Hate Detection in Meme using Multimodal Sentiment Analysis

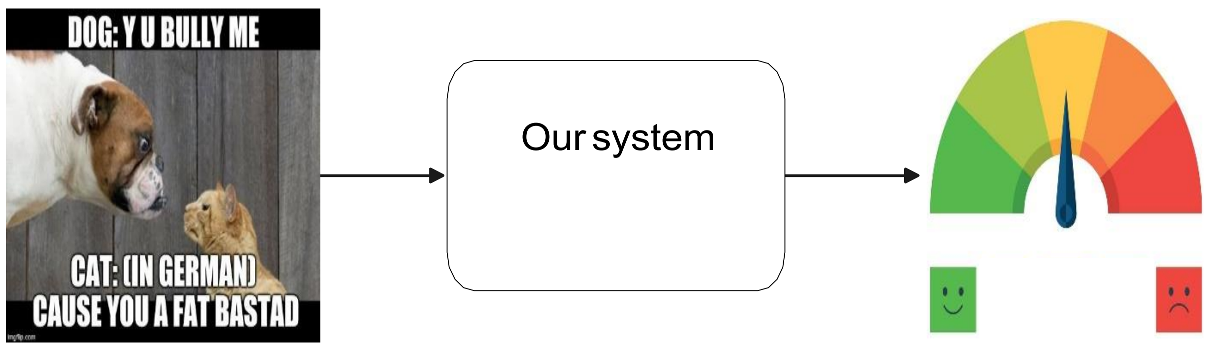
Team Name: Unplugged

Abid Hassan, Ankush Karunakara Shetty, Manikanta Chunduru Balaji, Suvimal Yashraj, Pooria Namyar

USC Viterbi School of Engineering

Motivation

Memes enable people to express their opinion freely through social media, which may create a hostile environment for users. So, it has become crucial to detect and filter such hate instances.



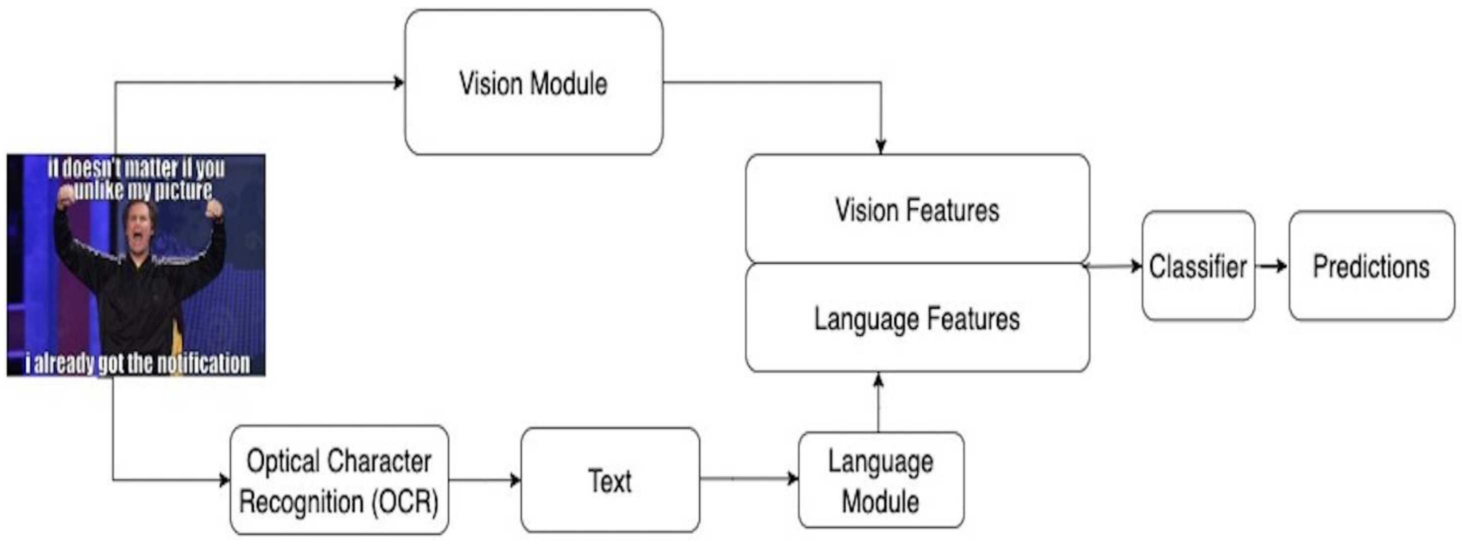
While hate speech detection has traditionally focused on language, we explore the impact of visual information for this task.

