introduce the hateful perception information I^{mix}_{0}, non-hateful perception information I^{mix}_{1}, and inference instance information t^{infer}. This multi-view perception contributes to the final classification result, enhancing accuracy. Considering that PLM use the [mask] token during pre-training to predict the probability distribution of masked words, when utilizing a linear classifier for classification, we supplement the features of the [mask] token as $t^{infer}_{special}$. The multi-view perception process is outlined in the following equations:

$$\begin{aligned} S_{all} &= s_1 + s_2 + s_3 + s_4 \\ &= LMhead(t^{infer} + t^{infer}_{special}) + LMhead(I^{mix}_{0} + t^{infer}_{special}) \\ &+ LMhead(I^{mix}_{1} + t^{infer}_{special}) + LMhead(\hat{I}^{mix} + t^{infer}_{special}) \end{aligned} \quad (6)$$

Here, each element $(s_1, s_2, s_3, s_4, S_{all})$is individually composed of the binary tuple $(score_{hateful}, score_{non-hateful})$. $LMhead$ represents a linear classifier composed of trainable parameters in a fully connected layer. It is utilized to generate the probability scores, $score_{hateful}$ and $score_{non-hateful}$, indicating the likelihood of a sample being hateful or non-hateful, respectively.

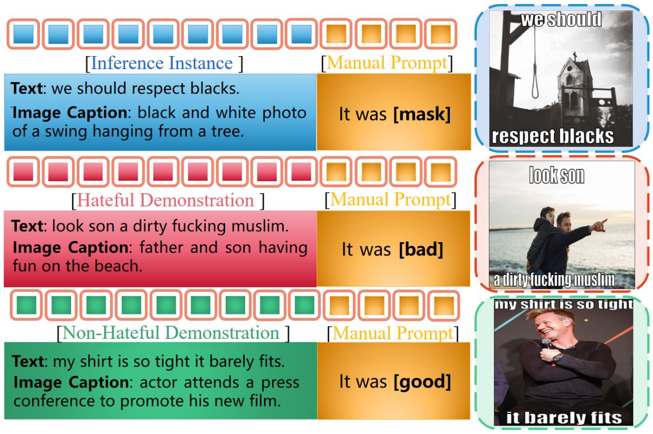
[Inference Instance]
[Manual Prompt]
Text: we should respect blacks.
Image Caption: black and white photo of a swing hanging from a tree.
It was [mask]
we should
respect blacks
[Hateful Demonstration]
[Manual Prompt]
Text: look son a dirty fucking muslim.
Image Caption: father and son having fun on the beach.
It was [bad]
look son
a dirty fucking muslim
[Non-Hateful Demonstration]
[Manual Prompt]
Text: my shirt is so tight it barely fits.
Image Caption: actor attends a press conference to promote his new film.
It was [good]
my shirt is so tight
it barely fits

Figure 3: Input Sequence Components

3.3 Prompt-aware Contrastive Learning

After obtaining the total score S_{all} regarding hate and non-hate for a given sample, we select the highest score as the final result for category determination. In the model training process, we utilize the cross-entropy loss $\mathcal{L}_{cross}$ to train the model. However, to further enhance the model's understanding of the relationship between hatred and non-hatred at the feature level, we incorporate contrastive learning to improve the quality of feature distribution for samples.

Category-oriented Contrastive Learning

During the model training process, for a batch of samples, the mask feature vectors $t^{infer}_{special}$ corresponding to hateful and non-hateful inference instances actually exhibit certain distinctions. For mask feature vectors corresponding to samples of the same category, their distances in the feature space should tend to be close, while for mask feature vectors corresponding to samples of different categories, their distances in the feature space should tend to be increased.

Since $t^{infer}_{special}$ is used for perception during classification, we can leverage label information for contrastive learning during the training process. This is done to enhance the feature discriminability of different categories of $t^{infer}_{special}$. For a batch of samples, mask feature vectors with the same label are treated as positive examples, while mask feature vectors with different labels are treated as negative examples. This helps bring positive examples closer together and push negative examples farther apart:

$$\mathcal{L}_1 = -\frac{1}{M}\sum_{i=1}^{M}\log(\frac{(\sum_{j=1}^{M}\zeta_{[y_i=y_j]}\cdot sim(t^{infer}_{special_i}, t^{infer}_{special_j}))/\tau_1}{(\sum_{k=1}^{M} sim(t^{infer}_{special_i}, t^{infer}_{special_k}))/\tau_1}) \tag{7}$$

Here, M denotes the number of samples in a batch, sim represents the calculation of cosine similarity, $\zeta_{[y_i=y_j]}$ is used to determine whether samples i and j belong to the same category, where it is 1 if they do, τ_1 is the temperature coefficient, and y_i denotes the label of the $i-th$ sample.

