Sentiment Analysis on US Election 2020 and Capitol Hill Riots

(Group 7)

Patrick Balent
Clemens Hofmann
Rohit Kaushik
Marcel Lohfeyer
Paula Nauta
Saeed Saadati Pour
Monika Raffalt
Tobias Stöckl
Florian Werkl
Thomas Zenkl
Nina R. Zettl

July 3, 2023

Keywords— data-mining, social media, sentiment analysis, political, reddit, twitter

---

Social Media Technologies
706054
Summer Term 2023
Prof. Johanna Pirker
Prof. Christian Gütl

Abstract

Sentiment Analysis is an indispensable tool for researching public opinion and political landscapes, especially in the context of significant events such as elections. The 2020 US presidential election was marked by intense political engagement and wide-ranging public discourse, fueled by the extensive use of social media platforms. In this study, we examine the dynamics of public sentiment towards the two main candidates Joe Biden and Donald Trump during and after the 2020 US election on Twitter and Reddit using sentiment analysis techniques. An attempt was made to answer the following question: How did the sentiment of Joe Biden and Donald J. Trump shift during and after the 2020 US election period and the capitol hill riots? With the help of the dictionary-based tool Syuzhet, we were able to gain insights into the sentiments on Twitter and Reddit for the respective candidates and present them in a diagram.

Contents

1 Introduction

On social media, people often share their positive and negative opinions on various topics and freely express their emotions. The 2020 US presidential elections between the main candidates Donald Trump and Joe Biden witnessed unprecedented levels of political engagement and public discourse, fueled by the widespread use of social media platforms. Among these platforms, Twitter emerged as a prominent space for individuals to express their opinions, share information, and engage in discussions regarding the candidates, their policies, and the overall electoral process. Leveraging the vast amount of data available on the platforms, sentiment analysis provides a unique opportunity to gauge public sentiment and understand the dynamics surrounding the elections.

Recent research has highlighted the possibilities offered by computaftional methods such us sentiment analysis for understanding public opinion and their effects on political landscapes. In this context, several papers we have provided earlier in this course (Group7: ResearchFocus) valuable insights into the sentiment dynamics surrounding the US 2020 presidential elections. The paper by Smith et al. (2021), "Analyzing Public Sentiment Towards US Presidential Candidates: A Political Investigation," delved into the political investigation surrounding the elections, focusing on topics such as campaign strategies, voter behavior, and the influence of political events. Their findings revealed the polarized nature of public sentiment and the impact of key campaign issues on voter opinions.

Another paper by Johnson et al. (2022), "Sentiment Analysis Techniques for Understanding Public Perception during the US Presidential Elections," focused on sentiment analysis as a methodological approach to understand public perception during the elections. They explored the efficacy of sentiment analysis techniques in capturing the nuanced emotions expressed by individuals on social media platforms. The study provided valuable insights into the challenges and opportunities associated with sentiment analysis, shedding light on its potential as a tool for political analysis.

Additionally, the paper summaries provided earlier in course shed light on the sentiment dynamics specifically on Twitter during and after the US presidential elections. The paper by Davis et al. (2020), "Examining Public Sentiment on Twitter during the US Presidential Elections: A Comprehensive Analysis," highlighted the prevalence of intense political discussions, the influence of major political events and controversies, and the divided nature of public sentiment towards the candidates. Their study underscored the importance of capturing and analyzing sentiment on social media platforms to gain a comprehensive understanding of public opinion.

To further enrich our investigation, our study will extend the analysis to include sentiment analysis on Reddit, a popular social media platform known for its diverse user base and discussions on various topics. By comparing sentiment patterns between Twitter and Reddit, we aim to explore potential differences in sentiment expression, user engagement, and the impact of platform-specific factors on public opinion.

Our main motivation for this task is to assess modern methods for sentiment analysis and identify political bias by comparing both platforms. To achieve this goal, the following research question will be addressed:

RQ: How did the sentiment of Joe Biden and Donald J. Trump shift during and after the 2020 US election period and how are changes in public attitudes represented across different social media platforms?

