Twitter Sentiment Analysis

An analysis into the change in social media sentiment through the COVID-19 pandemic

Tucker Miles
Knoxville, Tennessee
tmiles7@vols.utk.edu

Vicki Tang
Knoxville, Tennessee
wph612@vols.utk.edu

***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 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 a 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. The first of these two groups consist of Twitter feeds from The University of Tennessee administration, as well as university-sanctioned accounts, such as departmental pages for the Tickle College of Engineering. The second group is that of people who are talking about the university but are not employed by the university, such as students.

We plan to collect data from these two groups from not only the COVID-19 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 analyzed and compared the sentiment of the two groups, and eventually moved on to other textual analysis such as word frequency.

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 surrounding The University of Tennessee community change before the outbreak and after. From there, we used 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 textual sentiment from users' tweets as it provides a wide array of 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 can 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. This tool is called GetOldTweets3.

We used GetOldTweets3 to collect the data for our first group of administration and university-sanctioned accounts and it worked perfectly for our purpose. Unfortunately, in early November, this tool stopped working correctly, and no fix seemed to be in progress. This forced us to pivot to another tool, known as snscrape, to continue with our data collection. We were able to build wrapper scripts around this command-line tool which enabled us to get mostly the same functionality we originally had with GetOldTweets3. We used this tool to successfully capture our data for group number two, the group of people Tweeting about The University of Tennessee, but are not employed.

Going into more detail about our data collection, one group was collected by pulling Tweets from specific users' pages, and the other group was collected via keyword argument search. The following are the accounts and keywords we used to build our datasets:

A. University Administration and University-Sanction Accounts (Account Search)

- @DondePlowman - Donde Plowman
- @randyboyd - Randy Boyd

- @tucarpenter - Tiffany Carpenter
- @UTIA_SVP - Tim Cross
- @KC4UTM - Keith Carver
- @utknoxville - UT Knoxville
- @utk_tce - UTK TCE
- @utkdos - Office of the Dean of Students
- @ut_admissions - UT Admissions
- @utk_asc - UTK Academic Success Center
- @UTKCEHHS - UTK CEHHS
- @utk_cfs - UTK CFS
- @UTKStudentLife - UTK Student Life
- @UTKCoAD UTK Arch + Design
- @UTKSOM - UT School of Music
- @tennalum - UT Knoxville Alumni
- @utknursing - UT College of Nursing
- @HaslamUT - Haslam Business

B. Non-University affiliated Tweets (Keyword Search)

- utk
- University of Tennessee Knoxville
- Knoxville Tennessee

For processing, most of the high-level bulk operations were done with scripts that we wrote ourselves to interact with the CSV files. For example, making the datasets exclusive from one another, as well as combining multiple CSVs into one. Once this bulk work was done, this data was processed and validated using Pandas and getting only the information we need, and removing any unnecessary or bad items.

IV. Methods

We used numerous powerful libraries throughout the project. To handle most of the data collection, formatting, and configuration, we heavily relied on Dataframes in the popular Python library known as Pandas. Moving on to our analysis algorithms, we used NLTK in Python to create a binary Naive Bayes Classifier. This is what allowed us to classify Tweets as either positive or negative. We also used word density analysis algorithms in NLTK, as well as their pre-built models for tokenization and normalization. Moving on, we used the vaderSentiment Python module to do VADER (Valence Aware Dictionary and Sentiment Reasoner) analysis. This allowed us to add a neutral rating to our Tweets, rather than only having a binary classification of just positive or negative. Along the way, when these algorithms were being used, we relied on MatPlotLib for our visualizations where there wasn't an immediately obvious visualization we wanted in the other mentioned libraries.

V. Responsibilities of Each Member

In the beginning, we delegated the numerous tasks of this project depending on who has had experience with the individual components of the project. Tucker was mostly focused on creating the scripts for data collection, and Vicki was mostly responsible for obtaining the data, as well as filtering it. As we progressed, we began to do mostly pair programming. Since there are only two of us on this team, it was very easy for us to plan to meet and program together. This was how most of the work was done throughout this project.

