Not a Safe Space: Identifying Hateful Tweets through Transformer-Based Architectures

David Bai, Allison Lim, Leslie Moreno, Aryan Gulati

CAIS++
CENTER FOR AI IN SOCIETY'S STUDENT BRANCH

Introduction

46% of U.S. teenagers ages 13-17 have been cyberbullied[1]. Teenagers are commonly targeted because of race or gender, with 22% thinking they were targeted because of their gender and 20% because of their race[1]. Cyberbullying has also been "linked with suicidal thoughts and attempts[2]." We propose that identifying and banning the malicious actors responsible for such content reduces cyberbullying and develop machine learning models capable of doing so.

Data

Our dataset[3] comprises 19,770 human-annotated tweets for sexism, racism, both, or neither, supplemented by 589 scraped tweets based off a separate cyberbullying dataset[4]. (though complete retrieval was limited by X's current policies). Of the total tweets, 2,071 contained images, with 1,028 successfully retrieved. To improve generalizability, we preprocessed the tweets by removing user mentions, links, and recurring hashtags (notably #mkr, present in ~2% of sexist tweets referencing a specific TV show). The external tweets were labeled to match our annotation scheme with GPT-4 few-shot prompting, with manual verification through sampling, serving as an additional test set for model evaluation.

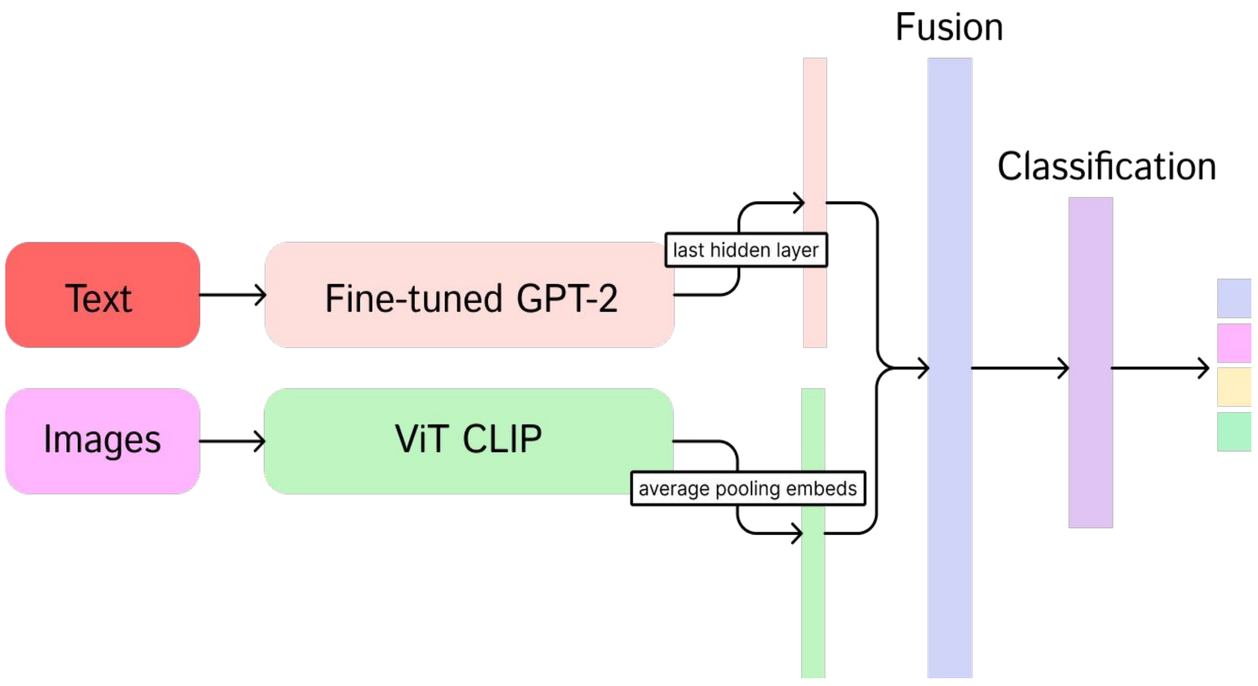

Figure 2. Multimodal Classification Model with GPT-2 fine-tuned on dataset

Results

MODEL	NEITHER	RACISM	SEXISM	BOTH
BERT Ensemble	0.8901	0.7554	0.7722	0.0000
BERT	0.8878	0.7505	0.7763	0.0000
GPT2 Ensemble	**0.9214**	**0.7941**	**0.8581**	**0.0444**
GPT2	0.8810	0.7577	0.7623	0.0000

Figure 3. Overall F-1 Scores, 5-Fold Cross-Validation

We found a fine-tuned GPT-2 works best across all four categories. For the multimodal dataset (i.e. all tweets with images), fine-tuned GPT-2 + CLIP worked best, though when we visualized the L2 Norm of normalized text and image embeddings (Fig.5), we found that the images contributed relatively little to the final classification. Regarding a low F1 score for the **both** class across several models, Fig.6 shows a unbalanced distribution across categories and a low percent of **both** tweets, meaning the models have little to work with.