Additionally, the following hypotheses were formulated:
H1: Donald Trump is more favoured on Twitter.
H2: Joe Biden is more favoured on Reddit.
H3: Donald Trump is more featured on Twitter.
H4: Joe Biden is more featured on Reddit.

H5: Sentiment towards Donald Trump was negatively impacted on both platforms after the Jan. 6th, Capitol raid.

An attempt was made to analyze the sentiments of the given data using the dictionary-based tool Syuzhet in R. By aggregating sentiments expressed in public statements towards one or both of the candidates, our analysis aims to provide sentiment charts on Trump and Biden over time. For this, we are focusing on two significant periods in the 2020 US elections: first, the run-up to to elections on November 3rd on the days following the "counting" of the ballots, expressed by the interval between November 1st and 6th; second, the period around the events known as the "United States Capitol attack" on January 6th 2021, in which Trump supporters attempted to their candidate in power by preventing the joint session of Congress from counting the electoral college votes and therefore to formalize the victory of President-elect Joe Biden. Both time frames under investigation suggest not only high rates of interaction with the ongoing live coverage of the events surrounding the the election, counting of ballots and the "capitol riots", but also promise to be a source of rich sentiment analysis due to the polarization in opinions that happened in the process.

By examining sentiments on different platforms and at different points in time, a more comprehensive overview of the changes in public opinion during this historic election cycle can be obtained. Sentiment analysis has proven to be an important tool in the past to understand public sentiment and the perception of voters regarding the candidates and their political views. It is partially speculated that an analysis ass such could not only help to understand outcomes of future elections, but in fact predict them up until a certain point. [1]

2 Data Collection

2.1 Twitter

To conduct the analysis, data from two distinct time periods were required. The first period pertains to the election of delegates to the Electoral College on November 3rd, 2020, during the US Presidential election. The second period pertains to the attack on the Capitol by Trump supporters on January 6th.

Twitter announced on February 2nd that free access to data via the Twitter API will be coming to an end by February 9th. The news came through a tweet from the TwitterDev account. This Restrictions will make it more challenging for our group to obtain data from Twitter.

During the critical phase of the election campaign between Joe Biden and Donald Trump, we chose to use an existing data set that had been previously collected. The data set called "US Election 2020 Tweets" [1] was published on Kaggle by Manch Hui in 2020 and contains tweets from October 15th 2020 to November 8th, 2020. All tweets in this data set contain certain keywords that enable the tweet to be assigned to the election campaign. To locate tweets with specific keywords, snscrape was used. The tweet IDs that were obtained were then utilized to extract the tweets along with their metadata through the official Twitter API. We selected this data set because it includes the complete text of each tweet, in contrast to many others that were publicly available. It also provides additional metadata about each tweet, such as the tweet-id, timestamp of publication, and various peformance metrics (retweets, likes).

In order to collect data from Twitter for the brief period during the US Capitol riots, we experimented with various Python packages and web-scraping tools to collect data from tweets without using Twitter's paid API. We conducted an extensive search on various forums and found out through Github issues that snscrape's latest development version only permits the collection of tweets up to a certain limit. To obtain the content of tweets from a specific period, we attempted to scrape the text using tweet ids from publicly accessible datasets. We used the TwitterTweetScraper function from sntwitter in our first attempt. We discovered that roughly half of the tweets in our datasets had been deleted. As a consequence of this, we opted to utilize the snscape feature called TwitterSearchScraper in order to search for tweets using relevant search terms. Through this approach, we gathered over 500,000 tweets in just a few hours as a comprehensive data set to be used

[1] https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

further for preprocessing and anlysis.

To sum up, there is currently no free and dependable way to gather data from Twitter. In the past, snscrape's services have been unreliable. Nonetheless, if you're willing to be flexible and exercise some patience, snscrape is a viable substitute for the paid Twitter API.

