US Politics by Social Media

Project Checkpoint Report

Rachel Ng Min Yee
University of Colorado, Boulder
Rachel.Ng@colorado.edu

Xinyi Lu
University of Colorado, Boulder
xilu2783@colorado.edu

Anuragini Sinha
University of Colorado, Boulder
ansi3987@colorado.edu

Abstract

The US election is the biggest event in US politics. In recent years, political parties have been actively using social media to promote their campaigns and reach out to voters. With a younger voter share, social media sites such as Twitter and Reddit have become the place for discourse. In our project, we aim to find out the sentiments of Twitter users towards the elections, how it influences their decision and finally how political parties can improve their campaigning to gain greater voter share in the future.

We utilized natural language processing techniques to perform sentiment analysis and opinion analysis on a collection of tweets specifically related to the two prominent candidates from the 2020 US presidential elections, Donald Trump and Joe Biden. We were able to discover the topics that voters held close to their hearts during the 2020 US Presidential Elections. Additionally, we also found that consistency is key during the entire duration of the election period and gaining favor from the people nearing the end of the election period is not sufficient to swing the vote. With this, parties and their public relations teams should rethink their strategies and personal branding tactics for building a stronger campaign in the future.

1 Introduction

Social media platforms are widely used by citizens during the US election period to express their opinions. This project examines the sentiments of Twitter users towards the different political parties and how Tweets can affect the outcome of the US presidential elections. With the use of natural language processing techniques such as keyword extraction, text classification, topic modeling and opinion analysis, we aim to extract key information from social media data and study the political behavior of Twitter users. This will be helpful for election candidates to assess their odds of success based on previous case studies and to help them tailor their campaigns based on the sentiments on the ground.

2 Related Work

Multiple projects have explored the sentiments of tweets related to the US election. Ali et al. proved that removed tweets posted after the 2020 US Election Day sided with Joe Biden while those before Election Day were more favorable about Donald Trump. This was done by performing sentiment analysis on tweets that were subsequently removed from Twitter and comparing their results to accessible tweets and accounts across time [1]. In another research project, Caetano et al. discovered that political homophily level rises when there are close connections and similar speeches. This was accomplished by analyzing the tweets of users during the 2016 US elections and classifying sentiments into six groups: Trump supporter, Hillary supporter, whatever, positive, neutral and negative [2].

Finally, Ansari et al. found out the dominance of support for a single party on Twitter for the 2019 General Elections of India by using the Long Short Term Memory (LSTM) classification model to predict the sentiments and results of the elections. Ansari et al. also compares the results from the LSTM model with other machine learning models [3].

Our approach is different from past projects as we aim to perform opinion analysis using natural language processing techniques to answer key questions. Some questions we can possibly answer are: which groups of people have these opinions? How do opinions change over time? Subsequently, with insights from the data, we can find out the possible factors that led to Trump's large voter share despite the controversies surrounding him in the 2020 US Presidential Election.

3 Methodology

3.1 Dataset

The dataset we chose to work on is a collection of tweets containing #DonaldTrump, #Trump, #JoeBiden, #Biden hashtags from the 2020 US presidential election [4]. There are two subcollections in the dataset - one for the tweets related to Donald Trump and one for Joe Biden. With 970,919 and 776,886 rows respectively, there is a total of approximately 1.72 million rows of data in our dataset. Of all the tweets, 394,400 or roughly 20% of them were tweets from the US.

The data set has a total of 21 columns and a Data Dictionary describing the names and definitions of the columns of the dataset can be found in the appendix. The specific columns that we found interesting were columns created_at, tweet, likes, reteweet_count, user_followers_count, user_location, lat, long, city, country, continent, and state_code.

3.2 Tools

- Pandas
- Numpy
- Langdetect
- Re
- Matplotlib
- Seaborn
- Plotly
- Ast
- Wordcloud
- Nltk
- Textblob

3.3 Main tasks

With the above dataset, we hope to perform text classification, opinion analysis as well as a case study on Donald Trump and his relatively successful voter share despite his abrasive personality and the serious controversies surrounding him. These analyses can be further split into the following subtasks:

Analysis	Subtasks
Text Classification	<u>Keyword Extraction</u> • Which keywords are the most important/relevant in the tweet? <u>Sentiment Analysis</u> • Is the tweet a positive, negative or neutral one? • What identity is the tweet supporting / criticizing?
Opinion Analysis	<u>Temporal Analysis</u> • What are the predominant opinions over time? <u>Categorization of Opinions according to:</u> • State • Democratic vs Republican
Case Study on Donald Trump	• How did opinions change over time? • Did the timeline coincide with certain events? • How did Twitter specifically help him/prevent him from swinging favor? • What factors helped him garner his large voter share?

