Comparative Analysis of Hate Speech on Social Media Platforms

Jessica Yeung and Anooshree Sengupta
{jyeung27, anoosh22}@stanford.edu

Stanford | ENGINEERING
Computer Science

Overview

Objective

Currently, models that identify hate speech contain racial bias or reflect other flaws in human decision-making. As recent mass shootings and other crimes are being increasingly linked to online hate speech, finding and analyzing accounts with racist or violent speech is more urgent now than ever before.

We ran two different models, a 1 vs. the rest SVM classification model and a BERT model, on our data sets in order to identify hate speech (defined to be directed speech with malicious intent) and pinpoint important features for classification. Ultimately, we hope that our model evades biases that exist in human classification of hate speech and provide insight into how and why classification differs across platforms.

Data

1) Twitter - 24,000 tweets classified as hate speech (0), offensive language (1), or neutral (2)
2) Reddit comments - 5,021 threads of comments with labels corresponding to which comments were flagged as hate speech
3) Stormfront text - 1,914 training files and 478 test files classified as hate or no hate

Because the Twitter and Reddit comments were not split into training and test data sets, we trained on 50% and tested on the remaining 50% of each data set.

One-vs-Rest Classification

- For each text class, used *sklearn* package to implement a LinearSVM model that uses one versus the rest classification, which creates a new binary outcome model for each class.
- The binary outcome models use feature vectors made up of words in each of the tweets, similar to sentiment analysis.
- Predict the label *k* for each unseen example based on confidence score from each binary classifier

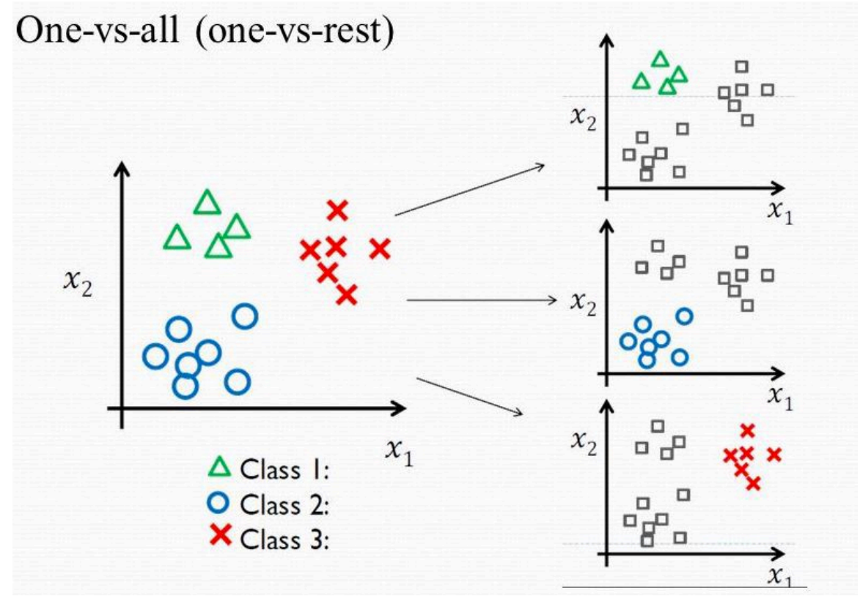

Figure 1. Visualization of 1 vs. the rest classification

BERT

Bidirectional Encoder Representations from Transformers is a new text analysis technique - instead of analyzing text in a singular direction, it is a language model that is bidirectionally trained, resulting in a better understanding of language. Unlike the models we analyzed in class, BERT reads all the words at once.

There are two training strategies used by BERT:

1. Masked LM (MLM) - about 15% of the words in the text are "masked," and the model tries to predict what the original value of the masked words. It only predicts the masked words.
2. Next Sentence Prediction (NSP) - pairs of sentences are fed to the model and the model has to predict the order of the sentences.

Both models use a classification layer of learned weights and matrixes when predicting outputs. The combined loss of MLM and NSP is minimized when training the BERT model.