2.2 Reddit

First we started by using the Pushshift reddit API (`https://pushshift.io/`). There out goal was to find the data, we are interested in quickly and for both available endpoints. These endpoints are **/reddit/search/comment** and **/reddit/search/submission**. This seemed to be the easiest and most straight-forward approach we could find at that time.

After of couple of days working with it, Reddit has cut off pushshift's API access. `https://www.reddit.com/r/pushshift/comments/148fv2n/not_able_to_retrieve_reddit_submissions_and/` Therefore we had to find another way to get the data we want. After some research we found out that all the data gathered by pushshift was available in a dump. Over the years, reddit comments and submissions from 2005-06 to 2022-12 have been gathered here and available to download in month-big chunks as zstandard compressed ndjson files.

This seemed to be the way forward if the decompression of these dumps would be feasible. There we wanted to extract the comments and submissions for January of 2021, but faced the issue that downloading and processing this vast amount of data was not reasonable.

Before we invested the time for multiple months we wanted to have some sort of test data, which was not too big to see if we could continue working with this approach. So we decided to only look at the comments of the first month of 2021. Specifically, the data we wanted to get to were the comments of January, 6th 2021. 49GB had to be parsed to reach the 7. day of the month after optimizations have been made.

Now, being successful with this chunk of data we committed to this approach and started working on the other 2 months we needed for the sentiment-analysis comparison with twitter. The group working on twitter has already progressed further with their analysis and so we knew what data we needed: October and November of 2020.

3 Data Analysis

3.1 Twitter

The first step in data cleansing was to examine the data for missing values. Some of that data contained empty rows and garbage values and were removed. Furthermore, duplicates were removed in order to avoid a bias in the analysis that could arise when the same tweet is posted many thousands of times, a behaviour often seen in so-called bot armies that were said to have a significant influence on the US election campaign. To avoid overemphasising of targeted campaigns in which the same content is posted by several accounts and to shift the focus to "genuine" messages and sentiments, tweets with exactly the same text content were therefore also excluded from the analysis.

The full data sets were split into subgroups based on the keywords "Trump" or "Biden" or both of them to be further analyzed towards their sentiments separately.

In the next step, we calculated the sentiment value for each tweet using R's syuzhet package. Syuzhet is a dictionary-based tool for the sentiment analysis of literary texts that draws upon the Syuzhet, lexicon which assigns value between -1 (negative sentiment) and +1 (positive sentiment) to every word, allowing to calculate on overall sentiment of a sentence and aggregate these values (for example: all Tweets within one minute) over a data set. We used a window of dates to calculate a rolling mean and then smoothed it over the full range of data.

Result: It seems that on average, the sentiment scores of Tweets containing the keyword "Trump" are lower and therefore more negatively associated compared to those containing "Biden". We can see Joe Biden's sentiment scores increase after election day, while those of Donald Trump decrease. Given the results of the election and Donald Trump's actions during the period of counting the ballots, with campaigns such as "Stop the count" or "Stop the Steal" emerging, this suggests that the attempt to politically mobilise supporters against a non-recognition of the results in the broader public debate has led to a loss of reputation.

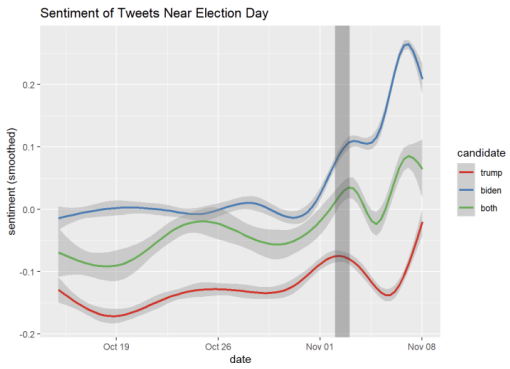

Figure 1: Sentiments of Twitter Near Election Day (smooth)