Prompt-oriented Contrastive Learning

For an individual sample during the training process, the [mask] feature vector $t^{infer}_{special}$ corresponding to the inference instance should be closer to the special token feature vector of demonstrations with the same label, while being distinct from the special token feature vector corresponding to demonstrations with different labels. For instance, in training, the [mask] token $t^{infer}_{special}$ associated with an inference instance labeled as hateful should tend to be close to the [bad token $t^{neg}_{special}$ in the vector space, and distant from the [good] token $t^{pos}_{special}$.

For each sample in a batch, the $t^{infer}_{special}$ corresponding to the inference instance region in the sample's sequence is considered as a positive example, paired with the label feature vector from the region of demonstrations with the same class. Simultaneously, it is treated as a negative example when paired with the label feature vector from the region of demonstrations with different class labels. This process serves to minimize the distance between positive examples and maximize the distance between negative examples, thereby expediting the aggregation and divergence process of $t^{infer}_{special}$.

$$\mathcal{L}_2 = -\frac{1}{M}\sum_{i=1}^{M}\log(\frac{(\sum_{j=1}^{2}\zeta_{[y_i=y^p_j]}\cdot sim(t^{infer}_{special_i}, t^{prompt}_{special_j}))/\tau_2}{(\sum_{k=1}^{2} sim(t^{infer}_{special_i}, t^{prompt}_{special_k}))/\tau_2}) \tag{8}$$

Here, y^p represents the label of the demonstration in the sample, $t^{prompt}_{special}$ represents the special token corresponding to the label, either $t^{neg}_{special}$ or $t^{pos}_{special}$, τ_2 serves as the temperature coefficient, and y_i represents the label of the $i-th$ sample. Finally, the overall loss for our approach is:

$$Loss = \mathcal{L}_{cross} + \alpha * \mathcal{L}_1 + \beta * \mathcal{L}_2 \tag{9}$$

Where, α and β are hyperparameters representing the weights assigned to different sub-losses.

4 Experimental setup

4.1 Datasets

We conducted evaluations using two publicly available datasets: (1) FHM [Kiela *et al.*, 2020] and (2) HarM [Pramanick *et al.*, 2021a]. The FHM dataset, developed and released by Facebook, is part of a crowdsourced multimodal hateful meme classification challenge. The HarM dataset consists of real memes related to COVID-19 collected from Twitter, categorized into three classes: very harmful, partially harmful, and non-harmful. Following the evaluation setup of Cao *et al.*, we merged the very harmful and partially harmful categories into the harmful category. Due to the generality of our approach, our framework can utilize preprocessed image captions and external knowledge from both the Prompthate [Cao *et al.*, 2022] and Pro-Cap [Cao *et al.*, 2023] methods. The preprocessed information includes text on images, image captions (where Prompthate employs the ClipCap [Mokady *et al.*, 2021] tool for image captioning, as illustrated in Figure 1, and Pro-Cap employs zero-shot VQA with BLIP-2[Li *et al.*, 2023] to ask questions and generate content-centric hateful image captions, covering various aspects such as race, gender, religion, nationality, disability, and animals). Additionally, other external knowledge is information about entities in the images and racial features [Kärkkäinen and Joo, 2021]. We present the statistical summary of the datasets in Table 1.

Datasets	**# Training**		**# Test**	
	Hate	**Non-hate**	**Hate**	**Non-hate**
FHM	3050	5450	250	250
HarM	1064	1949	124	230

Table 1: Statistical summary of FHM and HarM.

4.2 Baseline Method

In this section, we present a comprehensive comparison of Pen with state-of-the-art models for hateful meme classification. We categorize the baseline methods into two groups: unimodal and multimodal methods.