Results

One-vs-Rest Results

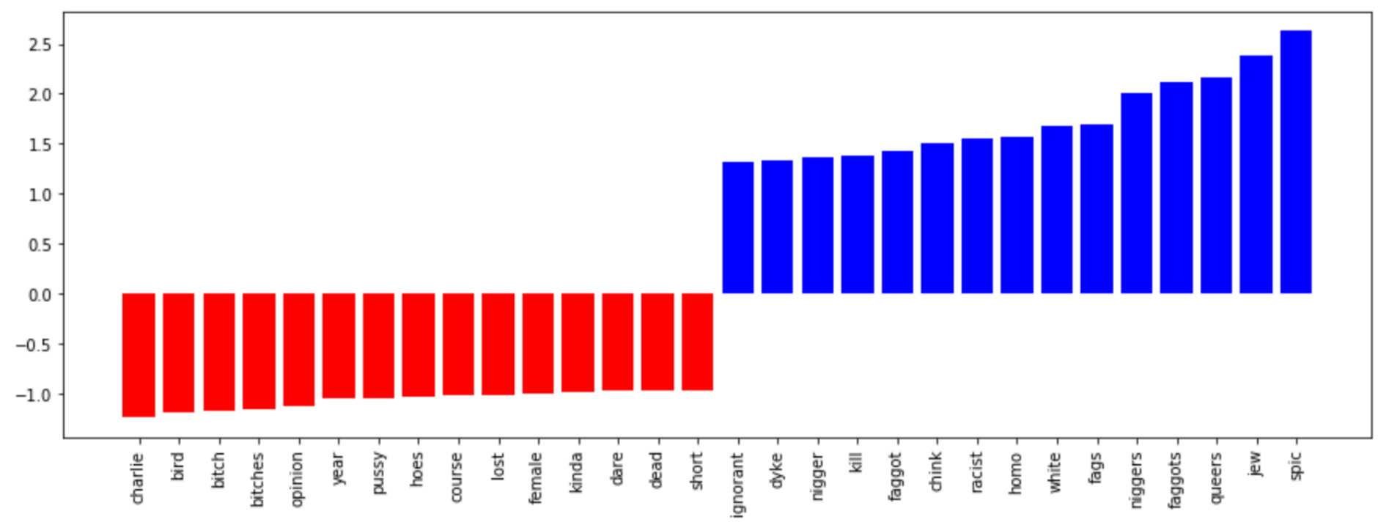

Figure 2. Top 15 Negative and Positive Word Features for **Hate** Speech Classifier from **Twitter**

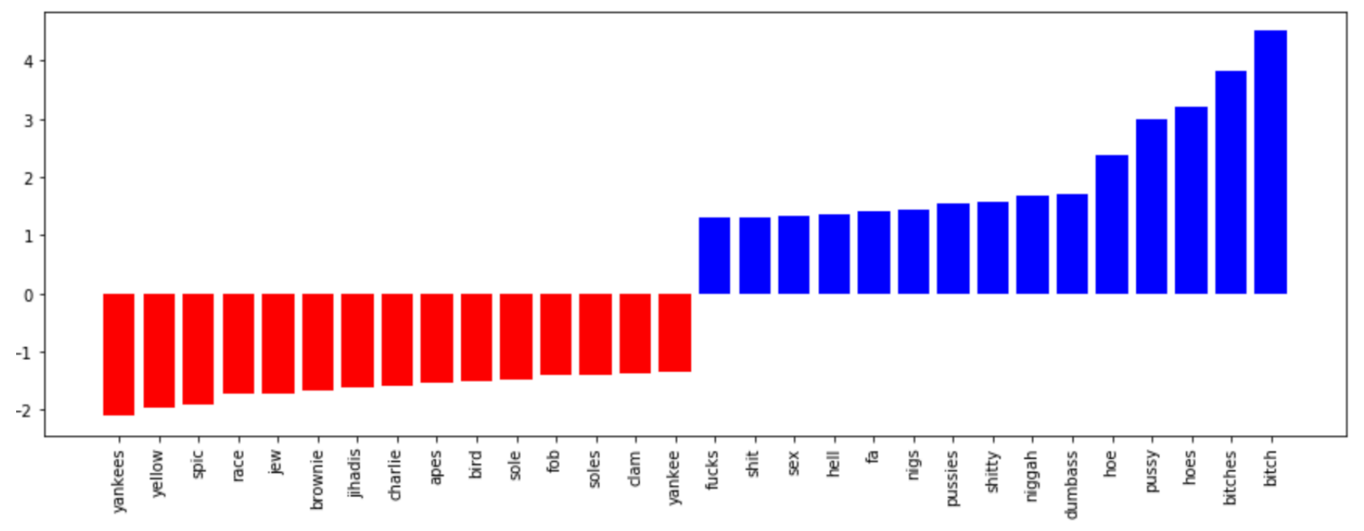

Figure 3. Top 15 Negative and Positive Word Features for **Offensive** Speech Classifier from **Twitter**

