Abusive Language Detection Against Immigrants and Women

Introduction

The aim of this project is to design and build the most effective classifier of offensive tweets towards the target groups of immigrants and women as possible. The project is based on a similar project run online, and as such I will be comparing my work and results to the results of other participants of the competition. As a result of there being other people working with the same data set that I have used for this project, one goal of the project was to beat the highest score of the competition, through a thorough understanding of natural language processing, and experimentation with different forms of classifiers, as well as neural networks.

The intended audience of this project is anyone interested in natural language processing, but also anyone interested in the practical application of offensive speech detection online. This is especially beneficial in the current world, given the rise of hate speech online on popular platforms such as Twitter. While this project was developed and tested on a specific set of tweets, with some modification it could be used to classify tweets in real time. Therefore, I would argue that the techniques and models explored in this project would be beneficiary to anyone looking at implementing an automated hate speech detector into their website.

The scope of this project adapted over time, due to the fact that at the start of the project the libraries and concepts I was going to be working with were largely unknown to me. Therefore, it was hard to estimate the amount of time it would take me to complete pieces of work.

The way that I approached carrying out the project was through regular meetings with my supervisor, coupled with a Kanban approach to working. This meant that I was focussed on small chunks of deliverable work, that I could then discuss with my supervisor and plan out what to work on next. This also meant that the scope was able to remain flexible throughout the project, and I was able to pivot quickly in the event that circumstances changed, which they did at some parts of the project.

[INSERT ASSUMPTIONS HERE]

In terms of the important outcomes from the project, there were a number of interesting outcomes, as follow:

* Word embeddings based on Word2Vec were the most effective way of determining features for the models to use.
* SVC was the most effective model

Background