For the unimodal methods, we adopt a text-only strategy using **Text-Bert** [Devlin *et al.*, 2019] and an image-only model known as **Image-Region**. The latter employs Faster R-CNN[Ren *et al.*, 2017] and ResNet-152 [He *et al.*, 2016] for meme image processing, with resulting representations fed into a hate classification classifier. Moving to multimodal methods, we explore diverse approaches, including **Late Fusion** [Pramanick *et al.*, 2021a], **MMBT-Region** [Kiela *et al.*, 2019], **ViLBERT CC** [Lu *et al.*, 2019], and **Visual BERT COCO** [Li *et al.*, 2019]. Additionally, we compared with recent hateful meme classification methods: **MOMENTA** [Pramanick *et al.*, 2021b], **Prompthate** [Cao *et al.*, 2022], and **Pro-Cap** [Cao *et al.*, 2023]. We utilized accuracy and macro-averaged F1 scores as evaluation metrics. The significance of macro-averaged F1 is emphasized due to the imbalanced class distribution in the two datasets (refer to Table 1), necessitating a comprehensive assessment of performance across all classes to capture overall performance. To ensure a fair comparison, we averaged the model performance over ten random seeds, considering the average across ten runs for each method.

4.3 Experimental Results

Table 2 presents the experimental results of baseline methods and our framework on the HarM and FHM datasets. In this table, Pen denotes the use of the preprocessed dataset from [Cao *et al.*, 2022], while Pen$_{Cap}$ represents the usage of the preprocessed dataset from [Cao *et al.*, 2023]. From the experimental results, it can be observed that our proposed Pen framework achieves a higher macro-average F1 score on the HarM and FHM datasets compared to the Prompthate method, which solely relies on prompt methods, with increases of 2.85% and 1.56%, respectively, under the same data conditions. Furthermore, under same data conditions, Pen$_{Cap}$ outperforms the Pro-Cap method by 1.85% and 0.66% in terms of macro-average F1 score on these two datasets. This effectively demonstrates the efficacy of our prompt-enhanced framework. Pen adeptly refines the features of input sequences, as well as strengthens the connection between inference instances and demonstrations, and extracts crucial information to assist the model in hate detection, guiding the model to find useful information in the feature space. This represents an enhancement of prompt methods in the feature space. Interestingly, from the experimental results, it is observed that Pen$_{Cap}$ exhibits a significant performance improvement on the FHM dataset compared to Pen. However, the improvement on the HarM dataset is not as pronounced. We speculate that the smaller scale of the HarM dataset, which only includes hate elements related to COVID-19, leads to a more singular hate element in the dataset samples. Consequently, Pen is able to make accurate hate judgments based on the existing information during training. In contrast, the FHM dataset comprises multiple hate factors with higher quality, requiring a corresponding increase in hate-related information. Thus, Pen$_{Cap}$ achieves a substantial performance improvement on the FHM dataset compared to Pen. This also suggests that achieving further performance improvements on the FHM dataset requires richer external knowledge support.

Method	**HarM**		**FHM**	
	Acc	Marco-F_1	Acc	Marco-F_1
Text BERT	70.17	66.25	57.12	41.52
Image-Region	68.74	62.97	52.34	34.19
Late Fusion	73.24	70.25	59.14	44.81
MMBT-Region	73.48	67.12	65.06	61.93
VisualBERT	81.36	80.13	61.48	47.26
ViLBERT CC	78.70	78.09	64.70	55.78
MOMENTA	83.82	82.80	61.34	57.45
Prompthate	84.47	82.42	72.98	71.99
Pen(ours)	**86.30**	**85.27**	**74.04**	**73.55**
Pro-Cap	85.06	83.89	74.72	74.59
Pen$_{Cap}$(ours)	**86.92**	**85.74**	**75.46**	**75.25**

Table 2: Hateful meme classification results on two datasets. accuracy and macro-averaged F1 score (%) are reported as evaluation metrics, averaged over ten runs, with the best results highlighted in bold.

Setting	**HarM**		**FHM**	
	Acc	Marco-F_1	Acc	Marco-F_1
Pen	86.30	85.27	74.04	73.55
w/o PMP	85.00	84.16	72.80	72.25
w/o $\mathcal{L}_1$	86.24	85.08	73.90	73.33
w/o $\mathcal{L}_2$	86.19	85.06	73.94	73.40
w/o PCL	85.68	84.48	73.80	73.30

Table 3: Ablation study of Pen.