3.4 Data pre-processing

Data cleaning is the process of identifying and correcting errors, duplicated and incomplete data from a dataset. To start off, date columns (ie. 'created_at', 'user_join_date', 'collected_at') are converted to datetime types and numeric columns (ie. 'tweet_id', 'likes', 'retweet_count', 'user_id', 'user_followers_count') are converted to integer types. Next, rows that have NA in the 'country' columns are dropped as we intend to analyze tweets by country. We also standardized country namings by converting 'United States of America' to 'United States'. As we are more interested in finding out the sentiments of Americans towards the elections, we filtered for tweets from the US. Irrelevant columns like tweet_id, source, user_name, user_screen_name, user_description, user_join_date and collected_at were dropped to reduce the size of data.

Following this, we conducted text cleaning on the 'tweet' column of the dataset. Using the package langdetect, we detected the language of the tweets and filtered for english tweets to facilitate sentiment analysis of tweets. Next, hashtags, mentions, retweets, links, punctuation and numbers were removed and tweets were converted to lowercase. As stopwords do not carry much meaning in a sentence, they were removed as well. Finally, we converted tweets to tokens and to their base form using stemming and lemmatization functions from the nltk library.

3.5 Sampling

Sampling is a necessary step in this project as the combined dataset is too huge (1.72 million rows). We used two approaches to sampling for this dataset. First, we sampled the first 10,000 rows of raw data from each Trump/Biden dataset and obtained a clean dataset of 8169 rows. In our second approach, we looked into tweets posted before Election Day (Nov 3, 2020) and split them into 3 subsets. In each subset, we then sampled 10,000 rows of raw data from each Trump/Biden dataset. After data cleaning, we obtained subsets containing 6382, 5935, and 5310 rows respectively. For evaluation purposes, we looked into tweets posted after Election Day and sampled 20,000 rows of raw data from each Trump/Biden dataset. After data cleaning, the combined dataset had a total of 5200 rows.

3.6 Exploratory Data Analysis

3.6.1 Bubble Map

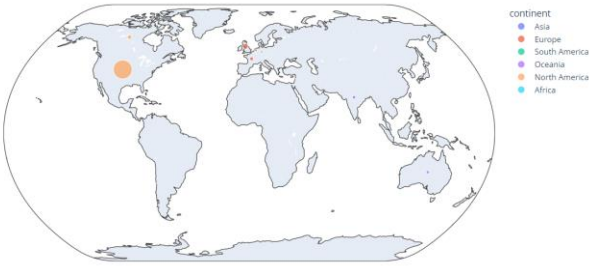


Fig 1: Distribution of tweets across continents

Fig 1 shows a bubble map of the distribution of sampled tweets across the world. As expected, the United States has the greatest number of tweets, followed by some in Europe and very little in the other continents. Since we want to analyze the sentiments of American voters who are mostly based in the United States, there will be less noise in our data and hopefully, this will give us better results.

3.6.2 Word Cloud

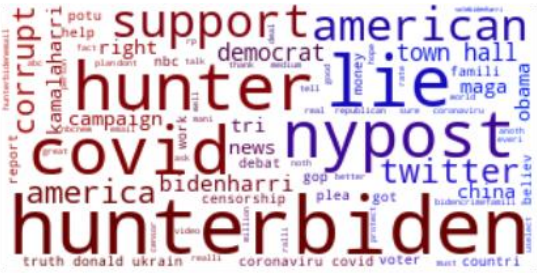


Fig 2**:** Overall Word Cloud of all Tweets

Fig 2 shows a world cloud for all the tweets in the sample dataset. This is helpful because it gives us an idea on the words that were most frequently used in discourse on twitter in the period leading up to the election.

From the overall word cloud, the most frequent words are “hunterbiden” and “hunter”. A cursory search shows this to be related to Joe Biden’s son, Hunter Biden’s tax crimes and false statement related to a gun purchase that came to rise during the period leading up to the 2020 US Presidential Elections. It is interesting that Joe Biden managed to win the election despite this being the trending topic regarding the elections on twitter. Perhaps, it is because people speak about it with sympathy towards Joe Biden. It could be useful to analyze how the media managed to separate Joe Biden from his son and diminish Biden’s culpability in this issue in the public eye.

Another big word is “lie”. Lies were a big thing in the 2020 US Presidential Elections with both candidates having their fair share of scandals. It would be interesting to analyze how their respective scandals affected voter perception.

“Covid” was another big word, not surprising due to the fact that the elections occurred in the middle of the pandemic. It would be insightful to see if covid did in fact play a part in Trump’s loss because of the way that he handled the pandemic in his term and the election period.

Fig 3: Word Cloud for Sample 1

Fig 4: Word Cloud for Sample 2

Fig 5: Word Cloud for Sample 3

Words like covid still remained prominent across the samples, but other words like the debate and american stood out more in some

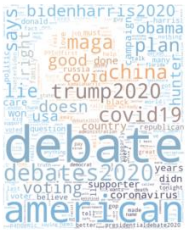
samples. We can also see the effectiveness of Trump's branding as his slogan “Make America Great Again” shows up quite strongly across all samples.

