Term Project Proposal

Fall 2021

Group 3

Group Members:

Shreya Bhootda, Casey Copeland, Ankita Kundra, Yanqi (Eden) Liang,

Problem Statement:

With the prevalence of social media in our lives, it is increasingly important that applications provide a safe and reliable environment for users. Facebook hosted a Hateful Memes data challenge in order to develop multimodal machine learning models to detect hate in memes. Hate speech is content on social media that spreads hate via images and text. Hateful memes can be difficult to spot due to their multimodal nature. One modality of the meme, the image, may appear non-hateful while the other modality, the text, is hateful. This means the machine learning model must classify the memes holistically, as humans do, taking into account the multimodal nature and nuances of humor, such as sarcasm and changing meanings. The goal is to achieve the highest AUC and ROC score when classifying a test set of memes. Accurately and efficiently detecting hateful content is vital for Facebook as they receive criticism from all angles in regards to the safety and accuracy of their platform.

Data Available:

The hateful memes competition provides an ongoing dataset for future development and research purposes. This dataset is built by the Facebook AI team, and this category of cases are referred to as benign confounders. It includes an image directory (hateful memes) and multiple jsonl files containing the id, img (reference to actual image), and the string representation of the raw text embedded in the meme.

The dataset can be downloaded here: https://hatefulmemeschallenge.com/#about

Because it is a completed competition, the 1st - 5th place winners have repos in Github with their publications, summaries, and code. We plan on using these model examples to help us create our approach to developing a multimodal model. We want to put an emphasis on hateful content on the internet, specifically Facebook, and how these accurate models will create a safer social media space. We hope to add to the published models via accuracy, additional insights, or other uses. Using machine learning, Facebook will be able to detect hate speech more quickly and improve their platform.

Possible Approaches:

Planned steps we will take to solve the problem:

1. Data preprocessing and feature extraction

The first steps will be downloading the Facebook hateful meme database and pretrain model provided by Facebook AI Research Scientist Dr. Douwe Kiela. To preprocess the data, the memes, we have looked into Detectron2, which is a pytorch based modular object detection library. This library built by Facebook's AI Research team could help us extract features from the images, which allow for a faster training process later while obtaining the relevant information. We could have feature extraction with different models to create a diverse set of attributes.

2. Model Approaches

There are a few approaches to modeling we have considered based on the winners' published work and research. The Facebook AI Research team has also provided baseline pre trained data for the sake of the competition :

- Running the image and text separately through different transformers, then aggregate them at the end to generate a final result. This could be easier to do but the result accuracy might not be as good.
- Running the image and text at the same time with a single transformer. Some notable single stream models we looked into include VisualBERT.
- Create multiple models (Single or Dual Stream) and utilize ensemble methods at the end to determine the final result.

References:

"Detectron2: A PyTorch-Based Modular Object Detection Library." Facebook AI, https://ai.facebook.com/blog/-detectron2-a-pytorch-based-modular-object-detection-library-/.

DrivenData. "Competition: Hateful Memes: Phase 1." DrivenData, https://www.drivendata.org/competitions/64/hateful-memes/.

DrivenDataOrg. "GitHub - Drivendataorg/Hateful-Memes." GitHub, https://github.com/drivendataorg/hateful-memes/.

HimariO. "GitHub - HimariO/HatefulMemesChallenge." GitHub, https://github.com/HimariO/HatefulMemesChallenge.

Muennighoff. "GitHub - Muennighoff/Vilio: Vilio: State-of-the-Art VL Models in PyTorch & PaddlePaddle." GitHub, https://github.com/Muennighoff/vilio.

Research, Facebook. “GitHub - Facebookresearch/Detectron2: Detectron2 Is FAIR’s next-Generation Platform for Object Detection, Segmentation and Other Visual Recognition Tasks.” GitHub, https://github.com/facebookresearch/detectron2.

---. “Mmf/Projects/Hateful_memes at Main · Facebookresearch/Mmf · GitHub.” GitHub, https://github.com/facebookresearch/mmf/tree/main/projects/hateful_memes.

“VisualBERT — Transformers 4.11.3 Documentation.” Hugging Face – The AI Community Building the Future., https://huggingface.co/transformers/model_doc/visual_bert.html.

Links to References:

1. https://www.drivendata.org/competitions/64/hateful-memes/
2. https://github.com/facebookresearch/mmf/tree/main/projects/hateful_memes
3. https://github.com/drivendataorg/hateful-memes/
4. https://huggingface.co/transformers/model_doc/visual_bert.html
5. https://arxiv.org/pdf/1908.03557.pdf
6. https://github.com/facebookresearch/detectron2
7. https://ai.facebook.com/blog/-detectron2-a-pytorch-based-modular-object-detection-library-/
8. https://github.com/Muennighoff/vilio