Detecting Hate Speech in Tweets: Datasets & Methods for External Evaluation

Debra Cooperman

Abstract

Elon Musk's recent takeover of Twitter, with its accompanying loosening of content restrictions, has provoked assertions that hate speech is now more prevalent on the platform. At the same time, the proprietary nature of the algorithms used by companies such as Twitter and Facebook, the sheer amount of data needed to train models, and nuances of defining hate speech above and beyond a list of direct keywords, all make the process of detecting and addressing online hate speech from "the outside" very difficult. This project attempts to replicate and reinforce the results from a recent study that addresses some of these issues. Specifically, the study authors present "ETHOS", a textual dataset that is generated through an "active sampling" procedure and extensive human annotation protocol that allows them to effectively train hate speech models with a small amount of labelled data. This dataset was most effective when used in conjunction with BERT, a deep learning language model, which provides much more context than frequency-based or purely sequential models. In my project, I replicated the high accuracy, recall, and f1-score of the BERT model trained on the ETHOS training data and validated on the ETHOS test data, and maintained the high performance of this model on two additional datasets.

1. Introduction

According to Wikipedia,. hate speech is "a form of insulting public speech directed at specific individuals or groups of people on the basis of characteristics, such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity." With the prevalence of online social platforms that have the ability to propagate hate speech more broadly than ever before, the question of who is responsible for properly policing this speech while simultaneously avoiding censorship and promoting the flow of ideas has become increasingly important. The choice becomes whether to depend on the largest of these social platforms (Twitter, Facebook, YouTube, etc.) to self-police, at the potential expense of profits, or address the challenges of detecting hate speech using publicly available data and methods.

Some of these challenges include: gaining access to data, being able to identify true hate speech vs. other offensive language[1], the sheer amount of data needed for training, and class imbalance and bias due to the reality that most online speech is not hate speech.

The researchers behind "ETHOS: a multi-label hate speech detection dataset"[2] have focused on the issues of class imbalance and bias through generating their dataset via an admittedly "time-consuming" process, but one that allows them to gain valuable learning from a small amount of labeled observations (~1,000 for the binary classification dataset). An important key to their approach is the idea of "active sampling", meaning data is sampled daily from YouTube and monthly from Reddit, is put through a simple model to establish an initial "hate" score, with a "stopping" threshold that occurs when the data becomes balanced both in terms of classes of hate speech (e.g., Gender, Race, etc.) and in terms of achieving diversity of the comments per class. By prioritizing having more diverse ways in which a certain class of hate speech can be expressed, this dataset is able to "cover more ground" with a smaller amount of observations. Once the threshold of observations is achieved, it gets sent through several layers of human annotators who have been vetted through continued evaluations, and the ultimate score of "isHate" or "isNotHate" is the normalized distribution of several annotators votes.

2. Replication of ETHOS experimental results

2.1 Transformer-based approach. The ETHOS study authors experimented with several different types of modeling, from simple "bag of words" and frequency models to Transformer-based models. DistilBERT got the best results across all the metrics for binary hate speech classification. DistilBERT is a "distilled" or less expensive version of BERT (Bidirectional Encoder Representations from Transformers), which is based on the Transformer architecture of neural networks. Transformers are predicated on the idea of evaluating relationships between words in context, dynamically updating the significance of these relationships as they gain more information. BERT in particular, through a process called "masking", *only* focuses on the *relationships* between words vs. relying on the fixed meaning of the word independent of context. BERT's ability to stack multiple Transformers and read bidirectionally at the same time means the full context is brought to bear, making it an extremely valuable approach for interpreting hate speech from context.

2.2 Methodology/Additional Datasets. I used a keras wrapper called ktrain to train and validate a DistilBERT model on the ETHOS dataset, then tested that model on two additional datasets to truly test the generalization of the model:

- Twitter Sentiment Analysis: The test set of this kaggle dataset is 31,962 tweets that have been labeled as "hatred(racist/sexist)" and "non-hatred(racist/sexist)"
- Hate Speech and Offensive Language Dataset: This kaggle dataset is 24,783 tweets labeled as "hate-speech", "offensive language", and "neither". (I combined the "offensive language" and "neither" categories into one "isNotHate" category).

One of the most important inputs to training in a neural network is finding the right learning rate to efficiently minimize the loss function. If the learning rate is too low, it will take too long to converge. If the learning rate is too high, there's a risk of overstepping and not minimizing the loss at all. I used ktrain's simulator to determine the best learning rate to use. I also experimented with different ways to cycle the learning rate, landing on the "one cycle" policy first introduced by Leslie Smith [3]. Using this policy, a maximum learning rate is first chosen using the simulator, then a lower learning rate of 1/5th of max learning rate is chosen. Then the cycle goes from lower to higher learning rate in one step and back to lower learning rate in a second step, with the full cycle length less than total number of epochs. In the last iterations, an extremely low learning rate is chosen (1/10th). This helps to get out of saddle points, and when the learning rate is higher in the middle of learning, that works to regularize and avoid overfitting. In my experiments using a non-cyclical learning rate adjustment, all of the scores were lower and took longer to converge. Using ktrain's `one_cycle` method, I was able to achieve comparable scores to the ETHOS study for all three datasets in 6 epochs, using a batch size of 6. (See *Table 1*).

