Derek Powell

Springboard Data Science Track

Capstone 2 – **Detecting Abusive Speech**

Final Project Report

As the world of digital media continues to grow and permeate every part of our life, certain aspects of communication remain the same. We should seek to communicate with strangers, colleagues, associates, and generally most others in a manner which is courteous, professional, inviting, and effective.

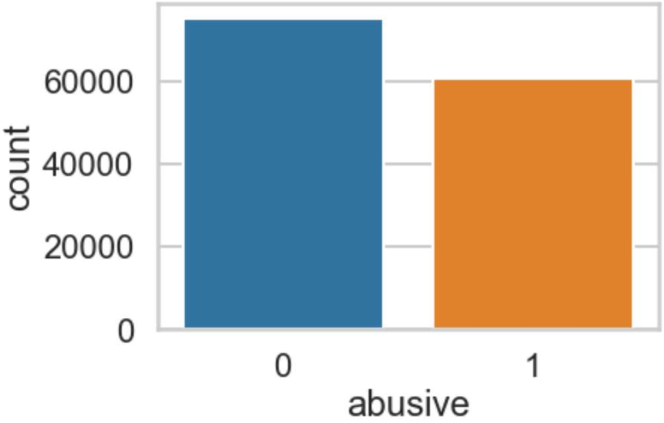
Effective communication leads to human interest and engagement. Conversely, ineffective, rude, off putting, or abusive language has the opposite effect. It tends to deter human interest and engagement. If a business featuring an online communication platform is to be successful, effective communication is what they want to encourage. It then becomes obvious that there is a growing need for companies to be able to effectively (and efficiently) moderate the communication of those using the platform; to restrict abusive communication, and limit or eliminate other forms of ineffective communication.

This is the goal of my project, to develop a machine learning model which has been trained to detect and report on abusive communications in digital text. In starting a project of this nature, it was critical to start with a labeled dataset of significant enough size to allow for proper model training. To that end, I used a dataset consisting of 135,000 ‘rated’ tweets, from Twitter users. The tweets were annotated by a varied base of volunteers, with enough variation to reasonably offset individual bias as much as possible. The dataset is credited below.

The dataset featured a ‘Hate Speech Score’ which was a float value ranging from -8.00 to +6.00, and was the aggregate of a series of subjective judgements by the annotators. The higher this value was, the more ‘abusive’ the text was considered to be.

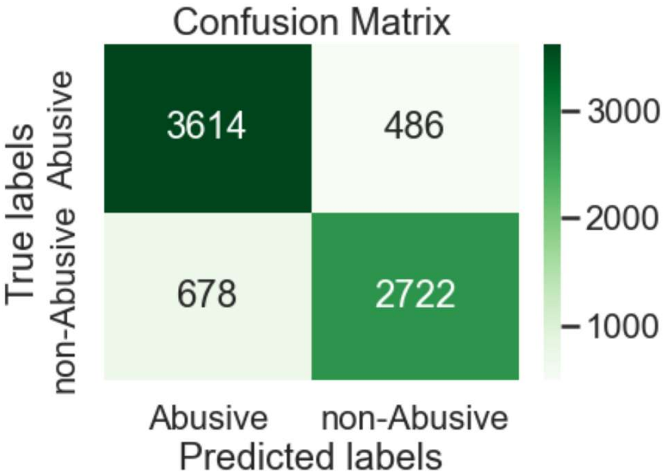
In this project, my goal was to predict the class of the text, whether it was 1 – Abusive, or 0 – Non-Abusive. As such, one of my first steps was to engineer a new binary feature: a value of 1 if the Hate Speech Score was >0 and a value of 0 if the Hate Speech Score was <0. From there, I dropped all columns of the data except the Text column, containing the tweets content, and the new binary ‘Abusive’ column.

In EDA, I explored my new Abusive binary column to see if I was dealing with imbalanced classes. Surprisingly, I found the classes to be quite close.

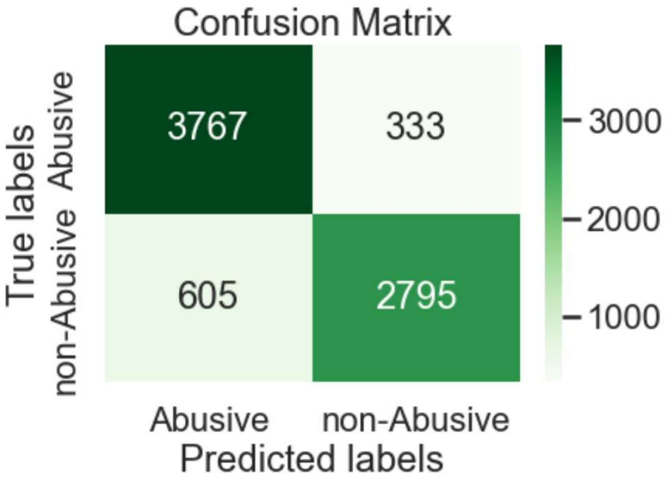


In Pre-processing of the text, I removed all special characters from the text, and equalized the Case for all letters (lower). Later, in creating the model, I applied further pre-processing by dropping English 'stop words' and limiting words from appearing in the vectorized corpus vocabulary unless they occurred in at least 10 tweets. All of these steps ensured I would have a consistent and clean vocabulary to train a model with. I used a TF IDF Vectorizer to make sure to capture important of certain words in correlation to their usage frequency and score significance.

The first model I used was a Logistic Regression model. Using fairly out-of-the-box parameters, this began us with a Accuracy of 84% and a Recall of 88% (predicting abusive texts correctly).



My initial goal was to achieve 90% or higher Recall, so I had more work to do. Using GridSearch and Cross-fold validation, I tried various other combinations of hyperparameters for my Logistic Regression model, but ultimately could not improve on the score significantly (achieved a 0.2% increase only. I moved onto a Random Forest model, but I started with some non-default parameters. This immediately gave us an appreciate improvement over the Logistic Regressor, boasting a confident 87% Accuracy and a 92% Recall.



As with the Logistic Regressor, I attempted to improve on this score with GridSearch and CV, but I similarly could not improve on the performance of the model with any significance. The final best-result model was capable to predict Abusive texts with 92% accuracy, leaving only 8% of Abusive texts misclassified.

Additional work that could be done to possibly improve further upon the results would be to try additional models (e.g. Ridge Regressor), or to use the original space matrices (post-vectorization) in the model training instead of casting them into data frames. Additional text could be brought in from other non-Twitter sources to diversity the nature of communication to more closely align to a different platform (e.g. gaming industry).

@article{kennedy2020constructing,
title={Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application},
author={Kennedy, Chris J and Bacon, Geoff and Sahn, Alexander and von Vacano, Claudia},
journal={arXiv preprint arXiv:2009.10277}, year={2020}