Fig 6: Word Cloud for Trump

Fig 7: Word Cloud for Biden

Fig 6 and 7 show word clouds for each of the candidates. This is useful in helping us compare between the topics of discourse for both candidates. Using “covid” as an example, we can see that those who talked about Biden cared a lot more about the coronavirus than those who talked about Trump, something to think about for each party as they move forward and tackle more health crises. “Maga” is also very prominent on Joe Biden’s word cloud, highlighting the success of Trump and his branding. It might be helpful for the Democratic party to look towards their own branding to help them in future elections.

Fig 8, 9 and 10: Word Cloud for Trump from Sample 1, 2 and 3 respectively



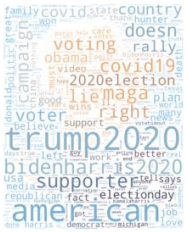


Fig 11, 12 and 13: Word Cloud for Biden from Sample 1, 2 and 3 respectively

Overall, these word clouds help give a big picture of the topics of discourse on Twitter in relation to the 2020 US Presidential Elections that we can focus on in our future analyses.

3.7 Sentiment Analysis

Sentiment analysis determines whether a text is positive, neutral, or negative based on its overall valence. The sentiment analysis is based on assessing text valence differently depending on the library or algorithm. These algorithms include the bag-of-words approach used by TextBlob and Vader algorithm, where the text is considered the sum of its constituent words, and valence or sentiment is calculated for words and combined to get a representative valence for the sentence, which informs if the text is positive, negative or neutral. Another approach is the word embedding-based model, where words are represented as vectors of numbers in an n-dimensional space This mapping from individual words to a continuous vector space can be generated through various methods: neural networks, dimensionality reduction, co-occurrence matrix.

We will use TextBlob and Vader models to analyze the sentiment of tweets for Biden and Trump. We will also use a deep learning model such as Flair or an alternative approach that uses aspect-based sentiment analysis. Flair requires a lot of computing power, so we may have to limit the sample size we can analyze.

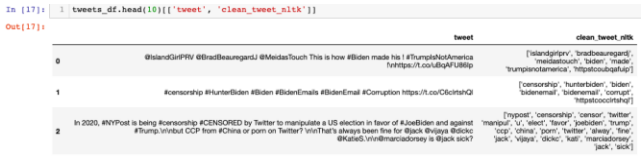
We add a column for data sources to identify tweets from Biden and Trump. This will allow us to see the sentiments of tweets from their supporters.

All three models, TextBlob, Vader, and Flair output a continuous number between -1 and 1. For our study, we will use classification and will convert these numerical values into categories. For example, we can classify them as negative if the score is less than 0, positive if the score is greater than 0, and neutral if the score is 0. These buckets for categorization can be changed as needed for analysis.

- TextBlob

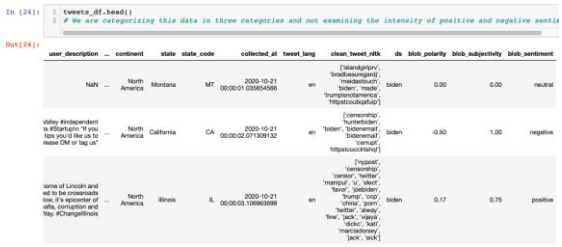
TextBlob is a simple model that calculates the polarity and subjectivity of a tweet. Subjectivity estimates how factual versus opinionated a text is and polarity estimates how positive, negative or neutral is the text. We are using a sample of the total dataset for

the computation. Clean tweet is a tokenization of the actual tweet. We started with 10,000 tweets for each candidate and cleaned them to arrive at this sample of data for analysis. A larger cleaned sample can be tried to see if the results vary a lot.



	tweet	clean_tweet_nltk
0	@IslandGirlPRV @BradBeauregardJ @MeidasTouch This is how #Biden made his f #TrumpIsNotAmerica h/thttps://t.co/uBqAFUB6lp	['islandgirlprv', 'bradbeauregardj', 'meidastouch', 'biden', 'made', 'trumpisnotamerica', 'httpstcoubqafulp']
1	#censorship #HunterBiden #Biden #BidenEmails #BidenEmail #Corruption https://t.co/C6cirtshQl	['censorship', 'hunterbiden', 'biden', 'bidenemail', 'bidenemail', 'corrupt', 'httpstcoccirtshql']
2	In 2020, #NYPost is being #censorship #CENSORED by Twitter to manipulate a US election in favor of #JoeBiden and against #Trump.\n\nbut CCP from #China or porn on Twitter? \n\nThat's always been fine for @jack @vijaya @dickc @KatieB.\n\n@marciadorsey is @jack sick?	['nypost', 'censorship', 'censor', 'twitter', 'manipul', 'u', 'elect', 'favor', 'joebiden', 'trump', 'ccp', 'china', 'porn', 'twitter', 'alway', 'fine', 'jack', 'vijaya', 'dickc', 'kati', 'marciadorsey', 'jack', 'sick']