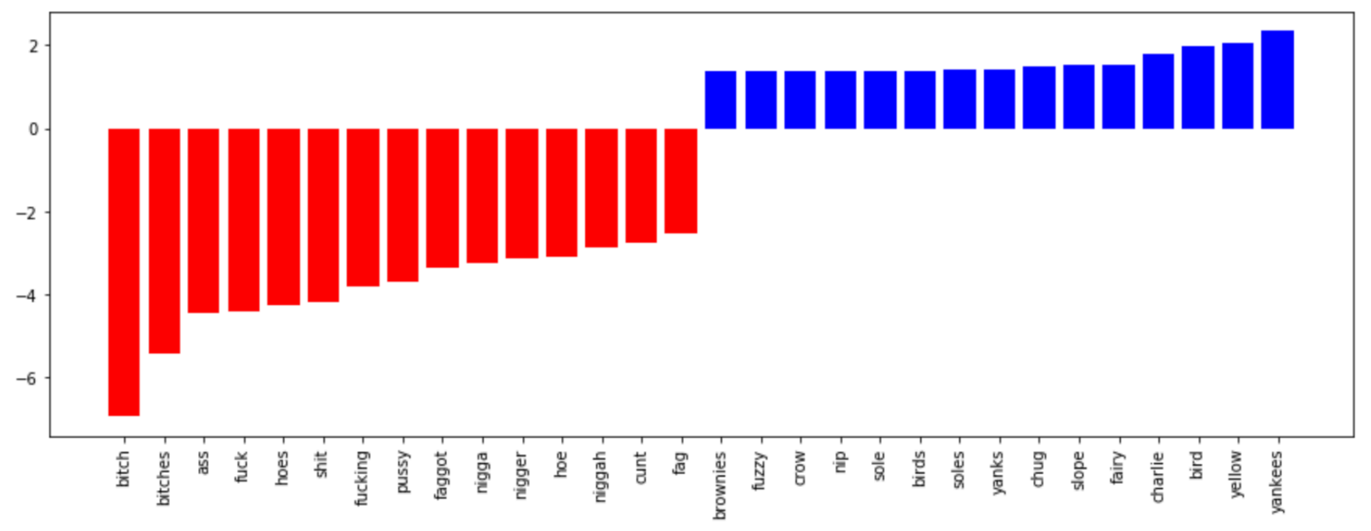

Figure 4. Top 15 Negative and Positive Word Features for **Neutral** Speech Classifier from **Twitter**

