Jacob Sutter, Jonathan Mendez, and Jonathan Schaeffer

Progress Report

The first step in being able to analyze all of our data was to clean it all. Two scripts were made that removed all of the deleted and empty comments. A tool called jq was utilized that is able to quickly and efficiently manipulate json files on the command line. Jq was able to identify comments that had been deleted and would delete all the keys associated with the array. After those scripts were run, we needed a way to find the total amount of comments for each json file. Each json file has all the public reddit comments from each month of the respective year. To get the comments, a linux command line tool called wc reads all the lines in a text file. Since we filtered out the data we don't need from each json file, we know that each array takes up 8 lines of text. So, whatever number wc returns, which is the total line count is divided by 8 to get the total amount of comments per json file.

The next step was to determine what exactly hate speech is. We utilized a tool written in Python called HateSonar which can be found on Github that will analyze a string of text and output whether it is Hate Speech, Offensive Language or Neither. So, a python program was written to take the comments from our json files and categorize them. The comments that were reported as Hate Speech were placed in a separate file called hateSpeechComments.txt and the Offensive Speech comments were placed in a separate file called offensiveSpeechComments.txt. Not only were the comments placed in the files but also information such as author and subreddit. The 2005 and 2006 data ran very quickly but as we got to 2007 and 2008, things took a lot longer. It turns out by default Python will not run on multiple cores, so we had to adapt the program to run on 4 cores to get through the data four times as fast. But once we got more years in, the program still run slow. The problem now was that too many write commands were slowing down the system. So we decided to forgo keeping the comments and other data in return for a just a count of hate comments and offensive comments. It's just unfeasible with the hardware that we have access to, to keep all that data and get our processing done in an appropriate amount of time. We are now reaching a problem of the json files reaching a file size that is too much for python to handle and is too much for our hardware to handle. So instead of grabbing the entire file, we are only taking in 100,000 lines of code at a time. This saves the amount of RAM consumption there multiple gig files takes up when python inputs them.

The way we are collecting this data is through google sheets. We are collecting the total comment count of each month, the count of hateful comments and the count of offensive comments. Since the data grows in size exponentially each year, we needed a way to fairly determine if there has been a rise in hate speech over the years. While there are definitely more hate speech comments, there are more comments in general. So we take a percentage of hate comments and a percentage of offensive comments compared to the total amount of comments on Reddit. The main problem with this is that we may be missing out on a lot of hate comments that had been deleted by moderators on subreddits. We are not totally done parsing through all of the comments that we have, this is mostly due to hardware limitations but we are going to begin sharing the load between the three of us to get the last years tone, which will allow us enough time for analysis.

The plan is to take the data from the google spreadsheet and export it as a csv file. We will then analyze on a month by month basis to see if there has been a rise in hate speech in terms of raw comments. Then, we will look at it yearly. This will give us an idea if the volume of hate speech has been on the rise over the years. Then we will find out what percentage of comments on reddit are hatespeech. This is more fair of an assessment because the growth rate of reddit comments by year (and by month) is exponential. If there is a percent increase or decrease then it would be easier to say, yes there is a rise in hate speech or no, there isn't. It may also be worth while to take the percent of hate speech comments by the entire year to see at a macro level if hate speech has been on the rise.

We are assuming that there may be months that hate speech use peaks, or increases drastically. It might be possible to look at past world events such as elections or other polarizing events. People may feel more inclined to post hateful things on the internet when tensions are at there highest. Also, maybe peeks in hate speech could have lead up to mass shootings or terroristic events across America. All our comments are in English and only from reddit.com so its save to assume everything is internal to America. So not are we only analyzing hate speech, but we are also hoping to find connections with the possible rise of hate speech and events that took place in America.

For our analysis, our plan is to look at both the percentages of hate speech and offensive comments on a per month and per year basis. We plan to come up with a weighted scale to equally compare months across years since there are more total comments in each year. We will look for trends related to the timing of changes in percentages, the amount of change, and the overall direction over time. Our plan is to create a number of separate graphs that show trends within each year and to also create a graph that shows each year as a separate color so that the years can be compared by month. We will run correlation analysis, determine a confidence interval, and look at a regression model for the data.

The most helpful thing that we can do in terms of analysis will be to look for trends on both a monthly and yearly scale. We will look for trends across all the years, in each month of a year, in each month over all the years, and between specific pairs of months and years that coincide with an important world event to see if that event had any impact of the comments that people were posting. What will actually make sure that this analysis is helpful will be to carefully separate and label all of the individual months and years since we are working with a lot of data.