Project 2.3: CS415 DS Team Report on Sentiment Analysis of Twitter and Reddit Data

Chris Wetzel
Binghamton University
Binghamton, New York, United States
cwetzel2@binghamton.edu

Tyler Gabriel
Binghamton University
Binghamton, New York, United States
tgabrie2@binghamton.edu

Paul Maino
Binghamton University
Binghamton, New York, United States
pmaino1@binghamton.edu

ABSTRACT

In an attempt to analyze sentiment towards public figures on both Reddit and Twitter, data scrapers were implemented that searched for keywords on Reddit posts and tweets. The effect of social media and the osmosis of ideas throughout it have had a large impact on politics specifically, so many of the key figures our implementation focused on are U.S. politicians. Our methodology focused strongly on associating sentiment with each of these key figures. Also, we attempted to collect data in order to compare sentiment between Reddit and Twitter, with checks on the same keywords on the different websites. After storing large amounts of data during the height of the 2020 U.S. election, libraries and services such as the Google Natural Language API and VADER were used to analyze the sentiment of this data in bulk. This gave us average sentiment scores for different key public figures on both Twitter and Reddit. While we used both VADER and GCP data in this study, we found that VADER offered less accurate analysis, so GCP was utilized more. The results are also discussed and noted that Twitter sentiment scores fluctuate wildly, Reddit tends to dislike Donald Trump and Mike Pence, and that Donald Trump and Joe Biden dominate most of the social media discussion within the scope of this study.

1 INTRODUCTION

Methods of analyzing intent, tone, and positivity for online texts are extremely worthwhile considering the impact social media has on modern-world affairs. Data scraping methods allow for mass collection of opinionated and divisive tweets, so analyzing the sentiment of this data gives insights into viewpoints of entire populations of online users.

Common sense suggests that users will interact on different platforms for different reasons; factors such as website format affect a user's experience and cause them to form a preference. Therefore, understanding what types of users and their preference towards certain topics. Particularly of interest is measuring preference towards political topics, since social media is, by definition, a method of communicating ideas across social circles.

In our research, we aimed to observe how sentiment on social media changed towards public figures over time, namely political ones. We also attempted to compare sentiment towards certain figures between different user bases, namely Twitter and Reddit. Lastly, we also wished to observe political leanings of groups that are outwardly 'apolitical,' such as subreddits for topics that are not explicitly related to politics and government.

2 BACKGROUND/RELATED WORK

The study of political leanings and preferences has taken new meaning since the U.S.'s 2016 election, where 'alt-right' movements on 4chan had a surprising yet substantial effect on Donald Trump's public image. 4chan users' ability to create and distribute content to various other websites and provided a diffusion of ideologies that was not immediately noticeable to the average website browser [1]. Regardless of the corresponding political ideology, study and analysis of online preferences and spread of information becomes increasingly important as time progresses, since events such as this are not yet predictable or foreseeable. The effects of social media on the 2016 election were influential, but also were shocking at the time since nothing of the sort had occurred before, at least at this level of impact.

Diffusion of ideas between social media means necessitates an understanding of the userbases of different groups and their preferences. Reddit and Twitter have consistently been in the top ten most used social media platforms for several years [2], implying their influence won't be dissipating in the near future. Study of these sites' user base and their preferences offers clear benefits given the above.

3 DESCRIPTION OF DATA SETS

In the initial part of our implementation, we developed a crawler to aggregate Twitter and Reddit data. After running three separate programs, we obtained three collections: "compressed_tweets" ,"compressed_comments", and "compressed_posts". These posts represent zlib compressed data from Twitter, Reddit, and Reddit, respectively. At this point in the pipeline, the data is not processed at all.

After performing analysis on each set we have four major datasets: "tweets_gcp", "tweets_vader", "posts-vader", and "comments_vader". Each of these datasets contain documents with unique identifiers, their respective text, a sentiment score based on that text, a list of tagged key individuals, a created_at timestamp, and information related to their platform such as subreddit or hashtag. These four data sets provide a sufficient basis for analysing user sentiment across each platform.

As of 11:00pm on November 29th, 2020, the "compressed" collections contain 1710933, 53945, and 973 documents, respectively. The analyzed datasets contain 284557, 1652842, 49227, and 874 documents, respectively.

4 METHODOLOGY

The primary focus of our research is to perform sentiment analysis on social media data. While many different options exist in order to obtain this information for each document, the researchers agree that one powerful option is Google Cloud Platform's Natural Language API. Another alternative to this is VADER sentiment analysis libraries for Python. While GCP is likely a more accurate solution

Figure 1: Variation between GCP and VADER sentiment scores on 1000 tweets)

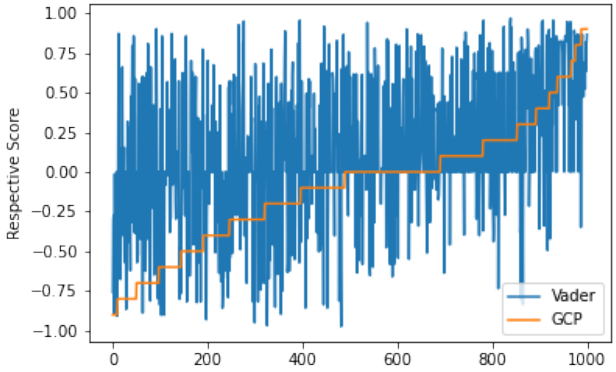
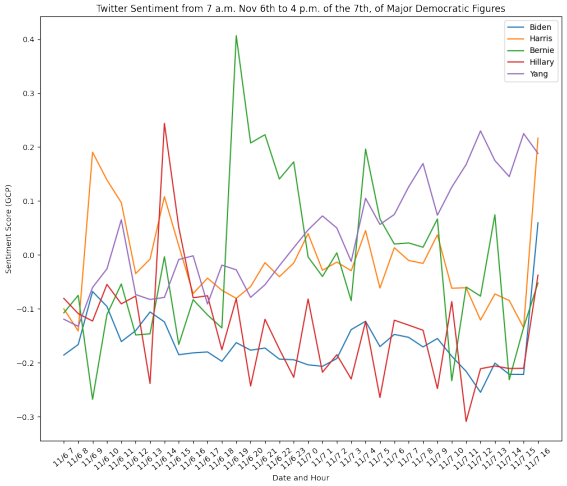


Figure 2: Democratic Individual Sentiment Scores from November 6th (7:00am) to November 7th (4:00pm)



due to its backing by Google, we do note that VADER is still viable for sentiment analysis and able to be used in our study. That being said, Figure 1 shows the variability between GCP and VADER on a small sample of (1000) tweets. For this reason, for our primary twitter-only analysis, we will use GCP for sentiment analysis and VADER elsewhere.

To answer our first research question we first aggregated tweets from the period of November 6th at 7am to November 7th at 4:00pm. We then piped these approximately 280k tweets into Google Cloud Platform's Natural Language API to obtain sentiment scores for each tweet text. We then grouped our key individuals into four groups (Democrats, Republicans, World Leaders, and Unaffiliated). The groups were as follows Joe Biden, Kamala Harris,

Figure 3: Republican Individual Sentiment Scores from November 6th (7:00am) to November 7th (4:00pm)

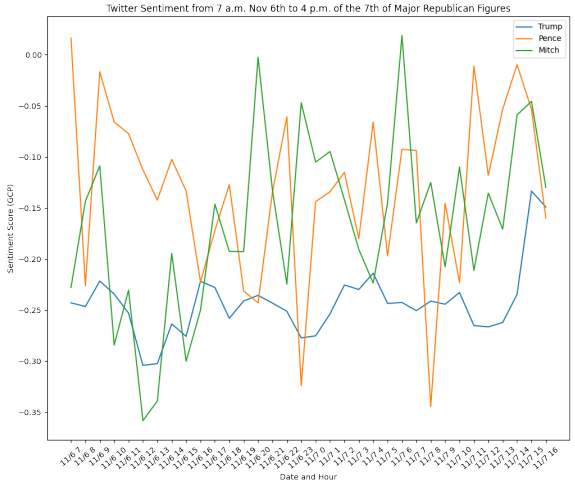
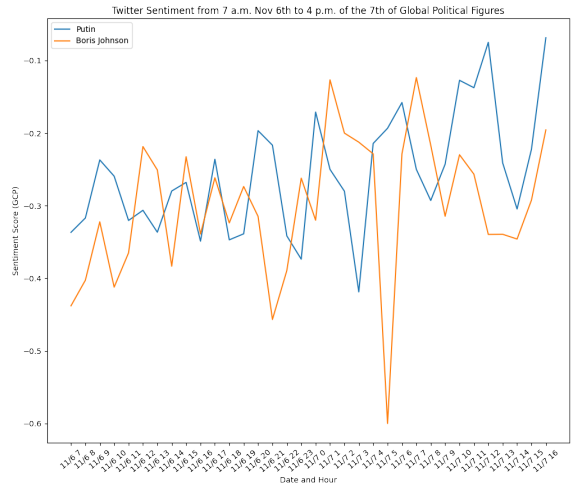


Figure 4: World Leaders Individual Sentiment Scores from November 6th (7:00am) to November 7th (4:00pm)



Bernie Sanders, Hillary Clinton, and Andrew Yang for Democrats, Donald Trump, Mike Pence, and Mitch McConnell for Republicans, Vladimir Putin and Boris Johnson for World leaders, and Elon Musk, Jeff Bezos, Anthony Fauci, Mark Zuckerberg, Kanye West, and Jeffrey Epstein for Unaffiliated. We filtered out any tweets with a GCP confidence level of below .25. These results were then graphed in Figures 2,3,4,5 for each respective group.

Figure 5: Unaffiliated Individual Sentiment Scores from November 6th (7:00am) to November 7th (4:00pm)

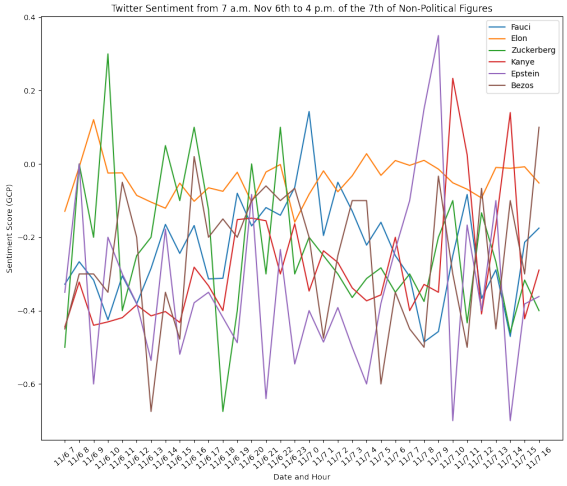
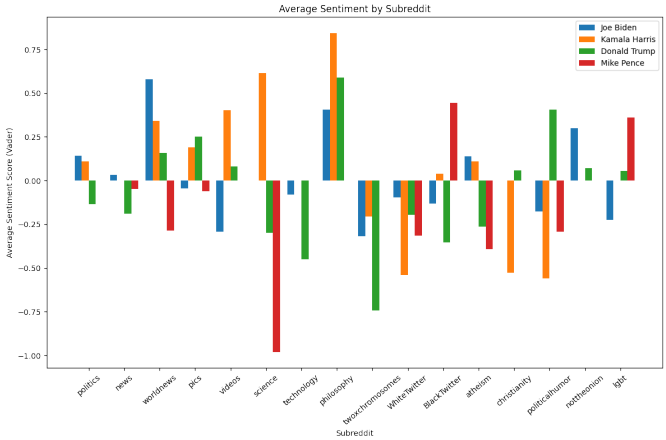


Figure 6: Average Reddit Comment and Post VADER Sentiment Scores by Subreddit and Individual)



To answer our second research question, we decided to use VADER sentiment analysis on Reddit posts and comments previously aggregated. After performing this on all of our posts and comments, we selected four key individuals (Donald Trump, Joe Biden, Kamala Harris, and Mike Pence) to measure bias in various subreddits. These results were calculated by computing the average sentiment score of comments and posts in said subreddits and matching them to certain individuals if the entry text contained their name. The results of these calculations can be seen in Figure 6, which also allows us to compare not only said individuals in a subreddit, but also between subreddits.