I then used ktrain's `view_top_losses` function to identify the top 5 values in each dataset where the model was mistaken. (See *Figure 2* for examples from the ETHOS dataaset).

2.3 Results. As mentioned previously, on the ETHOS test data I was able to replicate comparable scores to the original ETHOS study.

	Precision	Recall	F1-score	Support
0	83%	80%	81%	85
1	75%	78%	77%	65
accuracy			79%	150
macro avg	79%	79%	79%	150
weighted avg	79%	79%	79%	150

Table 1: Scores for ETHOS dataset. The scores generated in this experiment are comparable to the scores generated by the ETHOS paper authors using DistilBERT.

In identifying Hate Speech, a large amount of false positives are a concern because we want to err on the side of not censoring speech. The confusion matrix for the ETHOS test set has a reasonable amount of false positives at 11.33%, although there's room for improvement (See *Figure 1*).

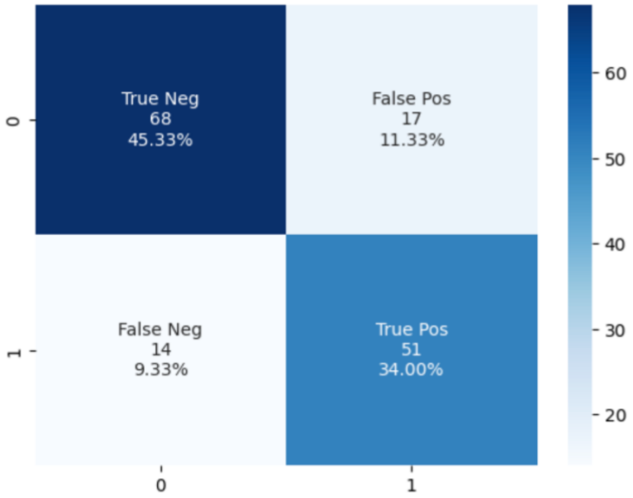

Figure 1: Confusion Matrix for ETHOS dataset

False positive example: `I wouldn’t like this girl as a police officer. I want police to keep me safe, she and every other woman cannot keep us safe they are week`

False negative example: `Deport the fucking brainless worthless pieces of shit freeloaders after they are all shot`

Figure 2: FP and FN examples for ETHOS dataset

In the trained model, an observation is labelled “hate speech” if the overall score is >= .5. In the false positive example above, this probably has a relatively low score (closer to .5), as it is sexist but is not as strong a statement as the false negative example. In the false negative example, this may have been predicted as not being hate speech because it is not specific enough in its targeting of a particular population, so it may be that more data needs to be gathered related to hate speech against immigrants in particular, as this tweet appears to be addressing that population.

Scores for the Twitter Sentiment Analysis Dataset and the Hate Speech and Offensive Language Dataset were also comparable to the original scores achieved by the ETHOS study authors (See *Tables 2 and 3*).

	Precision	Recall	F1-score	Support
0	94%	98%	96%	29720
1	44%	25%	32%	2242
accuracy			93%	31962
macro avg	69%	61%	64%	31962
weighted avg	91%	93%	92%	31962

Table 2: Scores for the Twitter Sentiment Analysis Dataset

	Precision	Recall	F1-score	Support
0	96%	82%	88%	23353
1	11%	38%	17%	1430
accuracy			79%	24783
macro avg	53%	60%	53%	24783
weighted avg	91%	79%	84%	24783

Table 3: Scores for the Hate Speech and Offensive Language Dataset

3. Conclusions

As shown above, training with a dataset that is balanced both in class distribution and diversity of observations is an effective combination with the transformer-based approach of BERT. This approach, combined with learning rate optimizations, was able to perform well not only on the original test set, but generalized to two other separate test datasets that were unrelated to the original ETHOS training data. Given that BERT and related approaches are usually time-consuming and require large amounts of data, it is a promising development that faster conclusions can be achieved using smaller datasets in the future.

References

[1] Davidson, T, Warmsley D, Macy, M and Weber, I 2017 *ICWSM Proceedings* **1** 1703

[2] Mollas I, Chrysopoulou Z, Karlos S, Tsoumakas G 2022 *Complex and Intelligent Systems* **8** 4663

[3] Smith L 2018 *US Naval Research Laboratory Technical Report* 5510

Code and data sources accompanying this report can be found here.