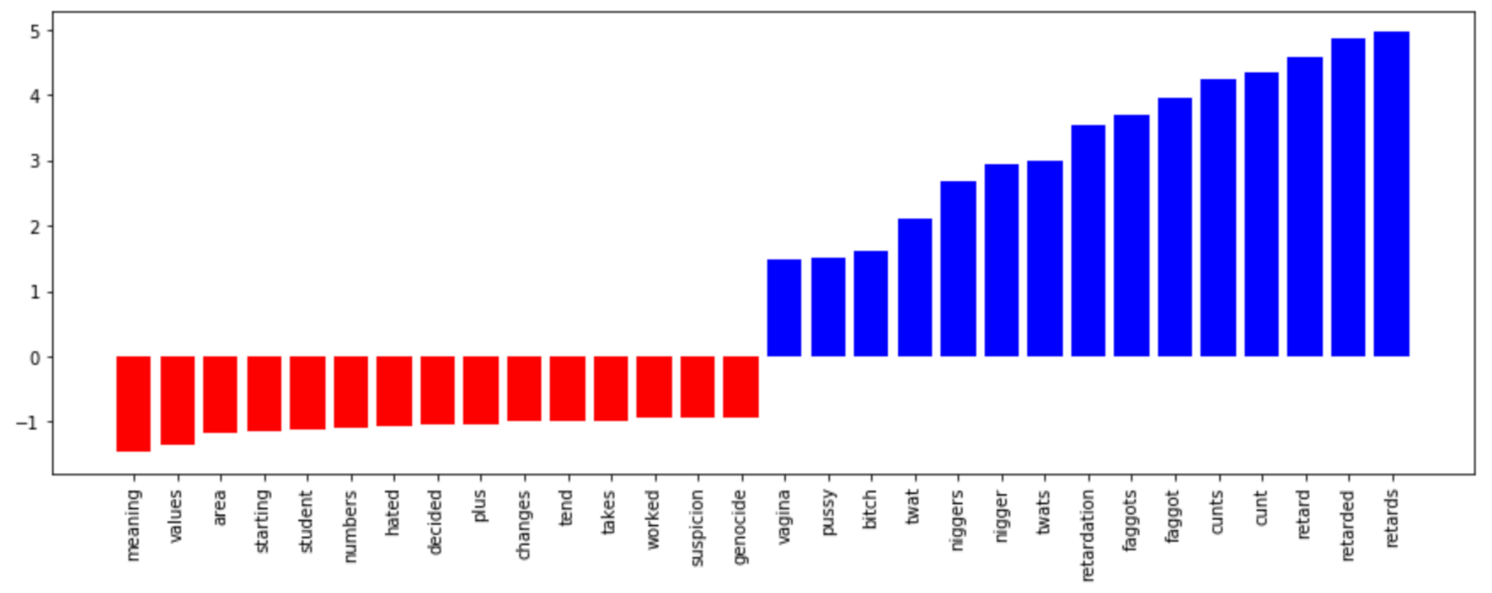

Figure 5. Top 15 Negative and Positive Word Features for **Hate** Speech Classifier from **Reddit**

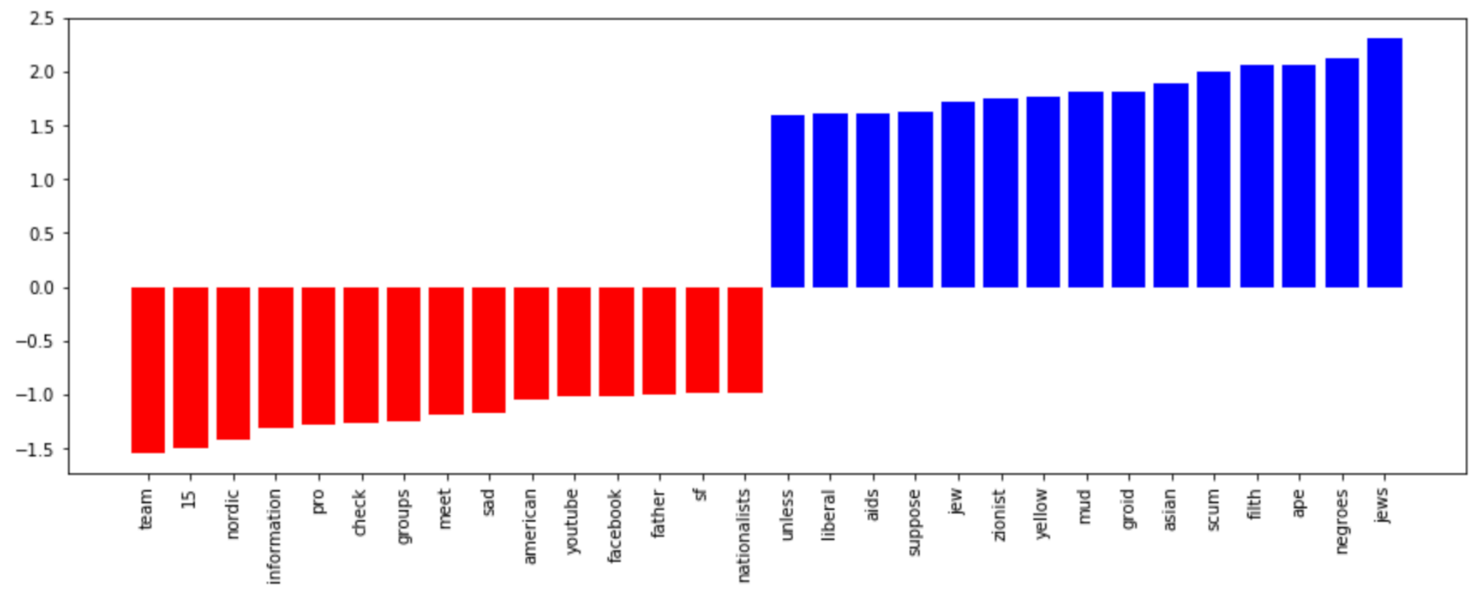

Figure 6. Top 15 Negative and Positive Word Features for **Hate** Speech Classifier from **Stormfront**

Twitter:	**Reddit:**	**Stormfront:**
Train error: 0.0726	Train error: 0.093	Train error: 0.112
Test error: 0.096	Test error: 0.132	Test error: 0.297

Challenges

- Not ideal to have to manually label other data sets to match initial data set's classification labels
- How can we have a model that can detect important features across multiple data sets?