Fig 14: Screenshot of cleaned tweets after using nltk



	user_description	...	continent	state	state_code	collected_at	tweet_lang	clean_tweet_nltk	ds	blob_polarity	blob_subjectivity	blob_sentiment
	NaN	...	North America	Montana	MT	2020-10-21 00:00:01.035654566	en	['islandgirlprv', 'bradbeauregardj', 'meidastouch', 'biden', 'made', 'trumpisnotamerica', 'httpstcoubqafulp']	biden	0.00	0.00	neutral
	Valley #independent is #Startuptin "If you tips you'd like us to lease DM or tag us"	...	North America	California	CA	2020-10-21 00:00:02.071309132	en	['censorship', 'hunterbiden', 'biden', 'bidenemail', 'bidenemail', 'corrupt', 'httpstcoccirtshql']	biden	-0.50	1.00	negative
	some of Lincoln and ed to be crossroads tow, it's epicenter of ulla, corruption and Way. #Changeillinois	...	North America	Illinois	IL	2020-10-21 00:00:03.106963698	en	['nypost', 'censorship', 'censor', 'twitter', 'manipul', 'u', 'elect', 'favor', 'joebiden', 'trump', 'ccp', 'china', 'porn', 'twitter', 'alway', 'fine', 'jack', 'vijaya', 'dickc', 'kati', 'marciadorsey', 'jack', 'sick']	biden	0.17	0.75	positive

Fig 15: Screenshot of data after applying sentiment analysis

From Fig 15, we also add a column named blob_sentiment, where we classify the numerical polarity values into categories named positive, negative, and neutral.

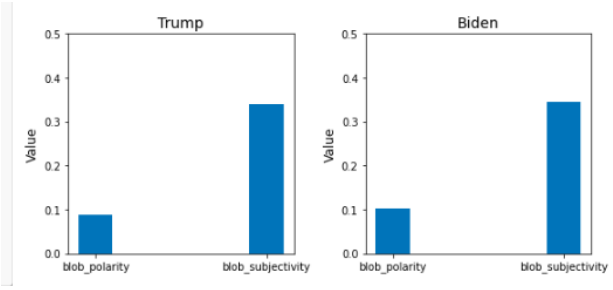


Fig 16: Tweet polarity for Trump and Biden

We can now plot the polarity and subjectivity of tweets in two ways: sentiment expressed per tweet and sentiment expressed per user. In Fig 17, we will compute an average polarity per candidate on a per-tweet basis. The limitation of this approach is that it does not handle spams. Imagine we have 1 user who tweeted 99 times, each having polarity -1 (opposer) and 1 user who tweeted once with polarity 1 (supporter). If we average across all tweets, we obtain -0.98. inferring support / opposing for the candidate would be limited in such a case as we are computing average sentiment per tweet.

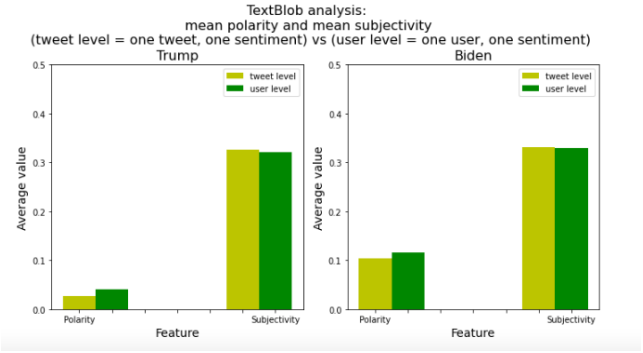


Fig 17: Polarity and subjectivity for Trump and Biden

Fig 17 shows the mean polarity and subjectivity for Biden and Trump. These are average values, so for Trump, the average polarity is a small positive number, whereas for Biden, the polarity is a slightly greater number for both per tweet and user basis. Biden supporters are tweeting more positive sentiments than Trump supporters.

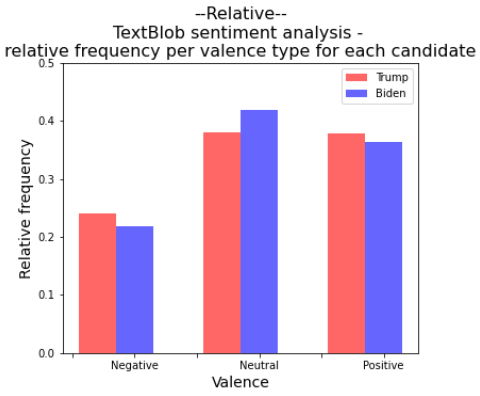


Fig 18: Relative frequency of sentiments for Trump and Biden

Fig 18 shows the relative frequency of sentiments and it is similar to the one in Fig 17.

