Comp 598 Final Project: An Investigation of User Engagement Regarding the candidates of the 2020 US election in the Subreddits /r/politics and /r/conservative

Marek Adamowicz 260738713 Romain Floreani 260847751

Sihan Wang 260767658

December 2020

1 Overview

The US presidential election, an event which occurred on November 3rd 2020, has been a polarizing topic of discussion across various social media platforms. Leading up to and following the election, the two primary candidates, incumbent president Donald Trump and democratic challenger Joseph (Joe) Biden, have been at the center of much of this discussion. In this paper, we analyzed the titles of Reddit posts from the subreddits /r/politics and /r/conservative in an attempt to determine the most salient topics encircling these two candidates. We then conducted an analysis of the words found in these post titles using a TF-IDF metric to find distinguishing vocabulary used by liberal and conservative voters regarding these topics. Our analysis shows that there were multiple differences in usage of language on both subreddits. The biggest concerns surrounding the US's past, present, and future did not always align, and their were opposing viewpoints regarding the legitimacy of the election, with conservative posts likely to push the claim of a fraudulent election as compared to liberal posts which held the opposite.

2 Data

Reddit is a website that provides an API for convenient access to its information. We were thus able to use the "requests" python library to obtain the required post data quickly and in a well-formatted way. Downloaded posts arrive in json format, in which we are given information pertaining to the author of the post, the urls it uses, its unique id and more.

Posts can be found in multiple different categories on Reddit. These categories are "hot", "new", "rising", "controversial", "top" and "gilded". We chose to collect our posts from "hot", a category displaying recent posts (usually only a few hours old) with high user interaction and positive approval by the community. We attempted to collect a total of 6000 posts found in "hot" from both the subreddits "r/politics" and "r/conservative", two popular sections on Reddit considered to be left leaning and right leaning respectively. The posts were collected from each subreddit in batches of 1000 over 3 consecutive days. We filtered the posts to obtain the title as well as its "upvote ratio", a metric which denotes the rate fraction of users who responded well to the post as opposed to those who didn't. As the posts were taken from "hot" on a popular subreddit, it is certain that the post received a sizeable number of votes from the community.

We encountered several issues in the post collection process. Firstly, the "hot" category saves only the 750 "hottest" posts. A call for more posts past this point returns posts that were previously downloaded (e.g. Asking for the 752nd "hottest" post will instead return the 2nd hottest post). This means that of the 6000 posts downloaded, over 25% were duplicates and would need to be removed. Secondly, posts that appear in the "hot" category on one day may also appear in the same category the next day. We felt that keeping duplicates in this manner could introduce bias into our results, so they would need to be discarded. Finally, a subreddit may also contain "stickied" posts, posts which moderators of the subreddit keep as the "hottest" post regardless of its score. We chose to keep these posts (with duplicates removed) as the content pertained to the current election. After addressing these issues, we were left with a total of 1000 unique posts which were collected over 3 days from each subreddit at 8p.m from Nov 18 to Nov 20 (on each day respectively we collected 333, 333, and 334 posts)

Next, we extracted the titles and upvote_ratios from the json objects. We converted the title to lower case and used regular expressions to determine if the titles contained the strings 'trump'

or 'biden'. We removed the titles that did not contain either of these terms. One drawback of this approach is that we could not keep posts referring to the two candidates only by their first names (Donald, Joe/Joseph) or through common nouns and pronouns (e.g. the president, he). Another drawback of this method is that it did not allow us to distinguish between the candidates family members who may also share their last names. Finally, we randomly selected 200 posts from each subreddit without duplicates, 1/3 from each day of our collection process, and saved the information in files using the csv format.

3 Methods

Multiple different techniques can be used to quantify the use of terms in our dataset. For the analysis of our Reddit post titles, we used the TF-IDF statistical measure. TF (term frequency), describes how common a word is in the data. For example, the word "in" is a common word that appears often in most texts while the word "report" does not. The former would obtain a high TF score while the latter would obtain a low score. IDF (inverse document frequency) describes how unique a word is to a document or category. For example, the word "the" is likely to appear in all texts and is thus not unique. If the word "report" appears often in a single category of words, it would obtain a high IDF score. The specific calculation of TF-IDF can be done in multiple different ways. The one chosen in our analysis is described as follows:

$$TFIDF(w) = \text{Freq}(w) log(\frac{|c|}{\text{Freq}(w|c)}$$

where

- w is a specific word found in the text
- c is the set of categories in which a post can be placed in
- Freq(w) is the count of word w in all of titles regardless of category.
- $|c|$ is the number of categories in the set
- Freq($w|c$) is the number of categories a word is found in

After multiple iterations of analysis, we decided to place the posts into exactly one of six categories: Political Opinion, Past situation in the United States, Current situation in the United States, Future of the United States, General election posts and Coronavirus.

We manually added 6 columns to the csv files; one for each of the categories. If the post belonged to a certain category, we coded a 'y', otherwise we coded an 'n'. The original intention was to assign titles to multiple categories, however we did not make use of this feature and instead each title was assigned to only one category. The computations of the TF-IDF scores were handled using data frame objects and python dictionaries.

4 Results

Below are the definitions of the categories as well as some examples to help with understanding (the examples below were manufactured)

Political Opinion

Def: The title contains the opinion of an individual or group about a political event. The entity expressing the opinion does not need to be explicitly named.

Examples:

- The US needs a national popular vote in order to prevent another Trump
- Majority of voters say special counsel should be convened to investigate Biden, Ukraine dealings

Past situation in the United States

Def: References events having taken place before the election. Sentences are in the past tense.

Examples:

- Trump ordered police to stop Black Lives Matter protests
- Report: Biden originally wanted Sanders as his VP
- Biden is past it (Incorrect category: "past it" is an informal way of saying "too old", it is not a period of time. This is a political opinion)

Current situation in the United States

Def: Events unfolding currently in the US that do not pertain to the election itself. Sentences are in the present tense.

Examples:

- The Trump Campaign Is Paying For A Recount In Wisconsin , But Only In Areas With Large Black Populations
- Federal Judge Blocks Trump Order That Has Already Led to 'Illegal' Expulsion of 13,000 Unaccompanied Children

Future of the United States

Def: Contains promises made my candidates and predictions regarding the fallout of the election. Sentences are in the future tense.

Examples:

- Trump wants to buy OAN
- Joe Biden plans to meet world leaders before inauguration
- Trump needs to think about the future of this country (Incorrect category: Though this does mention the future of the USA explicitly, this is not a promise or prediction of any kind. Rather, this is a political opinion)

General election posts

Def: An umbrella term for events which concern the 2020 US election. This includes references to voting, fraud, and announced election results.

Examples:

- Biden wins small Georgia county by 4 votes
- Video posted shows democrats burning hundreds of votes for Trump
- