Toxicroak

Description

- The need to identify toxic comments in conversation is more important as the number of users on the internet increases. Thus, this project hopes to use a text classification algorithm to better predict the use of toxic comments/hate speech in forums to facilitate better commenting and conversation online
    - The project was presented as a $35,000 competition on Kaggle and uses data as provided from Wikipedia's talk page edits
- The project involves building a multi-headed model that's capable of detecting different types of of toxicity like threats, obscenity, insults, and identity-based hate better than Perspective's current models

Approach(es) Summary

1. We will first approach this problem by using a Naive Bayes text classifier as it is one of the simpler text classifiers to implement.
2. We will implement support-vector machines to gain better accuracy on the text classification
3. **Stretch Goal** We would like to learn more about Google's word embeddings to improve our accuracy from support-vector machine text classification.

Dataset(s)

- The dataset that we're using can be found on Kaggle, under the name "Toxic Comment Classification Challenge".
    - In addition, there is a pre-processed dataset that can be found here that we could use instead.
    - A participant in this competition also created data sets by running the original training data through Google translate in different languages that can be used for augmentation, which can be found here. However, this could possibly result in overfitting to those specific examples.

Evaluation Metric(s)

- Since this is a multilabel classification problem we intend to use the following evaluation metrics for our model:
    - Hamming Loss to evaluate our model's ability to correctly label data overall.
        - Another metric that we could use is the Jaccard Similarity Coefficient Score.
    - In addition, we will be evaluating individual labels using typical precision and recall.