Report : Analysing Hateful Memes

Alex Thuruthel

15th February , 2024

Introduction

This task revolved around analysing memes , a seemingly trivial but very complex task. Analysing memes requires understanding the context with which it was made. The implementation mainly uses computer vision techniques but at some points certain NLP techniques were used as they were the most effective at doing the required task

Task A : Object Detection

Object Detection is a core part of this task as the objects within the image shape the analysis of it to a great extent. **YOLOv8** was the model chosen after extensive testing. Since the task revolved around hateful memes, race, gender, and emotion add very important context that would be lost with more general context detection. After extensive research, a model was found called **Deepface** that was designed for this task. The results of the model are later used in the classification system.
Brief overview of the models:

- **YOLOv8** YOLOv8 is a new state-of-the-art computer vision model built by Ultralytics. It is the pretrained model used for object detection.
- **Deepface**: Deepface is a hybrid face recognition framework wrapping state-of-the-art models: VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace, Dlib, and SFace. The Facial Attribute Analysis Feature from the model was used in the task.

The above models were chosen after extensive comparison with other available models, some models that were considered for the task but were not used were:

- **Faster R-CNN**: Limitations of the model included a lack of extensive labels more specific for the task.
- **Grounding DINO**: This is a zero-shot learning model that overcomes the challenge on object classes, however, after empirical testing with the dataset, accuracy was not satisfactory.

Task B : Caption Impact Analysis

Captions are an integral part of any meme , however due to their positioning in the foreground of the image , they might block out important sections of the object. Humans will be able to understand what the object that is being obstructed via context , but any object detection model will not be able to do so. Therefore, studying the impact of captions on object detection is essential.

The method used to eliminate captions from the image was a basic inpainting model. Inpainting to remove text from images involves process where first the pixels associated with the text to be removed are identified . Then, using deep learning techniques, the system analyzes the surrounding pixels to understand the texture, color, and patterns of the background. Based on this understanding, it generates new pixels that best match the surrounding area to fill in the gap that will be left by the absence of the text.

After inpainting was applied to images of the dataset , to assess the impact of the caption on object detection , the same object detection was run on the images without the text . All the tests were run on the dev images of the dataset , which consisted on 500 randomly chosen images.

Performance of object Detection per image : It was found that of the 500 images on which the test was run on , exactly 100 showed better results through the object detection model after inpainting. **Peformance of Object Detection per object** : Due to difference in scale of counts of persons and other objects , person count was plotted differently.

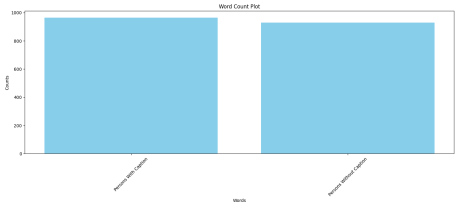

Figure 1: Comaparison of detection of people

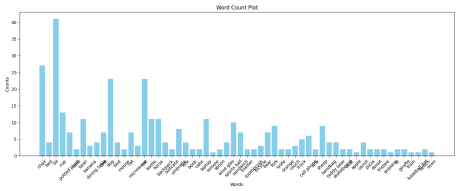

Figure 2: Detection with Captions

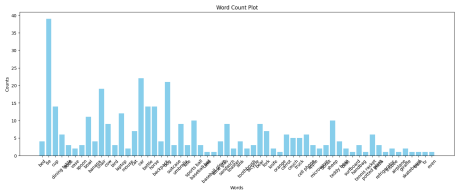

Figure 3: Detection without Captions

Task C : Classification System

The classification system designed for the purposes of this task determines whether an images is a meme or not. It does this by using the following : a caption generating model , a race and gender detection model and an object detection model , **BERT** . They are used for the following purposes:

YOLOv8 - Object Detection : Used to construct the prompt for the caption model

Deepface - Age/Gender Detection : Used to construct the prompt for the caption model

PromptCap - Caption generating model : The image (after inpainting if applicable) along with the prompts constructed in the above step will be given as input and it will output a caption for the image.

BERT - Sentence similarity : Bert is used to find the similarity between the caption generated by the captions generating model and text in the image. For the meme dataset the text is given , but OCR was implemented in a separate file if needed for other datasets. It then checks the cosine similarity of both the sentences. The lower the similarity , the more likely it is to be a meme , as if the similarity was high the text would just be a caption for the image and have the image would have no sarcastic meaning.

The object detection and race/gender detection were deemed necessary as they add a lot of the context to the meme . With the addition of their results to the prompt of the caption generating model a lot of information will be lost, especially the aspects that make the meme hateful.

Task D : Bonus Task

In this task , the task was to determine whether a given meme was hateful or not , effectively reducing this multimodal problem to a unimodal task. Even though the image contributes significantly to the meme , results from this task show that is possilble to determine the hatefulness of a meme solely from the text. BERT was used to run sentiment analysis on the text from the

image. It classifies the text into positive , neutral and negative by assigning it a number out of 5 , with 1 being the most negative and 5 being the most positive. Now to classify the text as hateful , a cutoff had to be determined to differentiate between hateful and not. Here the results of the classifier:

- With labels 1,2,3 being categorized as hate:
  - 50.4% accuracy on the training set.
  - 51.2% accuracy on the development set.
- With labels 1,2 being categorized as hate:
  - 54.77% accuracy on the training set.
  - 52% accuracy on the development set.
- With only label 1 being categorized as hate:
  - 56.35% accuracy on the training set.
  - 52.6% accuracy on the development set.

Clearly in this dataset , when only text is considered , only the most negative statements are considered hateful . Possible reason reasons for the low accuracy in some of the results is that the images contribute significantly to it be hateful. Thus even though there is a clear indication that text can be used to classify memes as hateful or not , the need for multimodal classification where the images are factored in is necessary

Limitations and Possible Improvements

Task E : Paper Reading Task

Summary

This paper introduced **MEMEX**, a task designed to extract the context behind memes using a combination of a meme and a related document. Due to a lack of datasets for this task , the researchers developed the **MCC** (Meme Context Corpus), a new, manually-annotated multimodal dataset comprising of memes on a preselected set of topics and a related document which

contained the context to explain the meme. They then introduced **MIME** (MultImodal Meme Explainer), a multimodal neural framework, which leverages a knowledge-enriched representation of memes and a layered approach to understand the semantic relationships between the meme and its context. **MIME** distinguishes itself by outperforming both unimodal and multimodal benchmark systems, results for which are displayed in the paper. This marks a substantial advancement in understanding and interpreting the nuanced communication conveyed through memes.

Strengths

Here are some of the strengths of the paper , this is not an exhaustive list.

- Having a context document from which the context of the meme is extracted , ensures that the context for the meme is extracted from an unbiased and verified source.
- Comparisons with popular Baseline models , gives the reader a clear and concise understanding of the accuracy of the model
- A detailed explanation of the methodology of implementation and data collection helps the reader gain a clearer understanding of the model and the results generated.

Weaknesses

Here are some of the Weaknesses of the paper:

- The selection of memes is restricted to very specific topics like History , politics (specifically US politics) and geo-politics via Google images and Reddit. Although they represent topics where a bias is likely and hence the need for **MemeX**, this results produced for this dataset will be inaccurate for more general memes as availability of content to derive context from and understanding of the content by the annotators might be limited.
- The content document is taken from a single Wikipedia page. A potential issue with this approach is that for more complex memes ,the ideal content document would involve information from multiple pages

, which the current approach fails to achieve, potentially losing very important context from the meme.

- When context for a particular meme is not found on Wikipedia , a google search is conducted. This can result is biased information. Heavily biased information might be flagged during the annotation process , but even subtle bias will negative effect the accuracy of **MemeX**. This issue is exacerbated due to the selection of memes for the dataset . Topics like History and Politics were preferred, these will often have heavily biased information online which will affect the integrity of the data.

- The Platform used by annotators required them to enter a specific text box which contained the content that was deemed as relevant content. This limits them to one block of text and renders them unable to add multiple lines from different sections as they deem necessary.

Improvements

- The platform for annotation must be redesigned to account for multiple context documents as well as multiple excerpt to given as input by the annotators

- The memes must be collected on a wider and more divers set of topics. The data also should be collected from multiple different different sites to prevent bias of users and the content moderation policy of the site to seep into the dataset.

[5] [1] [3] [2] [4]

References

[1] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.

[2] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.

[3] 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. 2020.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[5] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4. IEEE, 2021.