Indian Institute of Technology Bhilai

DS(L)-605: Deep Learning for Low Resource NLP

Instructor: Dr.Rajesh Kumar Mundotiya

MultiModal Hateful Meme Classification

Group 1 - Bot Army

Kammela Aditya Vardhan

Reddy Rishitha Reddy

G.S.V Raghavendra

Abstract

In the dynamic landscape of content moderation and social media analysis, the emergence of memes as a unique form of communication has sparked a need for innovative approaches. This project introduces "HateDetector," a multimodal prompt-based model designed for unsupervised topic modeling of memes. HateDetector achieves an impressive AUC score of 82.38, outperforming state-of-the-art baselines in classifying hateful memes. Meanwhile, HateDetector utilizes multimodal prompts to seamlessly extract and cluster topics from memes, demonstrating superiority over existing baselines in real-world meme datasets. Qualitative analysis underscores the model's capacity to identify culturally relevant meme topics. These advancements contribute significantly to our understanding of multimodal content analysis, classification, and the nuanced themes within contemporary meme communication. [Disclaimer: This paper contains sensitive content that may be disturbing to some readers.]

Contents

1 Introduction

In the ever-evolving landscape of social media communication, internet memes have emerged as a prominent and pervasive form of expression. While often intended for humor or satire, a darker underbelly exists, with malicious users generating and disseminating hateful memes targeting individuals based on race, ethnicity, religion, and other characteristics. The virality of these memes poses a unique challenge, as they can be reposted and shared across various contexts, potentially amplifying their harmful impact. To address this issue, social media platforms, such as Facebook, have released large hateful meme datasets, challenging researchers to develop automated solutions for effective classification. However, the complex nature of hateful memes, incorporating both visual and textual elements, necessitates innovative approaches that comprehend and reason across multiple modalities, demanding contextual background knowledge for accurate classification.

In response to the formidable task of hateful meme classification, previous research has explored multimodal approaches, utilizing pre-trained visual language models fine-tuned for the classification task. However, these approaches have limitations, as understanding hateful memes often requires additional contextual background knowledge. This study bridges these gaps by introducing a novel framework called HatePrompt, a multimodal prompt-based approach that harnesses the implicit knowledge within Pre-trained Language Models (PLMs) to enhance hateful meme classification. By converting images into textual descriptions and designing specific prompts, HatePrompt adapts and leverages the vast knowledge embedded in PLMs. This paper contributes a comprehensive analysis, demonstrating the effectiveness of HatePrompt through extensive experiments on publicly available datasets, showcasing its superiority over state-of-the-art methods in the challenging task of hateful meme classification. Fine-grained analyses and case studies further illuminate the efficacy of the proposed prompts in accurately classifying hateful memes, marking a significant advancement in the field.

2 Problem Statement

The proliferation of hateful memes on social media platforms poses a critical challenge for content moderation and societal well-being. These memes, disguised under the veneer of humor or satire, harbor malicious intent, targeting individuals based on race, ethnicity, religion, and other characteristics. The dynamic and viral nature of memes complicates the task of combating their spread, necessitating automated solutions for accurate and timely classification. Existing multimodal approaches, relying on fine-tuned visual language models, fall short in capturing the nuanced contextual background knowledge essential for understanding and classifying hateful memes. As such, there is an urgent need for innovative frameworks that bridge this gap, leveraging the latent knowledge within Pre-trained Language Models (PLMs) to effectively discern and categorize hateful content in the multimodal landscape of internet memes. The problem at hand demands a solution capable of comprehending both textual and visual elements, fostering a safer online environment by addressing the challenges inherent in hateful meme classification.

3 Earlier Works

In the realm of hate speech detection on social media, research has predominantly concentrated on text-based content, with limited attention given to text-embedded images containing hate speech. While several studies, such as those by Alam et al. (2022) and Chhabra and Vishwakarma (2023), have made significant strides in text-based hate speech detection, fewer efforts have been dedicated to the classification of images with embedded text for hate speech (Gomez et al., 2020; Bhandari et al., 2023). Recent scholarly interest has notably surged toward identifying hate speech within memes or images with text, as demonstrated by studies like Ji et al. (2023), Hermida and Santos (2023), Karim et al. (2022), Yang et al. (2022, 2019a), and Perifanos and Goutsos (2021). These endeavors highlight the increasing recognition of the need to analyze and combat hate speech within the specific context of text-embedded images, acknowledging the unique challenges posed by this multimodal form of communication.