- VADER

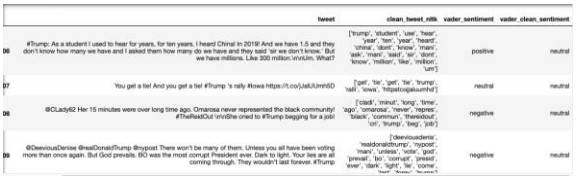
VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media and works well on texts from other domains. The valence for the words in the dictionary was empirically validated by multiple human judges who are familiar with this space of tweet and microblogging. It uses some heuristics to recognize word negations (“cool” versus “not cool”) and word intensifiers (“a bit sad” versus “really sad”) to understand the sentiment. The limitation is that it cannot distinguish spelling mistakes and will consider them out of vocabulary words.

Vader is a pre-trained model. Vader outputs something like this:

{'neg': 0.0, 'neu': 0.436, 'pos': 0.564, 'compound': 0.3802}

Negative, neutral and positive are scores between 0 and 1.

The compound value reflects the overall sentiment of the text. It's computed based on the values of negative, neutral, and positive. It ranges from -1 (maximum negativity) to 1 (maximum positivity). There is no standard way to interpret compound. One can decide that whatever is larger than 0 is positive and lower is negative, while 0 means neutral. But we can also decide to look only at more extreme values, like above or below +/- 0.8, for example.



	tweet	clean_tweet_nltk	vader_sentiment	vader_clean_sentiment
06	#Trump: As a student I used to hear for years, for ten years, I heard China! In 2019! And we have 1.5 and they don't know how many we have and I asked them how many do we have and they said 'sir we don't know.' But we have millions. Like 300 million.\n\nUm. What?	['trump', 'student', 'use', 'hear', 'year', 'ten', 'year', 'heard', 'china', 'dont', 'know', 'many', 'ask', 'many', 'said', 'sir', 'dont', 'know', 'million', 'like', 'million', 'um']	positive	neutral
07	You get a tie! And you get a tie! #Trump 's rally #Iowa https://t.co/jJaiUUmh5D	['get', 'tie', 'get', 'tie', 'trump', 'rally', 'iowa', 'httpstcojjaiuumhd']	neutral	neutral
08	@CLady62 Her 15 minutes were over long time ago. Omarosa never represented the black community! #TheReidOut \n\nShe cried to #Trump begging for a job!	['clock', 'minut', 'long', 'time', 'ago', 'omarosa', 'never', 'repres', 'black', 'commun', 'thereidout', 'cri', 'trump', 'beg', 'job']	negative	neutral
09	@DeeviousDenise @realDonaldTrump @nypost There won't be many of them. Unless you all have been voting more than once again. But God prevails. BO was the most corrupt President ever. Dark to light. Your lies are all coming through. They wouldn't last forever. #Trump	['deeviousdenis', 'realdonaldtrump', 'nypost', 'many', 'unless', 'vote', 'god', 'prevail', 'bo', 'corrupt', 'presid', 'ever', 'dark', 'light', 'lie', 'come', 'wouldnt', 'last', 'forever', 'trump']	negative	neutral

Fig 19: Screenshot of data after VADER

Vader recommends that data may be used uncleaned as cleaning strategies can introduce biases. For this analysis, we add a column in Fig 19 for clean and raw tweets to evaluate if they differ a lot.

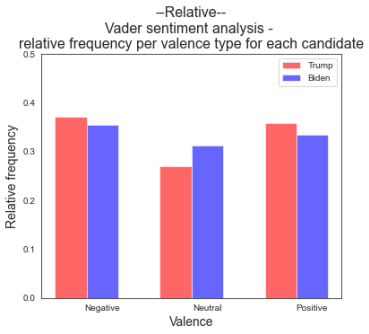


Fig 20: VADER sentiment analysis results

Fig 20 shows the VADER mode result, and it shows that a higher percentage of tweets for Trump is negative. On the other hand, Biden has a higher percentage of positive tweets. The broad patterns of results are similar to TextBlob. The results here show that both negative and positive trend for Trump is slightly greater than for Biden. This could be a result of using strongly positive and negative words in tweets as the Vader model is based on the analysis of the sentiment of each word which gets combined to get the sentiment of the sentence. Biden supporters are tweeting more neutral sentiment tweets based on this analysis.

- Comparison of TextBlob and VADER results

Our dataset is not labeled implying that we do not know by labels if a tweet is positive, negative, or neutral. So, there is no way for us to compare predictions to some 'ground truth'. We can, instead, compare each algorithm's predictions to the ones from the other two. This work remains to be done.