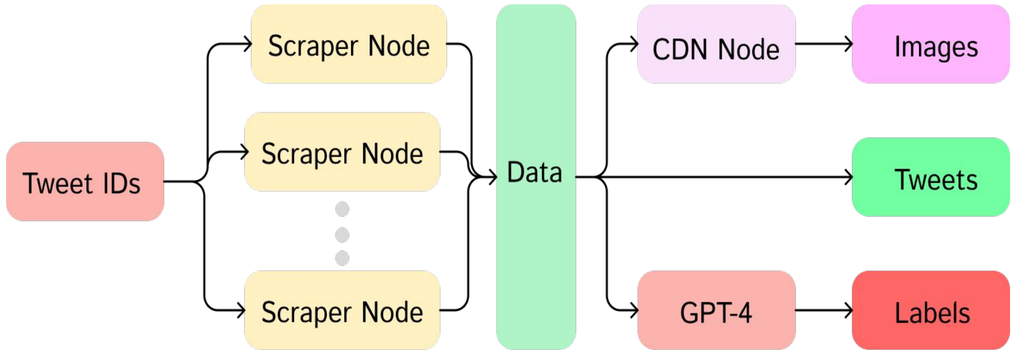

Figure 1. Tweet ID to Data Pipeline for Salawu Set [2]

Methods

For classification, we trained text-based models on the entire dataset and multimodal models on the subset of tweets with images.

Text-Based

We fine-tuned two pre-trained transformer models, GPT-2 and BERT. Both models were trained in two forms — ensemble and multiclass. The ensemble method combines the results of two fine-tuned (BERT/GPT-2) binary classifiers.

Multimodal

We trained a classification head consisting of a fusion layer(Linear, RELU, Dropout) and linear classification layer on top of the concatenated embeddings of a (optionally fine-tuned) GPT-2's last hidden layer and CLIP's vision transformer variant, for subset of tweets with images. This head allows us to consider both forms of media.

This model also had an ensemble and multiclass variant. We also tested classification purely off CLIP and off non-tuned GPT-2, both of which multimodal outperformed.

GPT2/CLIP	GPT2/CLIP ENSEMBLE	GPT2
0.864	0.830	0.857

Figure 4. Multimodal Accuracy

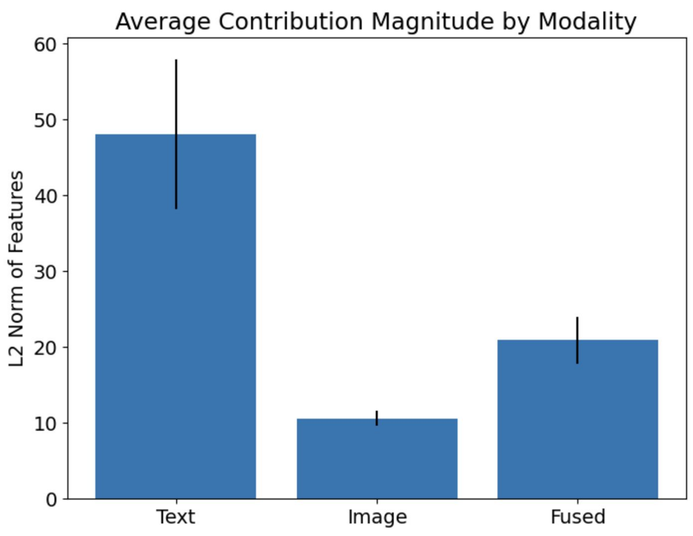

Figure 5. Embedding Magnitude Bar Chart

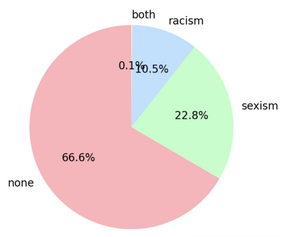

Figure 6a. Distribution of all Tweets

Figure 6b. Distribution of Tweets w/ Images

Conclusion

Cyberbullying is a pervasive problem and the detection of hate speech can be used to mitigate this harm. However, our dataset fail to consider marginalized communities outside of race and gender that are affected by hate speech such as the LGBTQ+ community. Our metrics for sexism consider a binary gender and thus fail to consider hate-speech against non-binary individuals. We also do not consider other types of cyber bullying, such as being shown unsolicited explicit images and revenge pornography. This work can potentially be generalized to other forms of social media such as Reddit, Instagram, and Facebook. These deficiencies may be mitigated by augmenting the dataset with real-world and synthetic tweets.

Works Cited

1) Vogels, E. A. (2022, December 15). Teens and Cyberbullying 2022. Pew Research Center. https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/

2) Reynolds, S. (2022, July 12). Cyberbullying linked with suicidal thoughts and attempts in young adolescents. National Institutes of Health. https://www.nih.gov/news-events/nih-research-matters/cyberbullying-linked-suicidal-thoughts-attempts-young-adolescents

3] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

4] Salawu, S., Lumsden, J., & He, Y. (2021). A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021) (pp. 146-156).