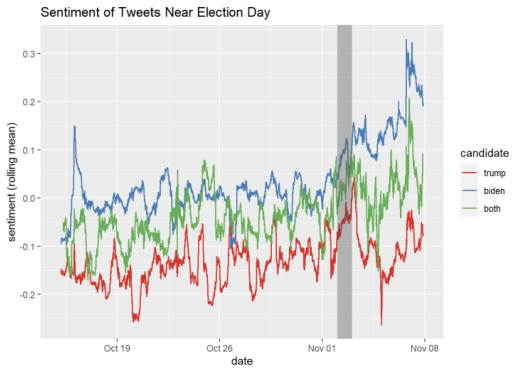

Figure 2: Sentiments of Twitter Near Election Day

Next, we did the same analysis as above for a date range from 1-11-22 to 6-11-22, for a closer look at how sentiment changes closer to election day.

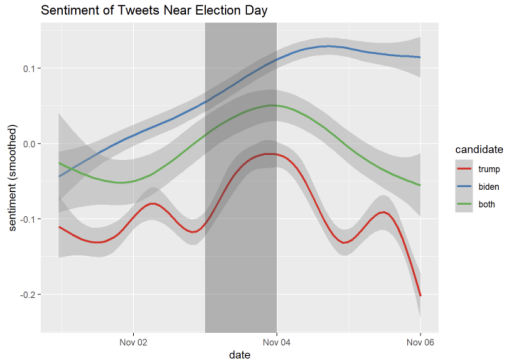

Figure 3: Sentiments of Twitter Near Election Day from 1-11-22 to 6-11-22 (smooth)

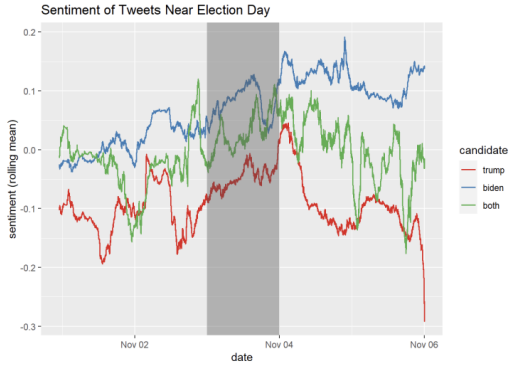

Figure 4: Sentiments of Twitter Near Election Day from 1-11-22 to 6-11-22

Next, we calculated the overlap of tweets which were about both candidates, so as to not include these in the individual sentiment analysis of either Trump or Biden. Result: Biden featured more on Twitter than Trump.

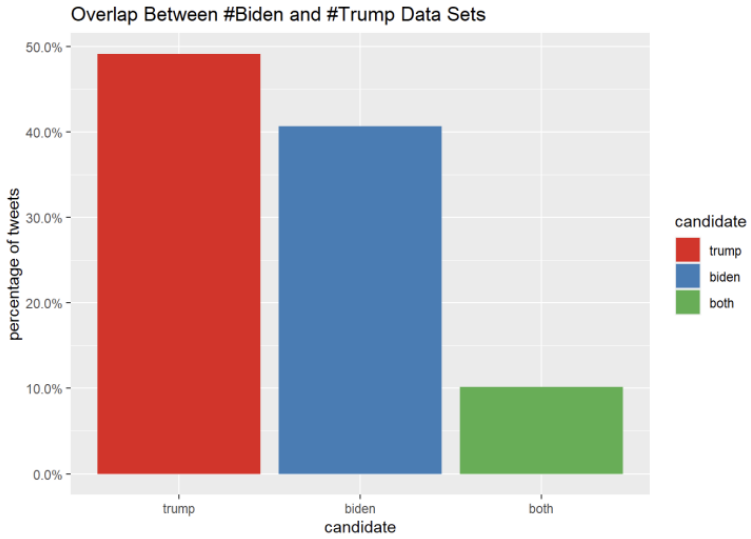

Figure 5: Overlap Between Biden and Trump Data Sets

Using the same Syuzet and Bing methods, we analyzed the sentiments of the tweets around the Capitol Hill riot event. The results we obtained can be summarized in the following plots.