To answer our final research question, we again decided to use VADER sentiment analysis on tweets, Reddit posts, and Reddit comments. Although GCP sentiment analysis was performed for tweets in our first research question, we did not want to compare two different sentiment analysis models between two different types of data. This process was completed by checking each type of document for a given individual's name and recording the sentiment analysis score, which was summed for each respective type and averaged across all of said individuals data. The results of this analysis can be seen in Figure 7. This figure allows us to accurately compare general sentiment for said figures between Reddit and Twitter.

Figure 7: Twitter and Reddit VADER Sentiment Scores and Statistics by Individual

Name	Average Twitter Sentiment	Average Reddit Sentiment	Number of Tweets	Number of Posts	Number of Comments	Total
Donald Trump	0.024115	-0.028956	1018573	594	29131	1048298
Joe Biden	0.126847	0.074727	509927	277	9847	520051
Mao Zedong	0.316925	0.427048	83523	1	1000	84524
Kamala Harris	0.158677	0.135791	82340	9	878	83227
Elon Musk	0.096703	-0.113324	36780	8	859	37647
Hillary Clinton	0.090088	-0.026662	21601	5	1093	22699
Bernie Sanders	0.175402	0.157604	21806	19	708	22533
Andrew Yang	0.220756	0.319357	22030	1	181	22212
Mitch Mcconell	0.049901	-0.012084	20089	15	1168	21272
Mike Pence	0.096894	0.012785	14141	3	542	14686
Boris Johnson	0.022187	0.145796	12382	0	24	12406
Kanye West	0.052168	0.058053	8765	0	57	8822
Vladimir Putin	-0.006590	-0.069444	7791	1	231	8023
Adolf Hitler	-0.084267	-0.097957	4106	1	477	4584
Anthony Fauci	-0.065660	0.109733	4060	7	312	4379
Jesus Christ	0.128972	0.041332	3049	2	181	3232
Jeffrey Epstein	-0.059963	-0.099384	2389	1	141	2531
Joseph Stalin	0.055651	-0.094206	1995	0	65	2060
Jeff Bezos	0.324049	0.027181	1605	0	54	1659
Mark Zuckerberg	0.004326	-0.037273	1459	0	37	1496
Tom Cruise	0.092755	-0.015640	115	0	5	120
Jeremy Blackburn	0.000000	0.000000	0	0	0	0

5 TWITTER TIMESTAMPS

While we have collected over 48 million timestamps, due to issues with the campus VPN and our VM, we were unable to provided either of the specified ranges. Thus we have provided three plots Figures 8,9, and 10 representing November 5th to November 11th. These plots are divided due to memory limitations while rending the timestamps.

6 DISCUSSION AND CONCLUSION

In our first research question, we can see that each individual regardless of group tends to fluctuate wildly before the election is called. One interesting insight is that Donald Trump's average sentiment score rises before the election is called.

In our second research question we can see that most subreddits tend to view Donald Trump and Mike Pence in a negative light, as noted by the low sentiment scores. Oppositely, Joe Biden and Kamala Harris are viewed very favorably. Notably, there are a few subreddits that have no data on at least one figure. In our third research question, we can see that Donald Trump and Joe Biden are the most commonly referenced figures and that Twitter is slightly positive to both while Reddit is negative to Trump and Positive to Biden.