Overview

- The goal is to build a system to detect and filter offensive memes.
- Humans understand the content of a meme holistically, however, NLP models cannot.
- Multi-Modal: Image and text modalities were combined to get holistic features.
- Hate Meme Detection: Combined features were used to classify a meme.
- Hateful Memes Dataset

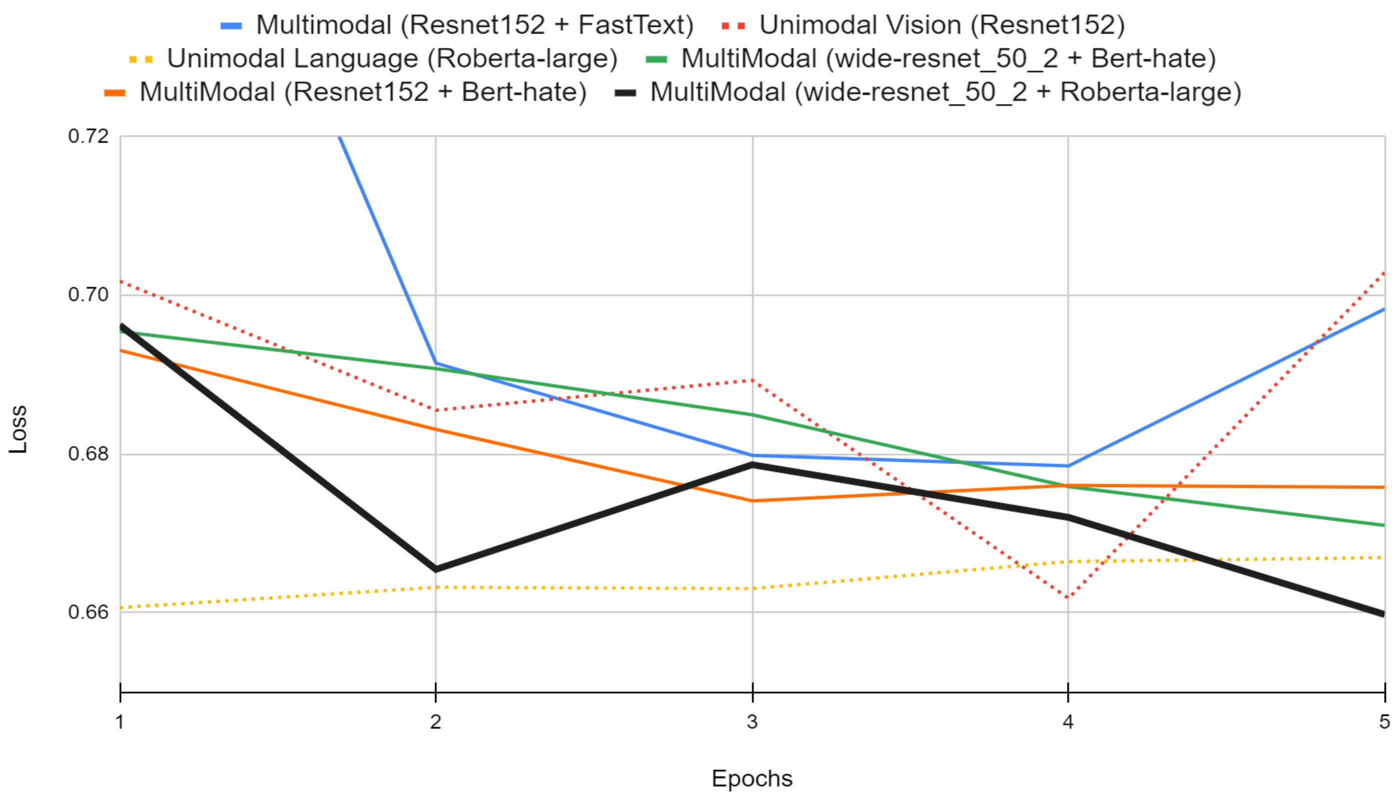
	😁	😡
Training	5481	3019
Validation	340	200
Testing	1250	750