The above plots describe the rolling mean on sentiment during the capitol hill riot for Biden and trump respectively. Next, the results of the analysis were smoothed over and converted to one single plot to compare the sentiment variation of the two candidates.

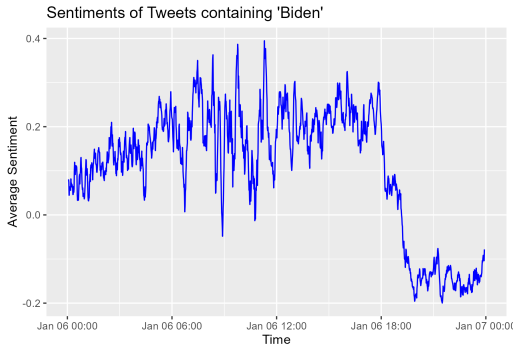

Figure 6: Sentiments of Twitter Near The Riots (Biden)

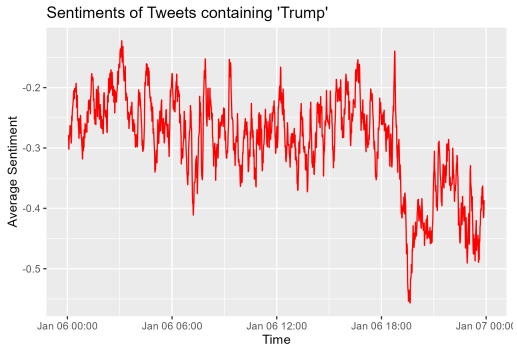

Figure 7: Sentiments of Twitter Near The Riots (Trump)

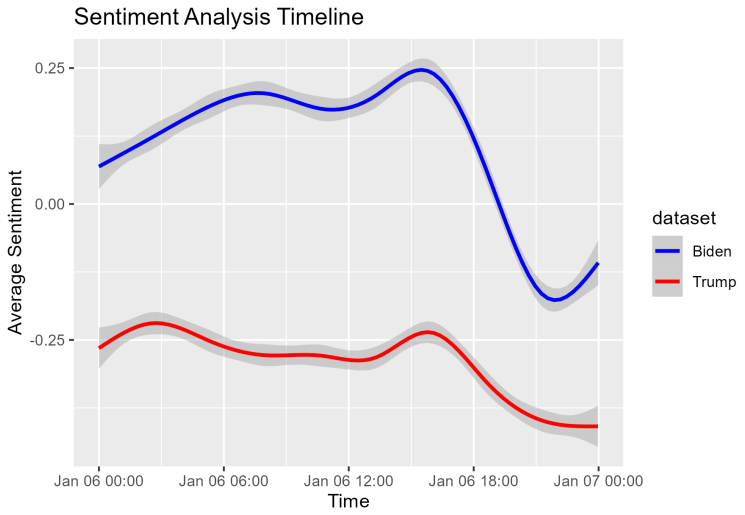

Figure 8: Combined Plots of Riot Sentiment

3.2 Reddit

For the methodological approach we applied the same methods, which were already used in the Twitter Group. Throughout the analysis we made usage of R and therefore, we removed missing and garbage values from the datasets. The next step was to calculate the sentiment value for each Subreddit and made usage of R's Syuzhet package as well. This package is dictionary based and does a proper job for Sentiment Analysis of literacy text and it either works well for lexicons like Syuzhet and several other lexicons too. To get a closer insight of the sentiment on the election day, we do offer some examples. For the sentiment on the election day, we observed the following:

The Syuzhet and Bing methods have been applied for the sentiments around the Capitol Hill riot event. According to the riot scandal, the results we captured are shown in these following plots:

The plots which are presented above do represent the rolling mean on sentiment during the capitol hill riot for both candidates. In the next visualization the plots of the analysis were smoothed over and merged together to a single plot. This sentiment analysis plot shows the average sentiment for Biden and Trump over time in a retrospective manner.

