Capstone Project Proposal: Detecting Hate Speech on Twitter

Twitter usage has proliferated in the past decade, with 500 million new tweets posted daily as of 2020 (https://www.dsayce.com/social-media/tweets-day/). Unfortunately, hate speech can go undetected in this mountainous stream of content. Thus, Twitter has an ethical responsibility to create more robust processes for detecting, flagging, and even removing these toxic posts.

Towards this end, I will be developing a Twitter hate speech detection algorithm. This will be a binary classification problem, where the algorithm will take as input a single tweet and return a label of either “hate speech” or “not hate speech”. Additionally, the algorithm will return a confidence level for the given label (e.g. “hate speech, 84%”).

For training the algorithm, I will use a supervised learning approach. I have found numerous datasets that comprise tens of thousands of labeled tweets each. Some of these datasets contain the original tweet text along with a label, while others only contain a unique tweet ID number with a label. For the latter, I will use the Twitter API to scrape the corresponding tweet texts for each ID number. Eventually, I will feed the data to a deep neural network for training.

The final deliverable will be two-fold: an API along with a web application. Once my model is ready for deployment, I will package it into an API that can be called to return a label and confidence level for a given tweet. Then, I will build a web interface that takes a Twitter account name as user input, then leverages the API to produce a “hatefulness score” for the given account.

While my combined data contain over 120 thousand samples of labeled tweet data, they only take up a modest 18.3 MB. I anticipate the most computationally

expensive steps will include feature engineering, model training, and parameter tuning. However, the computational resources allotted to me by Springboard ($300 in credits) should more than suffice for the project.