Twitter Sentiment Analysis

Tucker Miles[1] and Vicki Tang[2]

***Abstract*— We will be collecting tweets from Twitter accounts of the administration, faculty, staff, and students of the University of Tennessee in Knoxville. The textual sentiments from these tweets will be calculated and compared based on the pre and post COVID-19 time periods. Using the comparison, we will be able to analyze the general attitudes of the tweets from the group's social media presence between the two time frames. If there is additional time, we will conduct further comparison of this data with other universities and analyze the word frequencies of the data for any significant results.**

I. INTRODUCTION

As everyone knows, there has been a tremendous amount of social media buzz surrounding the COVID-19 outbreak and other related topics. This has lasted from about March of this year and continues to produce significant social media discourse to this day. The information and opinions you see on social media will vary greatly depending on what you're interested in, your beliefs, who you follow, and many other things, just to name a few. What we want to do is collect data from primarily two different groups of users on Twitter. Those two groups being the staff/faculty/administration of the University of Tennessee and those who are either students or show some sort of affiliation with the university.

We plan to collect data from these two groups from not only the COVID-19 time period, but from before as well. We will then take this information and compare and contrast the social media presence of these two groups. We primarily plan to analyze and compare the sentiment of the two groups, but could eventually move on to other textual analysis such as word frequency. Even further, we hope to potentially gather data from other universities and compare their metrics to those of The University of Tennessee's. applicable criteria that follow.

II. MOTIVATION

As the COVID-19 outbreak has brought significant changes and challenges to our society, we want to investigate how it has affected the community we are a part of by analyzing their social media presence. Social media presence has been increasing drastically as more people are staying home and going out less. We are interested in how the attitudes of the University of Tennessee community change before the outbreak and after. From there, we want to use Twitter to conduct our data collection and analysis because of how the platform provides a public forum for users to communicate openly to their audience/public. It is a perfect resource to analyze sentiment text from users' tweets as it provides various content. By collecting the data on this topic, we want to understand how our community is reacting to this situation and investigate to see if there are any interesting significant results.

III. DISCUSSION OF DATA COLLECTION

For data collection, we plan to use a web scraper built in Python. With this tool, we are able to easily gather the historical information we need, put it into a .csv file, and conduct some analysis from there. When we pull our data, each tweet contains the following fields: date (when the Tweet was written), username (who wrote the tweet), to whom (who the Tweet may have replied to), replies (how many users replied to the Tweet), retweets (how many users retweeted the Tweet), favorites (how many users favorited the tweet), text (the text in the Tweet itself), mentions (who was mentioned in the Tweet), hashtag (Hashtags contained within the Tweet), ID (ID of the Tweet), and a link to the tweet itself. Despite the large quantity of data we are pulling in, we are primarily concerned with the username, date, and text fields. We will be using numerous Python libraries to form our analysis and generate our conclusions. Some of these libraries may include but are not limited to the following: Numpy, Pandas, Matplotlob, SciPy, and Tensorflow. This set of tools allows us to perform the research that we desire. Several of the analyses which we would like to complete are as follows:

- Sentiment comparison between group A: university staff, faculty, and administration and group B: university students and/or general population. Did their general attitude towards COVID-19 differ through the pandemic?
- For these same groups, what words did each group use the most?
- Did different universities and student groups have different overall opinions on the pandemic?

IV. RESPONSIBILITIES OF EACH MEMBER

We will both be responsible for our fair share of the project. As far as specific roles go, we will be rotating between roles like project manager, developer, etc. We plan to delegate the numerous tasks of this project depending on who has had experience with the individual components of the project. In the beginning, we will have several overlapping tasks, but as the timeline moves on and the work becomes more specific, tasks will be assigned accordingly based on aptitude.

V. TIMELINE

- Sprint 1: Establish data set sources, gather data, and begin parsing information. In addition to this, the necessary tools and libraries will be chosen and set up.

- Sprint 2: Begin work on sentiment analysis and determine the best way to visually represent our findings. This is where we should begin seeing correlations in our data.
- Sprint 3: Finalize sentiment analysis visualization and begin work on other analyses if possible (word frequencies, other universities, etc.)
- Sprint 4: Bring project to a close, refine models, and finalize conclusions. Complete final project report.

VI. Expected Outcome

We expect to see a rather dramatic difference when comparing Tweets between the groups that we have chosen. Specifically, we expect the sentiment of university staff, faculty, and administration to be more positive than the general population. On the other hand, we expect to see the student and general population of Knoxville to have a more negative reaction, where they might be more nervous or anxious of the situation. We also expect the community's overall attitude to be different from other universities due to their different values/reactions perhaps, especially handling the situation will affect the results. However, we cannot expect what words these groups will use to convey their reactions due to variance, but it will be one of the interesting aspects to investigate.