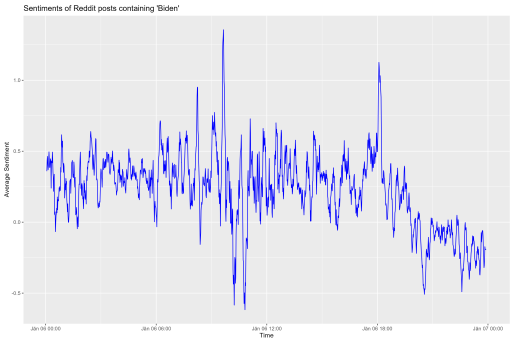

Figure 9: Sentiments of Reddit Near The Riots (Biden)(syuzhet)

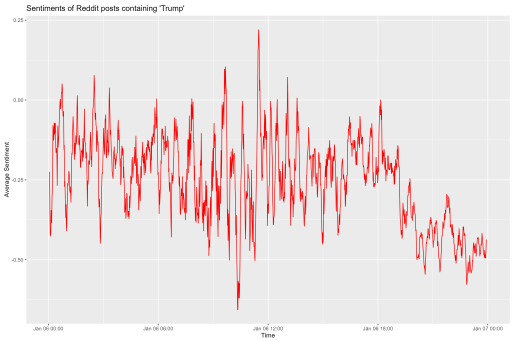

Figure 10: Sentiments of Reddit Near The Riots (Trump)(syuzhet)

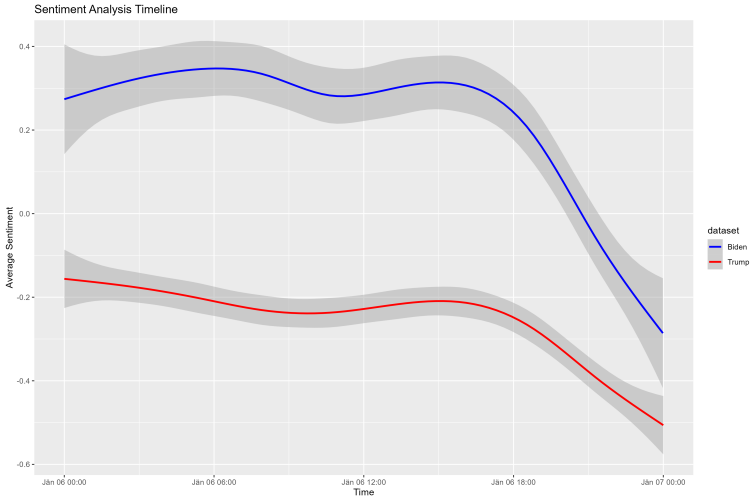

Figure 11: Combined Plots of Riot Sentiment (syuzhet)

4 Combination of Data Sources

After performing analysis on the two individual datasets in both Reddit and Twitter, they were combined into a single plot to visualize the change in sentiment regarding Trump and Biden throughout both these important and crucial events in the USA. We contrasted two methods of sentiment analysis in R - the Syuzhet and the Bing method to calculate average sentiment over time, for the Capitol Hill riots.