A few points worth mentioning/improvements in our experiment are firstly greater sample size to increase accuracy, an increased budget and time for the more fine-tuned GCP sentiment analysis

Figure 8: Twitter Timestamps from November 5th to November 7th

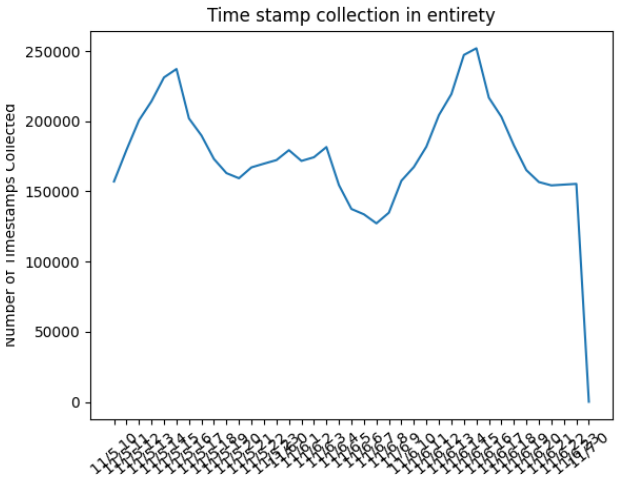
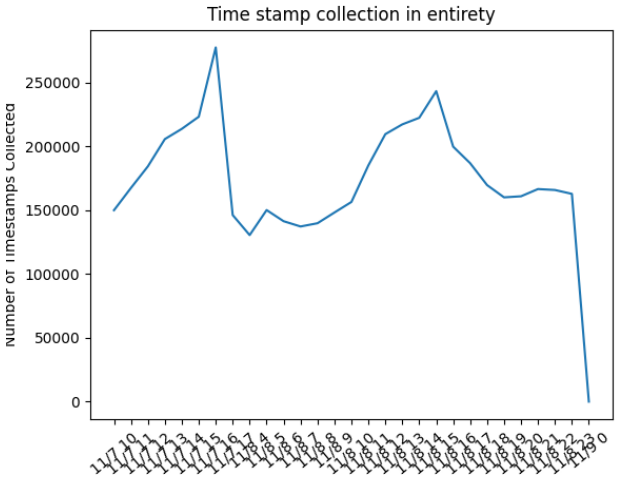


Figure 9: Twitter Timestamps from November 7th to November 9th

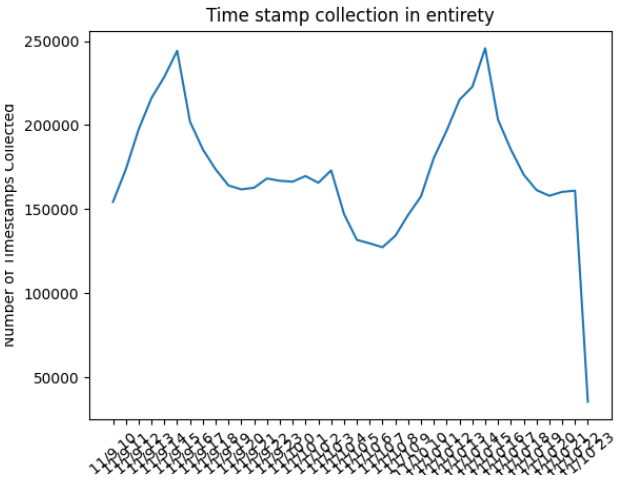


on all data, improved entity detection that does not rely on substrings, and an in depth preparation of the data to remove emojis, white space, and other conflating text. Overall, more resources will provide greater insight into the overall sentiments between Twitter and Reddit on various public figures.

7 REFERENCES

- 1. Gabriel Emile Hine et al. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. Proceedings of the 11th International AAAI Conference on Web and Social Media (2017).

Figure 10: Twitter Timestamps from November 9th to November 11th



- 2. Lilach Bullock, Gabrielle Wright, and James Gurd. 2020. Global social media research summary August 2020. (October 2020). Retrieved November 29, 2020 from https://www.smartinsights.com/so media-marketing/social-media-strategy/new-global-social-media-research/
- 3. https://pymongo.readthedocs.io/en/stable/
- 4. https://matplotlib.org/
- 5. https://cloud.google.com/natural-language/docs/reference/libraries
- 6. https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/
- 7. https://docs.python.org/3/
- 8. https://docs.mongodb.com/