CIS 5300 Final Project Milestone 3

Gokul Nair, Rohan Saraogi, Shivani Prasad Bondapalli, Yash Agrawal

December 12, 2023

1 Description

For this milestone we conducted two experiments to compare the performances of different embedding generation strategies.

1. **CLIP-Concat**: For the first experiment, we used OpenAI's CLIP (Contrastive Language-Image Pre-Training) model (https://github.com/openai/CLIP), a neural network trained on a variety of (image, text) pairs. In particular, we used the model to generate image and text embeddings for each (image, tweet_text) pair, and concatenated the generated embeddings.
2. **Text + Captions + S-BERT**: For the second experiment, we used Salesforce's BLIP (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation) model to generate image captions using nucleus sampling. Given an (image, tweet_text) pair, and generated image caption `caption`, we augmented the tweet text as `Tweet: tweet_text\n Image Caption: caption`. Finally, we generated sentence-transformer (https://www.sbert.net/) embeddings for the augmented text using the `all-MiniLM-L6-v2` model.

For both sets of embeddings we trained an elastic net model with different choices of `l1_ratio` $\in [0, 0.25, 0.5, 0.75, 1]$ to predict the `label` column and looked for the model that maximized the dev accuracy/macro F-score.

2 Results

For both experiments, `l1_ratio` $= 0$ marginally yielded the best results. Table 1 shows the results for this configuration (note that the precision, recall, and F1 scores are macro scores).

Table 1: Results

Embeddings	Data	Accuracy	Precision	Recall	F-Score	AUC
CLIP-Concat	Train	0.829	0.82	0.81	0.82	0.909
CLIP-Concat	Dev	0.806	0.79	0.79	0.79	0.869
CLIP-Concat	Test	0.785	0.77	0.77	0.77	0.860
Text + Captions + S-BERT	Train	0.783	0.78	0.76	0.76	0.850
Text + Captions + S-BERT	Dev	0.759	0.75	0.72	0.73	0.829
Text + Captions + S-BERT	Test	0.766	0.76	0.74	0.74	0.837

We observe that **CLIP-Concat** has consistently better performance than **Text + Captions + S-BERT**, and achieves a test accuracy and F-score of 0.785 and 0.77 respectively. This seems reasonable as, among other reasons, we would assume that there is more information in an image than just a caption for it, and CLIP would produce text and image embeddings that align well with each other. We also observe that both models have significantly better performance than our simple baseline, which had test accuracy and F-score of 0.61 and 0.38 respectively. This highlights the utility of these embeddings for our classification task. Additionally, we note that **CLIP-Concat** has roughly similar performance to our strong baseline, which had test accuracy and F-score of 0.841 and 0.78 respectively (this is caveated with the fact that the strong baseline was trained with the `label_text` column instead of `label`, and hence the comparison isn't entirely fair).