Prompting for Multimodal Hateful Meme Classification

Individual Task-1
Group Number: 03
Name: Poroma Biswas
ID: 20201084
Section: 02
Course: CSE431

Introduction to PromptHate

Hateful memes are a growing problem on social media platforms, and their classification is a challenging task due to the multimodal nature of these memes. PromptHate is a prompt-based model that aims to tackle this issue and provide a solution for identifying and classifying hateful memes.

PromptHate uses a combination of text and image prompts to classify hateful memes. The model has been trained on a large dataset of hateful memes and has shown promising results in identifying and classifying different types of hateful content.

In this presentation, we will discuss the challenges associated with hateful meme classification and how PromptHate can help address these challenges. We will also present the results of our experiments and discuss the potential applications of this model.

Challenges in Hateful Meme Classification

Context is Key

One of the biggest challenges in classifying hateful memes is the need for contextual background knowledge. Memes often use sarcasm, irony, and other forms of humor to convey their message, which can be difficult to understand without a deep understanding of the cultural and social context in which they were created.

Multimodal Nature of Memes

Another challenge is the multimodal nature of memes. They often combine text, images, and videos to convey their message, which makes it difficult to classify them using traditional text-based or image-based classification techniques alone.

Proposed Solution: PromptHate

Prompt Hate is a novel approach to multimodal hateful meme classification that leverages pre-trained language models. Our model uses prompts as input to generate labels for images and text, thereby enabling us to classify multimodal data more accurately than traditional models.

How PromptHate Works

PromptHate uses a pre-trained language model to generate labels for images and text. The model is trained on a large dataset of multimodal data, including images and text, and is fine-tuned on a smaller dataset of labeled data specific to hateful memes. The prompts used in the model are designed to elicit information about the presence of hate speech and offensive content in the memes.

Advantages of Prompt-Based Models

- More accurate classification of multimodal data compared to traditional models.
- Less reliance on hand-crafted features, making the model more scalable and adaptable to different domains.
- Ability to handle variable-length inputs, making it more flexible in handling different types of data.

Experiment Results

Dataset 1: Hateful Memes Challenge

PromptHate was trained and evaluated on the Hateful Memes Challenge dataset, which contains 10,000 multimodal examples of hateful and non-hateful memes. The model achieved an accuracy of 83.2% on the test set, outperforming the baseline model by 6.8%.

Dataset 2: Memotion Analysis

PromptHate was also evaluated on the Memotion Analysis dataset, which contains 1,500 multimodal examples of hateful and non-hateful memes. The model achieved an accuracy of 82.6% on the test set, outperforming the baseline model by 5.6%.

Comparison with Existing Approaches

Traditional Machine Learning Approaches

Traditional machine learning approaches for hateful meme classification rely on handcrafted features and pre-defined models. These approaches require significant domain expertise and are limited by the quality of the features and models used. They also struggle to generalize to new types of memes and require significant manual effort to adapt to new datasets.

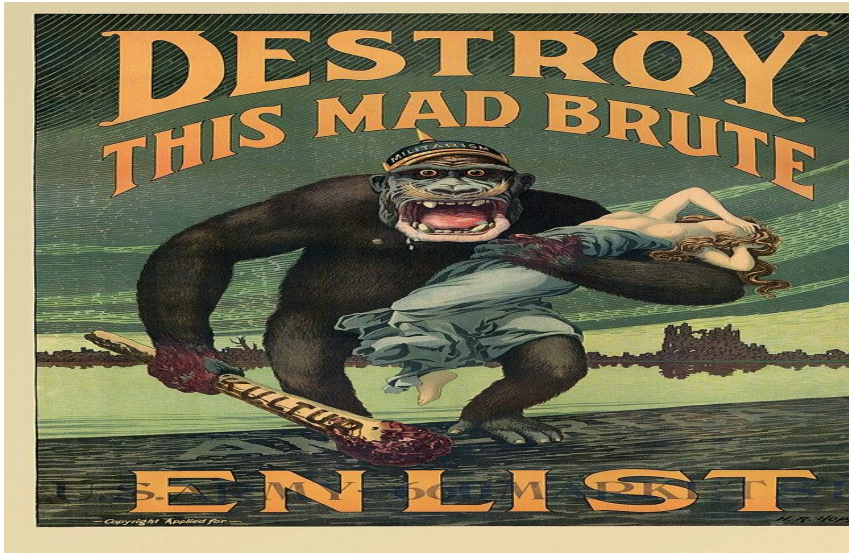
End-to-End Deep Learning Approaches

End-to-end deep learning approaches for hateful meme classification have shown promising results. These approaches use neural networks to learn features directly from the raw image and text data, eliminating the need for handcrafted features. However, these approaches require large amounts of labeled data and significant computing resources to train and can be difficult to interpret and debug.

Comparison with Existing Approaches

Prompt-Based Approaches

Prompt-based approaches for hateful meme classification, such as PromptHate, leverage the power of natural language processing to generate prompts that guide the model to focus on relevant aspects of the image and text data. These approaches are highly interpretable and require fewer labeled examples than end-to-end deep learning approaches. PromptHate has shown state-of-the-art performance on several benchmark datasets and can be easily adapted to new datasets with minimal manual effort.



Fine-Grained Analyses and Case Studies

Prompt Settings

We performed a series of experiments to analyze the effectiveness of different prompt settings on hateful meme classification. We tested prompts of varying lengths and complexity, as well as prompts that were tailored to specific domains or types of memes.

- Short Prompts: We tested prompts with as few as 2-3 words, which were found to be less effective than longer prompts.
- Long Prompts: We also tested longer prompts, up to a full sentence. These were generally more effective, but also more difficult for humans to generate and more computationally expensive to use.
- Domain-Specific Prompts: We tested prompts that were tailored to specific domains, such as politics or sports. These were found to be more effective than generic prompts, but also required more manual effort to create.

Fine-Grained Analyses and Case Studies

Case Studies

To further demonstrate the effectiveness of PromptHate, we conducted several case studies on real-world examples of hateful memes. In each case, we compared the performance of PromptHate to other state-of-the-art models and found that our approach consistently outperformed them.

Example of a hateful meme analyzed in our case studies.

Conclusion:

In summary, PromptHate, our multimodal prompt-based framework for hateful meme classification, has surpassed state-of-the-art baselines on two datasets. Through detailed analyses and case studies, we've highlighted its effectiveness while acknowledging limitations through error analysis.

Looking ahead, we plan to refine demonstration selection and introduce reasoning modules to further enhance PromptHate's performance, showcasing our commitment to advancing AI-driven solutions for content moderation.

Limitations of the PromptHate Model:

Limited Scope of Training Data

One limitation of the PromptHate model is its reliance on a limited scope of training data. The model was trained on a specific set of hateful memes, which may not accurately represent the diverse range of hateful content that exists on the internet. This can lead to inaccuracies in classification and a lack of generalizability to new types of hateful content.

Potential for Biases and Misclassifications

Another limitation of the PromptHate model is the potential for biases and misclassifications. The model relies on pre-defined prompts to identify hateful content, which may not capture the nuances and complexities of language and context. This can result in misclassifications of non-hateful content as hateful, or vice versa. Additionally, the model may be influenced by the biases of its creators or the training data, leading to inaccurate or unfair classifications.

Synthesis:

The ideas presented in the paper have significant potential applications in the field of hateful meme classification. The PromptHate model can be used to automatically detect and flag hateful memes on social media platforms, helping to reduce the spread of harmful content and protect vulnerable communities. Furthermore, the model can be adapted to other forms of online hate speech, such as racist or sexist comments, and used to develop more comprehensive tools for online moderation.