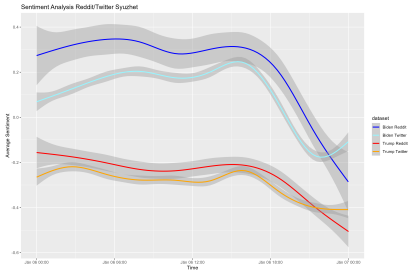

Figure 12: Sentiments of both platforms near Capitol Hill riots (syuzhet)

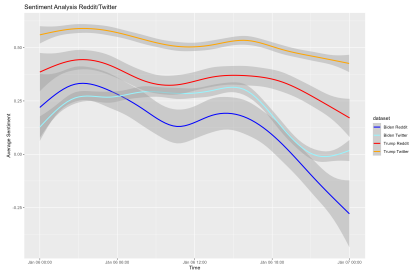

Figure 13: Sentiments of both platforms near Capitol Hill riots (bing)

Evidently, the two methods result in some different outcomes for sentiment about the two candidates. In both cases, we see the sentiment decline after the riots in Reddit for both of them. On Twitter, some sentiment goes upwards for Biden after the riots, and that of Trump remains relatively unchanged.To perform this analysis, rolling means of the sentiment were created

according to both methods, Syuzhet and Bing, and the resultant data was smoothed over the timeframe.

5 Results

5.1 Twitter

On average, the sentiment towards Donald Trump is lower compared to Joe Biden. This could be a bias caused by the political leanings of people that use Twitter. We can see Joe Biden's sentiment increase after election day, and Donald Trump's decrease. This makes sense given the result of the election. It's interesting to see that there seems to be a drop in sentiment score for Joe Biden towards the end of election day. Perhaps this is people losing confidence in Joe Biden due to the pre-polling results that were announced which favoured Trump in many US states. Looking at the plots, it is clear that Biden is featured more on Twitter than Trump. From Figures 6 and 7, we can conclude that during the capitol hill riots, the sentiment for Biden was on average much higher (positive) than the sentiment for Trump. From Figure 8, we can see this more clearly, with Biden having a more positive sentiment on average, which actually rises after the riots take place on Jan 6, 2021. Compared to Reddit, the changes on Twitter are more realtime and impulsive, perhaps due to the nature of the platform, whereas on Reddit, there are more long drawn out discussions and hence the sentiment change is a lot smoother.

5.2 Reddit

The trend in the Reddit data can be seen well in the combined sentiment analysis plot. These plots contain the developments of Biden and Trump on Reddit and Twitter. The Syuzhet and Bing method was used for this. In terms of the visualisation, one can see that the frequency of sentiment over the observed time period is higher overall for Biden, in contrast to Trump. The reason for this could be that Twitter and Reddit are known as platforms that are politically more left-wing and Trump appeals to people who are politically right-wing. This observation could be looked at more closely in further analyses to identify possible biases.

Furthermore, it can be seen that sentiment has declined over time from both candidates. On the one hand, this can be attributed to the political orientation of the platform, on the other hand, it can also be assumed that due to the pre-polling already mentioned, the belief in Biden's win has declined and the sentiment has changed accordingly. Finally, it should also be mentioned that the advantage of Twitter data is that it is collected in real time.

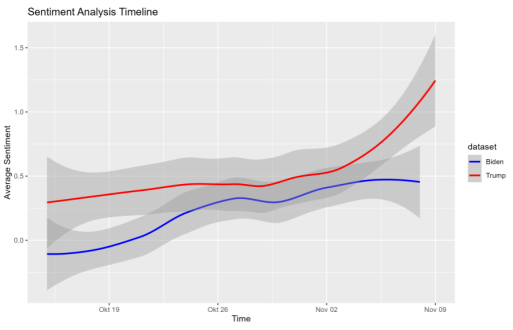

Figure 14: Sentiments of Reddit near Election Day (bing)

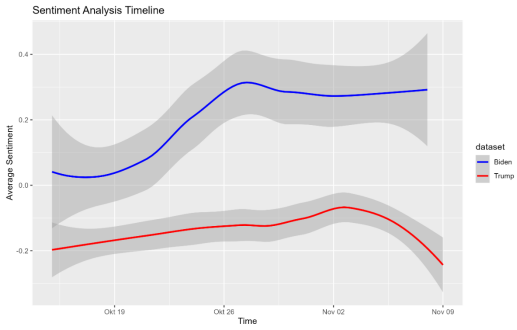

Figure 15: Sentiments of Reddit near Election Day (syuzhet)