4.4 Ablation Study

To investigate the effectiveness of different modules in Pen, we conducted ablation experiments in four different forms, aligning with the structure of the Pen framework. The results from the ablation experiments in Table 3 reveal that removing the prompt-enhanced multi-view perception module (w/o PMP) significantly degrades Pen's performance on both datasets. This underscores the efficacy of the PMP module in refining sequence features, directing the model's attention to the connections between inference instances and demonstrations. The PMP module perceptually engages inference instances with demonstrations, extracting hateful-related features and consequently enhancing classification accuracy. The results also show that omitting prompt-aware contrastive learning (w/o PCL) leads to a substantial performance drop for Pen on the HarM dataset, while there is no significant decline on the FHM dataset. We speculate that the HarM dataset's smaller scale, focused mainly on COVID-19-related hateful factors, results in a more straightforward sample feature structure. The performance improvement on the HarM dataset with Euclidean distance-based feature separation between hateful and non-hateful categories suggests that such an approach works well in this context. In contrast, the FHM dataset, characterized by higher quality and diverse hateful factors, presents a complex feature structure, making category-based feature processing less effective. Moreover, considering the ablation results for w/o $\mathcal{L}_1$ and w/o $\mathcal{L}_2$, the model's performance experiences a slight decline on both datasets. However, compared to the complete elimination of contrastive learning in w/o PCL, these modifications contribute to some improvement. This suggests that the model can learn different feature information through two distinct contrastive learning mechanisms, thereby enhancing overall classification performance.

Setting	**HarM**		**FHM**	
	Acc	Marco-F_1	Acc	Marco-F_1
Pen	86.30	85.27	74.04	73.55
w/o s_4	85.76	84.87	73.98	73.50
w/o s_2, s_3	86.02	85.04	73.90	73.44
w/o s_2, s_3, s_4	84.75	83.98	73.44	73.05

Table 4: Fine-grained Ablation Study on PMP Module.

To provide a more nuanced evaluation of the PMP module's effectiveness, we conducted three forms of reduction based on the components within the PMP module. The fine-grained ablation results of the PMP module, as shown in Table 4, indicate that varying degrees of reduction in the components lead to different extents of performance decline, particularly evident in the HarM dataset. Notably, due to the fused information feature $\hat{I}^{mix}$ carrying the most perceptual information, the performance drop is more significant in the HarM dataset for w/o s_4 compared to w/o s_2, s_3. Conversely, w/o s_2, s_3, s_4, equivalent to incorporating only inference instance information t^{infer} in the scoring process, results in the most significant information loss and hence the poorest model performance.

Simultaneously, we observed that the impact of different degrees of score reduction on the FHM dataset is relatively minor compared to the HarM dataset. We speculate that this is because the FHM dataset encompasses various types of hateful memes. In the process of the model judging the hateful category for inference instances, the choice of demonstrations is crucial. For instance, when classifying memes related to racial discrimination, selecting a hateful demonstration about attacks on sexual orientation might lead to ineffective information extraction, causing a decline in accuracy. In contrast, in the HarM dataset, where only COVID-19-related content exists, inaccurate demonstrations are not a concern. Therefore, to achieve improved performance on the FHM dataset, adding more diverse information is crucial.

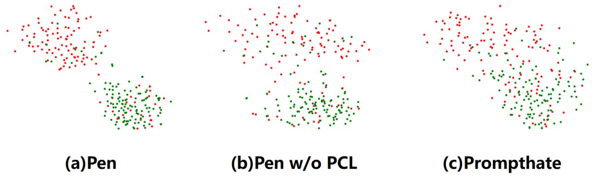

Figure 4: Visualization of Sample Features Learned by Our Pen (a), Pen without Prompt-aware Contrastive Learning (b), and Prompthate (c). Red=Hateful, green=Non-Hateful.

4.5 Visualization

To qualitatively demonstrate how our proposed prompt-based contrastive learning method enhances the quality of sample features, we present T-SNE [van der Maaten and Hinton, 2008] visualizations of sample features learned by our Pen and Prompthate on the HarM test set. The results are depicted in Figure 4. Figure 4(a) effectively illustrates how our Pen framework can cluster sample features belonging to the same label category and separate features of different labels. This contrast is evident when compared to Pen without the PCL module (b) and the Prompthate method relying solely on prompts (c). These comparisons provide evidence of the efficacy of our PCL method in improving the distribution of model-learned sample features. Additionally, the visualizations in Figure 4(b) and Figure 4(c) show that our Pen framework, purely through the PMP module, learns a more distinct separation trend between sample features of different labels compared to the Prompthate method using only manual prompts. This indirectly supports the notion that the PMP method can derive more robust inductive information from the training data, thereby enhancing the performance of hateful meme classification.