VI. Timeline

- Week of Sept. 28th - Determined a base collection for our Twitter user groups: University accounts and non-University accounts. We also began researching different tools for scraping data from Twitter and different methods for sentiment analysis.
- Week of Oct. 5th - Started to scrape tweets based on searched keywords for the non-University group and scrape tweets from specific users for the University group. We also started to use some sample data to test out different sentiment analysis tools.
- Week of Oct. 12th - Started to clean the sample data from the University group with various Python libraries (e.g. NLTK, VADER, etc.) to understand and test sentiment analysis methods.
- Week of Oct. 19th - Wrote scripts for tokenizing, normalizing, and removing noise from data. Then started to train a model for sentiment analysis and classifying if tweets are positive or negative.
- Week of Oct. 26th - Wrote scripts for further data manipulation and testing. We continued data collection for the respective groups and converted their tweets and info to CSV files.
- Week of Nov. 2nd - Began combining and filtering our collected data set into two CSV files and separated by their respective groups. We also classified the tweets and calculated their sentiment ratios for visualization.
- Week of Nov. 9th - Documented results and our process of the overall analysis on Jupyter Notebook. Further continued data analysis and data collection for visualization.
- Week of Nov. 16th - Fixed any loose ends with our analysis and finished up the documentation for presenting.

VII. 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.

VIII. Results

For our data set of administration and university-sanctioned accounts, the ratio of positive to negative tweets was very one sided in most months, leaning heavily on the positive side. In March, when many places began closing down and

switching up due to the pandemic, we saw the highest number of positive tweets out of all months. We propose this could be due to encouragement from campus leaders to stay strong when things were beginning to get rough, and to continue with the constant encouragement they tend to display. The negative tweets seem to be fairly constant throughout, but had peaks in the months of March and April, which was around the beginning of the pandemic.