3.8 Opinion Analysis

Opinion analysis, similar to sentiment analysis, studies the polarity of opinions (positive, neutral, negative). In this project, we explored aspect-based sentiment analysis. In another approach to analyze sentiments, we utilized the aspect-based sentiment analysis algorithm. In this analysis, certain aspects are considered. From the word cloud, we select an aspect that was very well discussed on Twitter - covid. The aspect-based sentiment analysis will take certain aspects of the full text and score those individually. After a score is given to each part of that text, a total score is then calculated, indicating whether the selected tweet text has negative, positive, or neutral sentiment for that aspect.

One example of this can be seen in the following sentence: “The food was good, but the service was terrible.” To do this aspect-based analysis, the two aspects that we would analyze would be food and service. In this case, the food part of the text would receive a positive sentiment score. While the service part of the text would receive a negative sentiment score as it contains the word “terrible.” Therefore, the overall sentiment of this text would be negative. We applied the same logic when further analyzing the tweets made for both President Trump and Biden. When doing this analysis, we used the covid aspect for tweets from the 2020 election.

Aspect Model Used for Analysis

We used a publicly trained model to analyze tweets for selected criteria. Description can be found at the link. We installed the transformers library along with the SentencePiece tokenizer (which is needed by some models of the library, such as DeBERTa). The absa_model and absa_tokenizer to test the deberta-v3-base-abas-v1.1 pre-trained ABSA model.

Covid Aspect

```
In [14]: covid_tweets_df.head(10)[['clean_tweet_nltk','covid_negative', 'covid_neutral', 'covid_positive']]
```

Out[14]:

	clean_tweet_nltk	covid_negative	covid_neutral	covid_positive
0	use honor refuse covid test enter studio hell away biden	0.924700	0.044610	0.030689
1	shock remake unoriginality hollywood day chrisfarley blacksheep bidenharris joebiden joewilleadus vote bidensunitingus bidenwillcrushcovid	0.915277	0.070782	0.013961
2	joebiden kamalaharris morningjoe maddow nbc amyconeybarrett acb bidentownhall kamalaharrisvp scotusheairing mikepence republicansforbiden republican covid electionday ballotboxes	0.005586	0.989257	0.005157
3	diary radio junkie day wake news amyconeybarrett scotusheairings obamacare aca roevwade pardon supremecourt france macron germany merkel borisjohnson italy conte belgium europecovid coronavirus biden painting	0.010428	0.979176	0.010395
4	diary radio junkie day wake news amyconeybarrett scotusheairings obamacare aca roevwade pardon supremecourt france macron germany merkel borisjohnson italy conte belgium europecovid coronavirus biden	0.008896	0.981861	0.009244
5	diary radio junkie day wake news amyconeybarrett scotusheairings obamacare aca roevwade pardon supremecourt france macron germany merkel borisjohnson italy conte belgium europecovid coronavirus biden biden	0.022043	0.968914	0.009042
6	trump say covid spike bad like gov great job believe weak lie economy amp job crash handle covid whats life trump nothing natl test plan	0.884577	0.111239	0.004184
7	voting vote election election politics democracy votingmatters votingrights covid electionday registertovote ivoted govote voterregistration news civilrights votebymail voter government votingday joebiden america coronavirus voteblue music votebynight	0.021070	0.938268	0.040662

Fig 21: tweets with covid aspect analysis

clean_tweet_nltk	covid_negative	covid_neutral	covid_positive
use honor refuse covid test enter studio hell away biden	0.924700	0.044610	0.030689

Fig 22: Covid aspect tweet

This tweet has 92% negative sentiment for covid aspect. Looking at the text that uses words like “refuse”, “hell” etc., this makes sense.

In summary, aspect-based sentiment analysis is a powerful tool to analyze sentiments for a certain selected criteria. The analysis is mostly in line with the sentiments discovered using the tool.

4 Evaluation

Key results

We have chosen the VADER algorithm to compare the tweets from Trump and Biden supporters.

- Temporal Analysis

Temporal analysis studies the change in a variable over time. First, we observe the daily average polarity score for tweets containing Trump and Biden respectively.

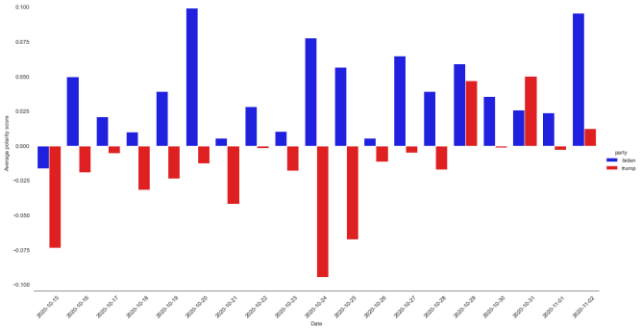


Fig 23: Distribution of daily average polarity score across time (pre-election)

From Fig 23, we can see that tweets containing Trump are mostly negative while tweets containing Biden are mostly negative. Other than this, we also observe some interesting insights. On October 24, there is the greatest difference in polarity scores. This could be due to the final presidential debate held on October 22 and news articles on the debate being published on October 23. In addition, on October 27, the White House published a press release naming the end of the COVID-19 pandemic as one of Trump’s top accomplishments. This could have resulted in the high polarity score for Trump on October 29 as Americans become more optimistic about Trump. Finally, on October 30, a Biden campaign bus was swarmed by Texas Trump supporters and Trump tweeted a video of the incident with the caption “I love Texas” the following day. This could have garnered attention and led to positive sentiments from Trump supporters, resulting in Trump’s polarity score exceeding Biden’s on October 31.