BERT Results

We ran BERT on the same Twitter data set we ran our one-vs-rest classification algorithm on.

Loss: 0.2225
Accuracy: 0.9123 (Test error: 0.0877)

The BERT model has 1% higher accuracy than the multi class model for the Twitter dataset, but expanding it to other datasets may require modifying layers in our neural network or changing the classifications in our datasets.

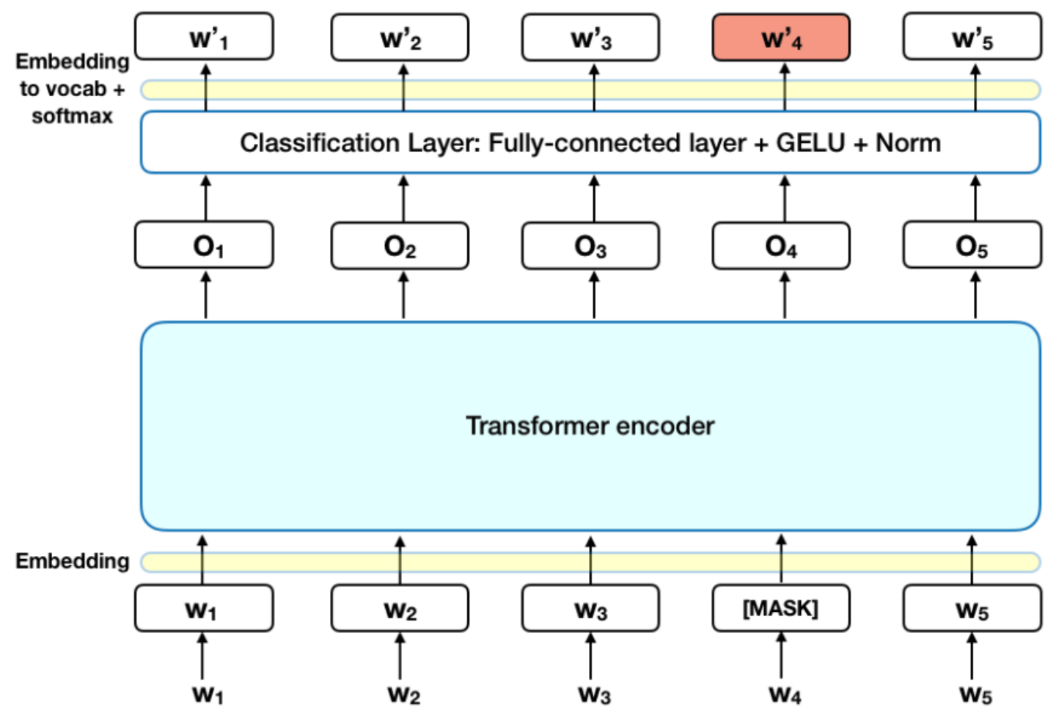
w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Figure 7. Visualization of MLM