Simultaneously, the exploration of memes and multimodal textual-visual data has primarily centered on general social media platforms. However, dedicated datasets and research within specific contexts have been limited. Initiatives by Pramanick et al. (2021a, 2021b) and Naseem et al. (2023) have made strides in understanding harmful memes related to the COVID-19 pandemic, memes related to the US election, and memes critical of vaccines. These studies lay the groundwork for context-oriented investigations, recognizing the importance of studying multimodal textual-visual data within specific thematic contexts. The shared task introduced in this study serves as a call to action, urging the research community to delve deeper into the examination of hate speech within text-embedded images and fostering context-specific analyses in the evolving landscape of online communication.

4 Novelty

We have meticulously compiled a dataset comprising Telugu-scripted memes sourced from diverse social media platforms, including Facebook, Reddit, Sharechat, and Pinterest. This dataset encompasses both hateful memes, addressing sensitive subjects such as religion, caste, gender, history, and race, as well as non-hateful counterparts. Our dataset aims to provide a comprehensive representation of Telugu-scripted memes, facilitating a nuanced examination of hate speech within this linguistic context.

Additionally, we conducted a rigorous process of hyper-parameter tuning to optimize the performance of our hate speech classification model, particularly focusing on the robust RoBERTa-large model. The tuning involved fine-tuning key parameters such as batch size, learning rate, and epochs to enhance the model's discrimination capabilities between hateful and non-hateful memes. Furthermore, we explored the computational efficiency and performance trade-offs by evaluating lighter models, including SqueezeBert, and analyzing the corresponding computation-performance graph. This comprehensive analysis seeks to identify the most effective model for hate speech detection within memes, aligning with our commitment to advancing accurate and scalable content moderation solutions within this linguistic domain.

5 Experimental Architecture

The proposed HateDetector leverages the prompting technique for hateful meme classification, incorporating manual construction of prompts with predefined label words and templates. This architecture guides the RoBERTa PLM to make predictions based on positive and negative demonstrations, providing a novel approach to the challenging task of identifying hateful content in memes. The study emphasizes the importance of label word and template choices, offering insights into the effectiveness of the proposed method.This multimodal approach, incorporating both textual and visual information, enhances the model's ability to capture the nuanced features of memes. By combining information extracted from the text and image captions, PromptHate aims to improve the accuracy of hateful meme classification.

The process involves extracting text from memes, in-painting to remove text from images, generating captions using ClipCap, and integrating these captions into the PromptHate architecture. This multimodal approach allows the model to leverage both textual and visual cues for more robust and nuanced classification of hateful content in memes.

1. Extracting Text from Memes:

The first step in preparing the input for PromptHate involves extracting textual information from the memes. This is accomplished using open-source Python packages, specifically EasyOCR2. EasyOCR is a tool designed for Optical Character Recognition (OCR), which enables the extraction of text from images. This step is crucial for incorporating textual information present in memes, which may include captions, comments, or any other relevant textual content associated with the images.

2. In-painting with MMEditing:

After extracting the text from memes, the next step involves in-painting the images

using MMEditing3. In-painting is a process of reconstructing missing or altered parts of an image. In this context, it is employed to remove the text from the memes, ensuring that the subsequent analysis is based on the visual content alone. MMEditing is an open-source multimedia editing toolbox that provides various functionalities, including image in-painting.

3. Pre-trained Image Captioning Model (ClipCap):

After extracting textual information and in-painting to eliminate text from images, a pre-trained image captioning model, ClipCap, is employed. ClipCap is tailored to produce high-quality captions for low-resolution web images. Subsequently, ClipCap generates captions that describe dominant objects or events within meme images. These captions serve as textual representations of visual content, essential for converting meme information into a format compatible with pre-trained language models like RoBERTa. This facilitates a multimodal approach in the analysis of memes, blending visual and textual information for a comprehensive understanding in the context of hateful meme classification.