To evaluate our method, we used the same metric (ie. daily average polarity score) to examine the tweets after election day.

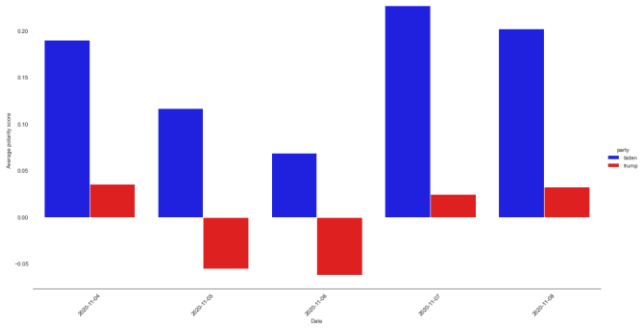


Fig 24: Distribution of daily average polarity score across time (post-election)

From Fig 24, we can tell that Biden tweets had the highest polarity score on November 7. This coincided with Biden’s projected victory in the election on the same day. In addition, there was a decrease in polarity scores for Trump in the days leading up to the announcement of the next President (Nov 4 to Nov 7), which could be due to Americans being pessimistic about Trump being re-elected.

Hence, temporal analysis is useful in monitoring the changes in polarity score as significant changes usually correspond to key events that occurred. This method can be utilized by the election team to improvise their campaign strategy and avoid events that can result in dips in sentiments.

- Categorization of opinions

We performed categorization of opinions based on the political party (Democrat vs Republican) and by state. In Fig 23 and 24 above, we can tell the average polarity score for each political party and conclude that Biden has more positive sentiments on Twitter during the election period.

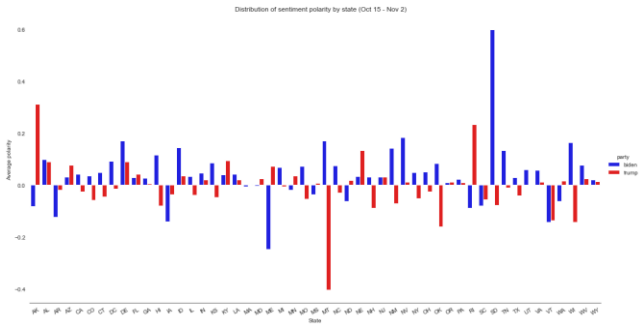


Fig 25: Distribution of average polarity by state pre-election

From Fig 25, for each state, there is usually one party that had a significantly higher average polarity score than the other. We concluded that for most states, the party which had a higher score was the winning party during the actual elections. Therefore, this is useful for political parties as they might want to focus on swing states where they do not have a clear advantage over their opponents yet. However, we acknowledge that there are discrepancies between our results and the actual election results,

and this could be because of sampling. The sample size may be too small, resulting in too few tweets from a particular state and consequently, biased, and inaccurate results.

- Case study on Donald Trump

There are four main questions to answer based on the results we obtained above.

1. How did opinions change over time?
2. Did the timeline coincide with certain events?
3. How did Twitter specifically help him/prevent him from swinging favor?
4. What factors helped him garner his large voter share?

For Trump, there were mostly negative sentiments from Twitter but switched to positive sentiments 5 days before the election period. However, we are not able to tell if this is due to more Americans supporting Trump or a result of Trump supporters tweeting more often. From our results, we could tell that the change in sentiments towards Trump coincided with events by the Trump party and hence, we can conclude that his strategies were useful in gaining confidence from Americans. This possibly resulted in a larger voter share for Trump in the actual elections. Overall, we hope that our project provided actionable insights into the 2020 US Presidential Elections by analyzing tweets from Americans.

5 Discussion

Key learning points

This project taught us how to use natural language processing (NLP) techniques to extract information from tweets and gain insights to help political parties in their election campaign strategies. We also explored sentiment analysis techniques extensively and understood the theory of aspect-based sentiment analysis. Moving forward, we are now equipped with the skills to perform text mining and apply data mining techniques in the real-world context. Apart from this, this project also taught us the importance of collaboration, communication and working under time pressure.

Challenges faced

Initially, we faced issues with text cleaning as we were unsure which other stopwords we should remove from our dataset. Fortunately, we were able to resolve this issue by coming up with word clouds to visualize the most common words mentioned. Moreover, we also had problems identifying topics from the topic modeling results and unfortunately, had to exclude our analysis from this report. Perhaps, we would have to understand the theory of topic modeling better to use it for our analysis.