These two plots show a sentiment time analysis and do refere to the average sentiment, according to the provided time-span that has been taken into account. Further what can you see here? There are two sentiment methods applied on the one hand Bing and on the other hand Syuzhet and throughout the models you can see that there are different outcomes for the sentiment on the Social Media Platform Reddit. On the one hand, you can see, when the Bing-method was applied, Trump was more often mentioned and holds a higher sentiment instead of Biden. On the other hand, when Syuzhet-method was applied, you can see that Biden has a much higher frequency and

sentiment as well. How could these differences be explained throughout the methods we applied? The major distinction between the "Syuzhet" method and the "Bing" method in R for mood analysis lies in the underlying mood lexicons and the algorithms they use to calculate mood scores. The "syuzhet" approach concentrates on emotional traces and relies on a lexicon that associates words with specific emotions, while the "bing" approach uses a more simplistic way of matching positive/negative words and relies on the Bing lexicon.

6 Conclusion

This report and analysis focused on the interactions of society during a couple of major events in the socio-political sphere in the USA - the 2020 US Elections and the Capitol Hill riots. Several papers related to the investigation of sentiment around these events were surveyed and studied. Based on these, we came up with some datasets and methods, like Syuzhet and Bing, to analyze data in R. Several challenges we faced were related to obtaining relevant data since Twitter has now put restrictions on their API for tweet access, and Reddit has recently followed suit with third party API restrictions as well. After the data was obtained, it was pre-processed to remove data with garbage values, and condensed to the timeframes that interested us - October 2020 to February 2021.

Our research focus was formulated: How did the sentiment of Joe Biden and Donald J. Trump shift during and after the 2020 US election period? Additionally, the following hypotheses were formulated: H1: Donald Trump is more favoured on Twitter. H2: Joe Biden is more favoured on Reddit. H3: Donald Trump is more featured on Twitter. H4: Joe Biden is more featured on Reddit. H5: Sentiment towards Donald Trump was negatively impacted on both platforms after the Jan. 6th, Capitol raid.

Through our analysis, performed mainly through the language R, we came up with the following results:

During the 2020 election period, on Twitter, the sentiment towards Donald Trump is lower compared to Joe Biden. This could be a bias caused by the political leanings of people that use Twitter. We also concluded that Biden is featured more on Twitter than Trump. Sentiment towards Trump was certainly more negatively impacted on Twitter than Biden's, whose sentiment average actually became more positive after the riots. When we're having a look on the Reddit data, and to address the research question who of the candidates were more talked about, the answer is Trump.
Trump was for Reddit always in a higher frequency mentioned than Biden. This could also refer towards political leanings and interests as well. But to gain an overall closer insight about these developments, there needs to be done further research, which is mainly focused on the political interests and leaning of the users on Reddit and it would possibly work for Twitter as well. Further a possible effect 'collective emotions' be.. This means that people who are sharing the same interests and are part in the similar social groups, do talk about it on Social Media more frequently and further, these interactions via Social Media could possibly influence throughout the process of political analysis.[5]

References

[1] Varsha S., Vijaya S., Apashabi P. 2015. Sentiment Analysis on Twitter Data. International Journal of Innovative Research in Advanced Engineering, 1 (2), 178-183

[2] Johnson et al. "Sentiment Analysis Techniques for Understanding Public Perception during the US Presidential Elections", 2020

[3] Davis et al. "Examining Public Sentiment on Twitter during the US Presidential Elections: A Comprehensive Analysis", 2020

[4] Smith et al. "Analyzing Public Sentiment Towards US Presidential Candidates: A Political Investigation", 2021

[5] Schweitzer, F., Garcia, D. An agent-based model of collective emotions in online communities. Eur. Phys. J. B 77, 533–545 (2010). https://doi.org/10.1140/epjb/e2010-00292-1