4. Integration with PromptHate:

The generated captions from ClipCap are then integrated into the PromptHate architecture. These captions, which encapsulate the semantics of the memes, become part of the multimodal input fed into the PromptHate model. The combined textual information from meme text and image captions, along with the constructed prompts, guides the PLM in making predictions regarding the hateful or non-hateful nature of the memes.

The core idea behind PromptHate is to construct a prompt that guides the PLM (RoBERTa) in making predictions about whether a given meme is hateful or non-hateful. The prompt contains three main elements:

a) Positive Demonstration : A representation of a normal meme, labeled as non-hateful.

Template: “the meme is good.”.

b) Negative Demonstration : A representation of a hateful meme.

Template: “the meme is bad.”.

c) Inference Instance : The meme to be predicted.

Template: “the meme is [MASK].”.

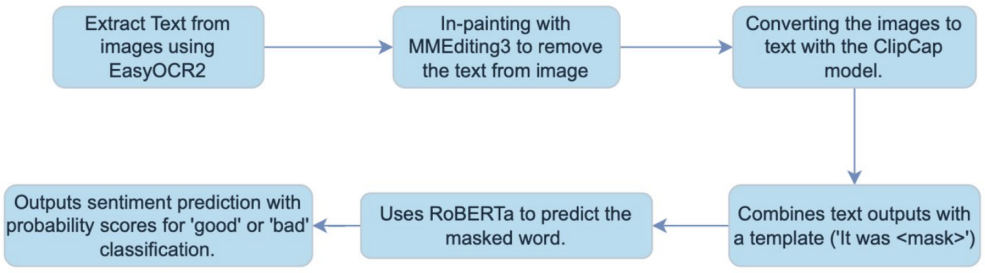

Figure 1: FrameWork

5. Prompting Hateful Meme

To guide the pre-trained language model (PLM) in inferring the label word, we provided positive and negative demonstrations to the PLM. The positive demonstration S_{pos} is generated as: $S_{pos}[SEP]S_{pos}[SEP]T(W_{pos})$, where S_{pos} and S_{pos} are meme texts and image descriptions respectively, $[SEP]$ is the separation token in the language model L, and $T(W_{pos})$ generates the positive label word W_{pos} into a sentence (e.g., “this is good”). A similar approach is used for the generation of negative demonstration S_{neg} and inference instance S_{infer} by replacing W_{pos} with W_{neg} and $[MASK]$, respectively. Inspired by Gao et al. (2021), we concatenate the demonstrations with the inference instance:

$$S = \texttt{[START]}\, S_{infer}\, \texttt{[SEP]}\, S_{pos}\, \texttt{[SEP]}\, S_{neg}\, \texttt{[END]} \tag{1}$$

where, S serves as the prompt fed into L, and $[START]$ and $[END]$ are start and end

tokens in L.

6. Model Training and Prediction

$$y_0 = P([\text{MASK}] = W_{pos}|S), (2) \quad y_1 = P([\text{MASK}] = W_{neg}|S). \tag{2}$$

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

$$\text{Loss} = y \log(\hat{y}) + y \log(1 - \hat{y}), \tag{3}$$

and the loss will be used for updating parameters θ in 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 $y_0 > y_1$, the meme will be predicted as hateful, otherwise, non-hateful.

6 Experimental Settings

We used PyTorch on an NVIDIA Tesla T4 GPU with 15 GB dedicated memory and CUDA 12.0 to train all the models. For pre-trained models such as Roberta and SqueezeBert, we used the transformers package(version 4.35.0) from HuggingFace. The learning rates were assigned empirically. For Bert-based models such as SqueezeBert, the learning rate was set to $1.3 * 10^{-5}$. For Roberta large-based models such as HateDetector, we tested learning rates ranging from 10^{-5} and $1.5 * 10^{-4}$ and reported the best performing one. We used AdamW as the optimizer for all the models. The mini-batch size was set at 16 during training. It takes one GPU eight minutes to train and validate HateDetector per epoch. For training HateDetector on the FHM dataset, it takes up to 15 GB of GPU RAM.