A. University Administration and University-Sanction user group balance plot

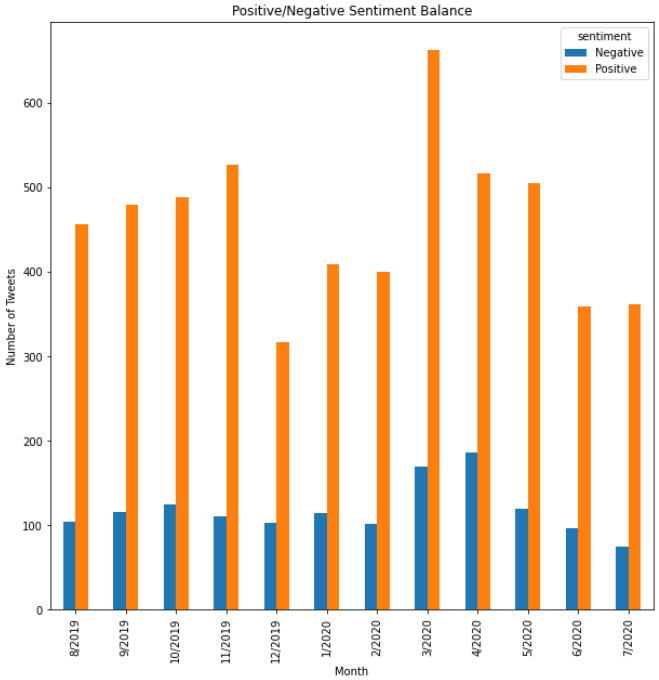

TABLE I
BALANCE PLOT - ADMINISTRATION/UNIVERSITY SANCTIONED

B. Non-University affiliated user group balance plot

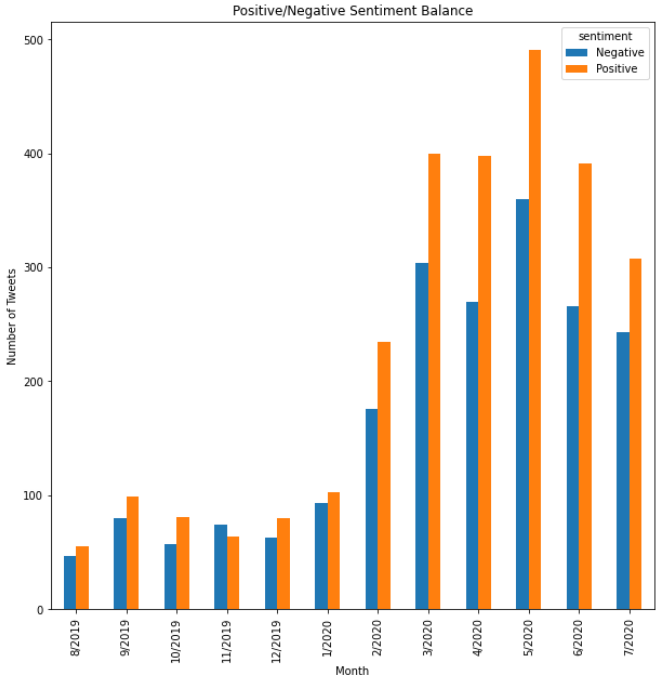

TABLE II
BALANCE PLOT - NON ADMINISTRATION/UNIVERSITY SANCTIONED

Moving on to our VADER analysis, the results showed no real significance, outside of the fact that the administration/university sanctioned dataset had a higher VADER compound rating for all months. In the below graphic, the blue line represents this dataset, while the orange line represents the non-University affiliated user group.

C. VADER analysis comparison

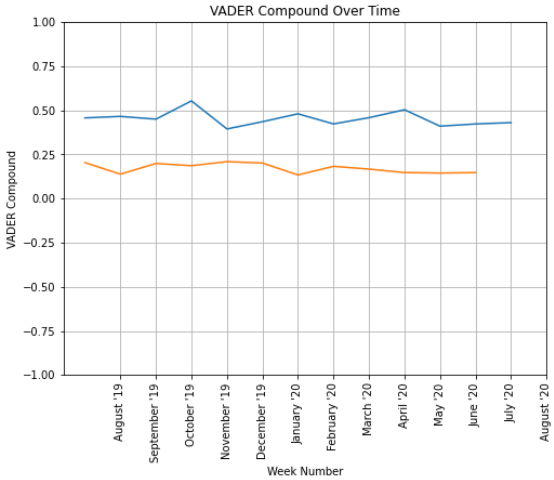

TABLE III
VADER COMPOUND OVER TIME

IX. Issues

Some of the issues we encountered when developing this project was looking for good specific keywords to do a query search for. The keywords have to be specific enough to get adequate results; otherwise, we had problems with retrieving results in foreign languages or off-topic. Also, there are a few other factors at play that could skew our analysis. For example, athletics is always a hot topic for social media, and could be considered separate from the data we were looking for. Another issue that we encountered was when we were starting out the project, we had some difficulties looking for a good, external Twitter retrieval API. We decided not to use Twitter's official API due to its limitations and constraints. However, we were able to find good resources such as GetOldTweets3 and snscrape to retrieve our datasets.

We also ran into an issue when originally modelling our data. As mentioned previously in this paper, we initially used a Naive Bayes Classification model to classify our Tweets as either positive or negative. This was a very large weakness in our analysis, as this model was forced to place a label on every single Tweet, even if it may be neutral. For example, a Tweet could be seemingly meaningless and neutral, but it was required for it to be labeled as either positive or negative. This led to issues with potentially generating misleading graphics and insights, and is what led us to using VADER analysis.

X. Future Work

After being able to analyze the sentiment of tweets within the community of UTK during the COVID-19 pandemic, we got to see interesting results of people's reactions towards it. It was a good sign to see that the University is being positive in their tweets to help uplift their community. After having the experience to analyze the sentiments of tweets within a community, we believe that this analysis can be used in other areas. One example is that we can see faculty's/students' reactions toward this new environment and system that the University implemented in having students learn virtually and access various resources on their