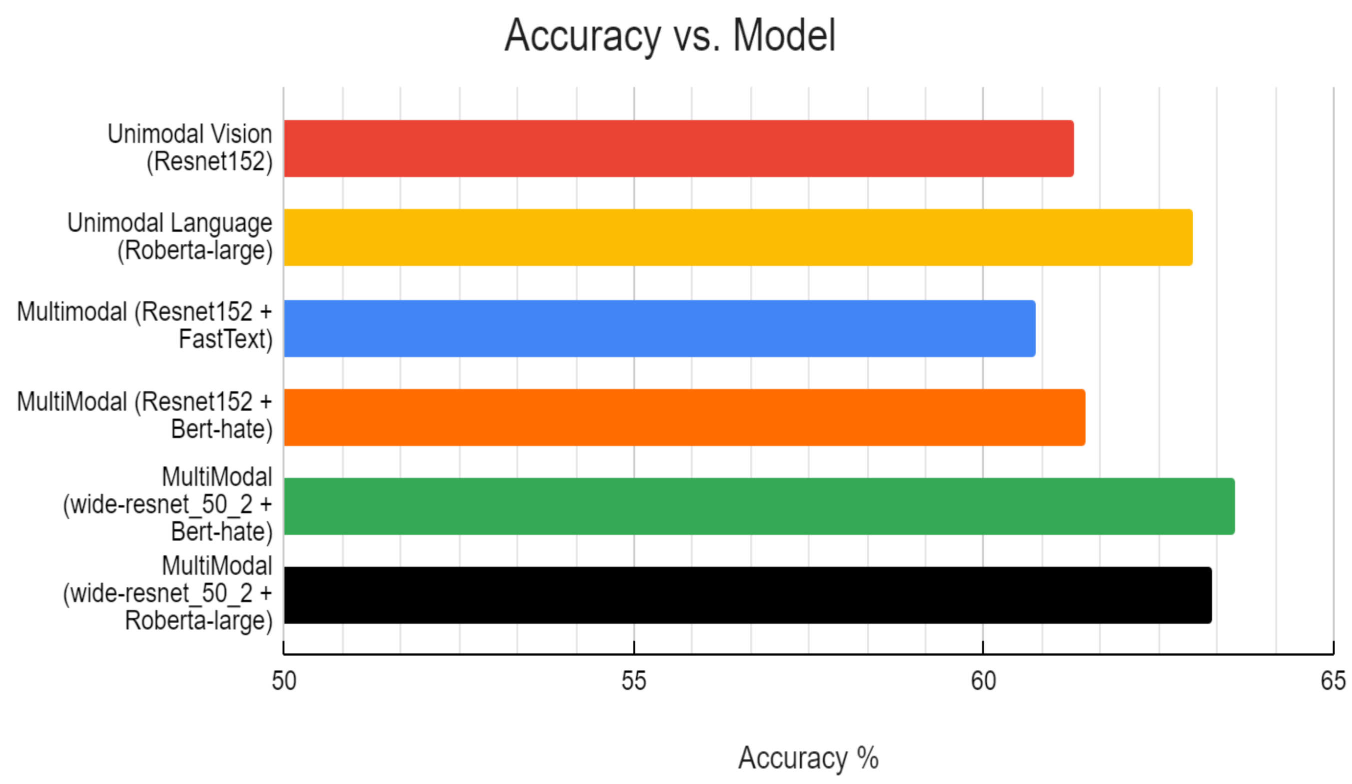
Methodology



Text Feature Extraction	**FastText model:** The outputs of the language module will come from a trainable Linear layer, as a way of fine-tuning the embedding representation during training.
Visual Feature Extraction	**ResNet152:** Pretrained on ResNet152 for input images of 224x224 Frozen weights. Input Feature size = 2048 while output feature size = 300.
Classifier	**Mid-Level Fusion:** We'll concatenate these feature representations and treat them as a new feature vector and send it through a final fully connected layer for classification.
Training	AdamW optimizer and Cross Entropy loss function. Converges after 5 epochs

Model Performance





Model	Accuracy
MultiModal (wide-resnet_50_2 + Bert-hate)	63.60 %
Multimodal (Resnet152 + FastText)	60.75 %
Unimodal Vision (Resnet152)	61.30 %
Unimodal Language (Roberta-large)	63.00 %
MultiModal (Resnet152 + Bert-hate)	61.45 %
MultiModal (wide-resnet_50_2 + Roberta-large)	63.25 %

Conclusion and Future Scope

Conclusion:

- Multi-modal outperforms traditional unimodal models.
- Language modality is more important than visual modality.
- Naively combining modalities can hurt the performance of the model.

Future Scope:

- Multimodal model slightly outperforms the unimodal model due to limited training dataset.
- Various fusing approaches can be used to combine the features.