In our experiments, we rigorously evaluated our method using two pre-trained language models: RoBERTa, known for robust bidirectional context understanding, and SqueezeBert, prioritizing model compression. This comparison aimed to assess the method's adaptability and effectiveness across diverse model paradigms, revealing insights into generalization capabilities and nuanced performance variations.

We utilized the publicly available dataset: the Facebook Hateful Meme dataset (FHM). To align with the binary classification of hateful and non-hateful memes, we grouped very harmful and partially harmful memes as hateful.

For evaluation, we adopted common metrics in hateful meme classification studies: Area Under the Receiver Operating Characteristic curve (AUROC) and Accuracy (Acc). To ensure reliable results, we averaged performance over ten random seeds. We compared PromptHate against state-of-the-art models like SqueezeBert. Additionally, we benchmarked PromptHate against fine-tuning RoBERTa (FT-RoBERTa), concatenating meme text and image descriptions for classification. The diverse set of benchmarks allows for a comprehensive evaluation of PromptHate's performance in comparison to existing models across various modalities and architectures.

7 Results and Analysis

Table 1 presents the outcomes of our experiments with the PromptHate framework. It includes the impact of hyperparameter adjustments, such as the learning rate modification, revealing insights into the model's sensitivity and convergence dynamics. Additionally, we explore the tradeoff between model complexity and efficiency by contrasting the performance of RoBERTa-large and the lightweight SqueezeBERT in the context of our multimodal hate meme classification.

Model	**Learning Rate**	**Batch Size**	**AUC**	**Accuracy**
Roberta-Large	1.30E-05	16	82.38	73.6
	1.30E-04	16	53.23	50.6
SqueezeBert	1.30E-05	16	74.47	66.8

Table 1: Results

In our experimental settings, we initially utilized a default learning rate of 1.30E-05 for training the PromptHate framework with the RoBERTa model, achieving a noteworthy accuracy of 73.6% and an AUC of 82.38. Subsequently, we conducted a variation experiment by increasing the learning rate to 1.30E-04. Surprisingly, this adjustment led to a substantial drop in performance, with the accuracy plummeting to 50.6% and the AUC decreasing to 52.23.

The unexpected decline in performance with the increased learning rate indicates that the model's ability to generalize and converge effectively was adversely affected. It suggests that the initial learning rate of 1.30E-05 was conducive to the model's training dynamics, facilitating better convergence towards an optimal solution. The decline can be attributed to issues such as overshooting optimal weights, failure to converge, and poor generalization, stemming from the abrupt change in learning rate. This underscores the critical importance of meticulous hyperparameter tuning to maintain stability and achieve optimal performance

in the training process.

When implementing SqueezeBERT, a lightweight alternative to RoBERTa-large, we observed a decrease in accuracy to 74.47% and AUC to 66.8. SqueezeBERT, known for its efficiency with fewer parameters, offers faster training times compared to the more complex RoBERTa-large. However, this speed advantage comes at the cost of reduced accuracy and AUC, indicating a tradeoff between model size and performance. SqueezeBERT's limitations in capturing intricate and complex features essential for our multimodal classification task underscore the necessity of balancing computational efficiency with the model's capacity to handle the inherent complexity of the data.

In conclusion, the HateDetector's experimental architecture strategically capitalizes on RoBERTa's advanced bidirectional context understanding. By seamlessly integrating textual information derived from both meme text and image captions, the model adopts a multimodal approach, thereby enriching its capacity for hateful meme classification. The bidirectional context understanding of RoBERTa empowers HateDetector to capture nuanced relationships within meme content, allowing it to discern subtle contextual cues indicative of hatefulness. This holistic fusion of diverse data sources significantly amplifies the model's discriminatory capabilities, culminating in a sophisticated and accurate classification framework.

8 References

1. Prompting for Multimodal Hateful Meme Classification Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, Jing Jiang.

2. Identifying Creative Harmful Memes via Prompt based Approach Junhui Ji, Wei Ren, Usman Naseem.

3. PromptMTopic: Unsupervised Multimodal Topic Modeling of Memes using Large Language Models Nirmalendu Prakash, Han Wang, Nguyen Khoi Hoang, Ming Shan Hee, Roy Ka-Wei Lee