Future works

As we could not identify the specific topics from our topic modeling section, we had to remove that section from our project. Hence, as future works, we can improve our analysis step through the use of n-grams, filter noun-type structures and optimize choice for number of topics through coherence measure.

In addition, we can also expand our opinion analysis to compare sentiments in the US vs outside of the US. Lastly, if given a device with better processing power or GPU, we can create larger samples and derive better results for our sentiment analysis and opinion analysis.

6 Conclusion

Our findings showed that sentiment analysis is a powerful tool for understanding voters' sentiments before and during the 2020 election. Sentiment analysis requires careful data preparation, exploration of data, word cloud building, analysis of data and results obtained. Certain key tasks, such as cleaning tweets using lemmatization and stemming, made the work of analysis streamlined as it removed data that were of low quality for sentiment analysis. Additionally, the exploratory data process was quite helpful, as we could utilize visualization of word clouds that were constructed to identify tweets of higher frequency and interest among users indicating engagement on those words. We observed that techniques such as sentiment analysis can be used to prioritize campaign expenditures. For example, if an area with mostly Biden or Trump supporters is experiencing low positive sentiments, more campaign funds could be invested to boost the positive scores. For our sentiment analysis, we used tools such as Textblob, VADER, and aspect-based sentiment analysis. While these methods are different, the analysis produces relatively consistent scores on sentiments, although there are some differences. The Textblob model calculates both the polarity and subjectivity of tweets, while polarity refers to how positive, negative, or neutral a tweet is. In contrast, subjectivity focuses on whether the tweet is an opinion or a fact, which was essential in helping us find which candidates were receiving positive, negative, or neutral sentiments from voters. Overall, we found VADER to be better than Textblob as it is based on a jury of people who specialize in those topics and have assigned a score to individual words. While aspect-based sentiment analysis was used to analyze certain aspects of tweets, whether they are positive, negative, or neutral. Furthermore, from our analysis, it was evident that voters who favored Biden tended to tweet positive or neutral sentiments regarding his policies and work. While for supporters of Donald Trump majority of the time, voters tweeted negative sentiments regarding his political work. This fact was also evident in our work using temporal analysis, which highlighted this as days toward the election date drew closer. The temporal analysis allowed us to monitor voters' sentiments just days before the election.

REFERENCES

[1] Ali, R. H., Pinto, G., Lawrie, E., & Linstead, E. J. (2022). A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-022-00633-z

[2] Caetano, J. A., Lima, H. S., Santos, M. F., & Marques-Neto, H. T. (2018). Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election. Journal of Internet Services and Applications, 9(1). https://doi.org/10.1186/s13174-018-0089-0

[3] Ansari, M. Z., Aziz, M. B., Siddiqui, M. O., Mehra, H., & Singh, K. P. (2020). Analysis of political sentiment orientations on Twitter. Procedia Computer Science, 167, 1821–1828. https://doi.org/10.1016/j.procs.2020.03.201

[4] Hui, M. (2020, November 9). *US election 2020 tweets*. Kaggle. Retrieved October 6, 2022, from https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

APPENDIX

Honor Code Pledge

On our honor, as University of Colorado Boulder students, we have neither given nor received unauthorized assistance.

Individual Contribution

Group Member	Work Done
Anu	• Sentiment Analysis with Texblob • Sentiment Analysis with VADER • Aspect Based Sentiment Analysis using Flair • Made graphs analyzing results for all 3 • Preparation and Delivery of Presentations • Writing of Report
Rachel	• Brainstorming and Ideation for the Project • Selection and Evaluation of Datasets • Exploratory Data Analysis (word cloud) • Topic Modeling (eventually left out of the report as the findings were less significant) • Preparation and Delivery of Presentations • Writing of Report
Xinyi	• Setting up GitHub repository • Data Cleaning • Sampling • Exploratory Data Analysis (bubble map) • Opinion Analysis • Preparation and Delivery of Presentations • Writing of Report

Data Dictionary

- created_at: Date and time of tweet creation
- tweet_id: Unique ID of the tweet
- tweet: Full tweet text
- likes: Number of likes
- retweet_count: Number of retweets
- source: Utility used to post tweet
- user_id: User ID of tweet creator
- user_name: Username of tweet creator
- user_screen_name: Screen name of tweet creator
- user_description: Description of self by tweet creator
- user_join_date: Join date of tweet creator
- user_followers_count: Followers count on tweet creator
- user_location: Location given on tweet creator's profile
- lat: Latitude parsed from user_location
- long: Longitude parsed from user_location
- city: City parsed from user_location
- country: Country parsed from user_location
- state: State parsed from user_location
- state_code: State code parsed from user_location
- collected_at: Date and time tweet data was mined from twitter