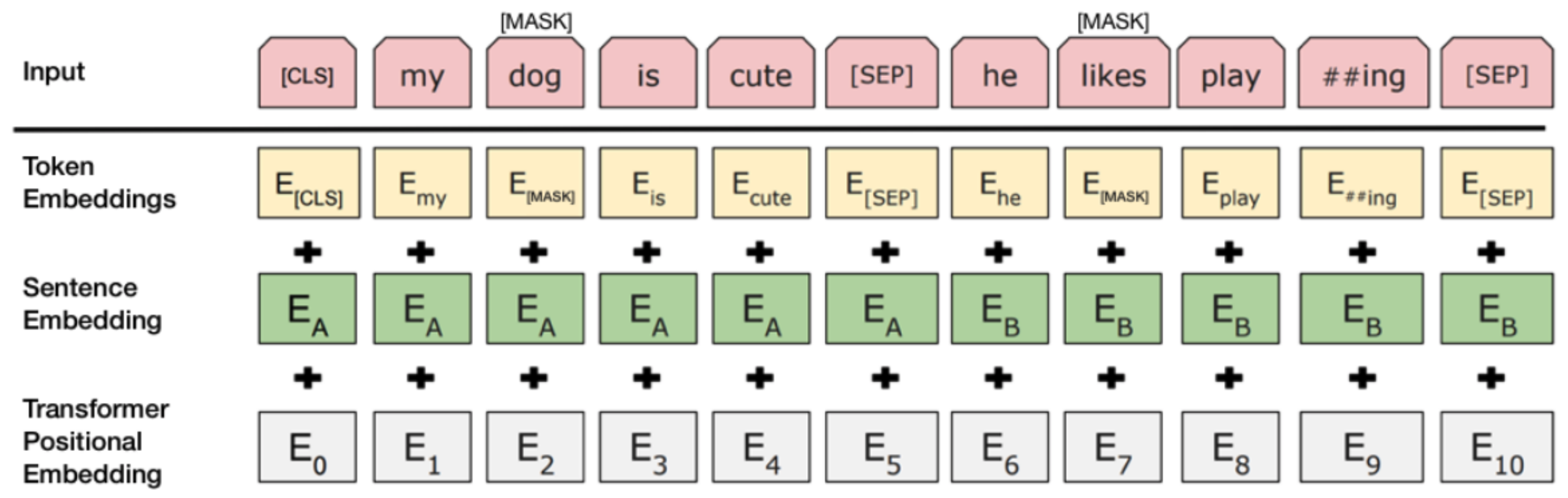

Figure 8. Visualization of NSP

Discussion

- Feature weights intuitively make sense: the binary classifier heavily weights those in both the Twitter and Reddit datasets that align with the classes we are using. For example, racial slurs are heavily weighted in hate speech, while neutral and positive words suggest normal speech.
- Disadvantages to BERT: BERT is comparatively slower, uses more memory, and the model needs to be saved. Running BERT with more than one epoch will vastly improve the performance of our model, but may not be feasible without using a powerful GPU.
- We hope to continue investigating the inherent structure of this problem: are there key words or phrases we need to control for? What features are being used to classify neutral or hate speech when they shouldn't be?
- With BERT, we can continue to incorporate additional features and compare accuracy between datasets, but we have less insight into the words that define each class of language.
- As we continue to incorporate more datasets, we hope to understand how our accuracy in classification differs from our original expectations for each platform—is that difference significant, and how does it manifest itself? Are the instances of hate speech inherently different on different platforms?
- Many of our heavily weighted features consist of racial slurs directed towards Asian and African American populations. If we were to "correct" for racial bias and remove those features, would we notice a difference in classification? Is there a different way to define hate speech outside of race-related language?

Future Directions

- Want to incorporate additional features for a more complete understanding of how hate speech is circulated. These features can range from a count of racial slurs to account for racially-charged language, to tagged accounts in a tweet, to information about an account, forum, or article that contextualizes a post.
- Investigate the inherent structure of this problem: are there key words or phrases that we need to control for? Tweets written by African Americans are more often flagged as offensive/hate speech. We think that we can evade this bias by eliminating certain types of language and words typically seen in those tweets and then retraining our model. Hopefully, we will observe a more accurate model, with fewer misclassifications for tweets from those accounts.
- We noticed that our results differ from our original expectations for each platform—for example, the one-vs-rest model performs poorly on Stormfront comments. We hope to understand why that difference occurs, and if it indicates a significant difference between the types of hate speech on each platform.
- Want to cluster hate speech among and within each platform to understand the different types of hate speech and their defining features. This could help us improve our model.
- Considering building on top of the BERT binary classification model (which is more accurate, and will be faster with a GPU), to run the same model across multiple data sets and conduct a thorough comparative analysis.

References

1. Bakharia, Aneesha. "Visualising Top Features in Linear SVM with Scikit Learn and Matplotlib." Medium, Medium, 1 Feb. 2016.
2. Demszky, Dorottya, et al. "Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings." Proceedings of the 2019 Conference of the North, 2 Apr. 2019, doi:10.18653/v1/n19-1304.
3. Ghaffary, Shirin. "The Algorithms That Detect Hate Speech Online Are Biased against Black People." *Vox*, Vox, 15 Aug. 2019.
4. Horev, Rani. "BERT Explained: State of the Art Language Model for NLP." Medium, Towards Data Science, 17 Nov. 2018.
5. Maiya, Arun. "BERT Text Classification in 3 Lines of Code Using Keras." Medium, Towards Data Science, 16 Oct. 2019.
6. Trivedi, Kaushal. "Multi-Label Text Classification Using BERT â€" The Mighty Transformer." Medium, HuggingFace, 13 Feb. 2019.