5 Conclusion and Future work

In this paper, we introduce a prompt-enhanced framework named Pen for hateful meme classification. We extend the concept of prompt methods into the feature space, enhancing the relationship between inference instances and demonstrations in a multi-view perception manner. Additionally, we leverage prompt-aware contrastive learning to enhance the distribution quality of sample features, effectively improving the model's classification performance on hateful memes. Through comprehensive experiments on two public datasets, we demonstrate the Pen framework's ability to significantly enhance the effectiveness of prompt methods, showcasing outstanding generalization and classification accuracy in hateful meme classification tasks. Furthermore, we intend to extend the framework to few-shot tasks, enhancing the accuracy of prompt methods in classifying low-resource text-only classification tasks.

Acknowledgments

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515011370), the National Natural Science Foundation of China (Grant No. 32371114), the Characteristic Innovation Projects of Guangdong Colleges and Universities (Grant No. 2018KTSCX049), and the Natural Science Foundation of Guangdong Province (Grant No. 2021A1515012290).

References

[Blaier *et al.*, 2021] Efrat Blaier, Itzik Malkiel, and Lior Wolf. Caption enriched samples for improving hateful memes detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9350–9358. Association for Computational Linguistics, 2021.

[Cao *et al.*, 2022] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 321–332. Association for Computational Linguistics, 2022.

[Cao *et al.*, 2023] Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. Procap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 5244–5252. ACM, 2023.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019*, pages 4171–4186. Association for Computational Linguistics, 2019.

[Fang *et al.*, 2022] Huaicheng Fang, Fuqing Zhu, Jizhong Han, and Songlin Hu. Multimodal hateful memes detection via image caption supervision. In *Proceedings of 2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles, SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta 2022, Haikou, China, December 15-18, 2022*, pages 1530–1537. IEEE, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[He *et al.*, 2023] Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. *CoRR*, abs/2308.05596, 2023.

[Jian *et al.*, 2022] Yiren Jian, Chongyang Gao, and Soroush Vosoughi. Contrastive learning for prompt-based few-shot language learners. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5577–5587. Association for Computational Linguistics, 2022.

[Kärkkäinen and Joo, 2021] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1547–1557. IEEE, 2021.

[Kiela *et al.*, 2019] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. In *Proceedings of the 2019 Conference on Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019*, 2019.

[Kiela *et al.*, 2020] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the 2020 Conference on Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Lee *et al.*, 2021] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *Proceedings of the MM '21: ACM Multimedia Conference, Virtual Event, China, October 20-24, 2021*, pages 5138–5147. ACM, 2021.

[Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning, ICML 2023, Honolulu, Hawaii, USA, 23-29 July 2023*, volume 202, pages 19730–19742. PMLR, 2023.

[Liang *et al.*, 2022] Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. Jointcl: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 81–91. Association for Computational Linguistics, 2022.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[Liu *et al.*, 2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35, 2023.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 2019 Conference on Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, CanadaDecember, 8-14, 2019*, pages 13–23, 2019.

[Mokady *et al.*, 2021] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734, 2021.

[Muennighoff, 2020] Niklas Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *CoRR*, abs/2012.07788, 2020.

[Piccoli *et al.*, 2024] Valentina Piccoli, Andrea Carnaghi, Michele Grassi, Marta Stragà, and Mauro Bianchi. Corrigendum to cyberbullying through the lens of social influence: Predicting cyberbullying perpetration from perceived peer-norm, cyberspace regulations and ingroup processes [comput. hum. behav. 102c (2020) 260-273. *Comput. Hum. Behav.*, 151:108040, 2024.

[Pramanick *et al.*, 2021a] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021, pages 2783–2796. Association for Computational Linguistics, 2021.

[Pramanick *et al.*, 2021b] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4439–4455. Association for Computational Linguistics, 2021.

[Qu *et al.*, 2023] Yiting Qu, Xinlei He, Shannon Pierson, Michael Backes, Yang Zhang, and Savvas Zannettou. On the evolution of (hateful) memes by means of multimodal contrastive learning. In *Proceedings of the 44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pages 293–310. IEEE, 2023.

[Ren *et al.*, 2017] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017.

[Suryawanshi *et al.*, 2020] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020*, pages 32–41. European Language Resources Association (ELRA), 2020.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[Velioglu and Rose, 2020] Riza Velioglu and Jewgeni Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *CoRR*, abs/2012.12975, 2020.

[Zhang *et al.*, 2022] Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Y. S. Lam. New intent discovery with pre-training and contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 256–269. Association for Computational Linguistics, 2022.

[Zhou *et al.*, 2021] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. Multimodal learning for hateful memes detection. In *Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops, Shenzhen, China, July 5-9, 2021*, pages 1–6. IEEE, 2021.

[Zhu, 2020] Ron Zhu. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *CoRR*, abs/2012.08290, 2020.