**CS410 – Data Science Project Plan Kareem Aboueich**

**Paper Topic**

For this project, I am planning on analyzing hate speech online – what qualifies, what doesn’t, and seeing what sorts of disparities exist between what crosses the hate speech boundary. This is something that, while companies have sworn to improve over the past few years, is still a struggle and raises many questions. I will be using a multitude of data from Twitter, some of which was declared hate speech and some which wasn’t, as well as a list of data sets that researchers in the past have used to distinguish what is hate speech from what isn’t.

**Research Planned**

The plan is to observe the data given and try and distinguish some of the differences between what qualifies as hate speech and what doesn’t. A website called hatespeechdata.com has several articles on research that has already been done on this field, in multiple languages, and which I will also try and summarize some main points from before beginning to hopefully be able to do some research that they might have missed or overlooked. I will also be exclusively focusing on hate speech in English, even though the problem of hate speech is present in many other languages and cultures. Since the main dataset I will be using is over 15,000 lines, I likely will select a small sample from it and analyze it extensively, trying to solve a few questions:

1. To a reader not familiar with the subject, does this sound like hate speech?
2. Does the context of the tweet play a role in why it was or wasn’t declared hate speech, and furthermore, what means and ways could have caught a tweet that slipped through the cracks?
3. How many of the tweets that were not declared hate speech violated some rule on Twitter’s terms of service?
4. Is it advantageous to declare certain speech hate speech and others just a violation of the rules for some other reason (even if the tweet should be qualified as hate speech)?

The hope is to come to some sort of an understanding on the intricacies of what qualifies as hate speech and means and ways of improving it, seeing as the current system is not as effective as it should be and can have catastrophic results when not effective. (See: [Facebook Myanmar](https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html))

**Team Structure**

I, Kareem, am working alone on this project, so there is not much of a team structure. I am consulting a friend of mine who did a project of hers on hate speech, as well as reaching out to a few other people who work in content moderation to get somewhat of a better sense of the job, my project, and ways to improve both. (hopefully)

**Milestones**

Thus far, I have gotten a few things done:

1. Obtained the research data I am planning on analyzing (<https://data.world/crowdflower/hate-speech-identification>)
2. Began the search for additional data. In addition to Twitter, I would ideally like to analyze data from Facebook as well, as it has billions of users instead of hundreds of millions.
3. Found a webpage with existing research on hate speech that I can use as scholarly articles (<http://hatespeechdata.com/>)
4. Began analyzing and taking notes on some of the interesting factoids I found in the articles from the site above.
5. Reached out to a few people who work in content moderation to discuss my topic, and hopefully to improve it or find more questions to ask.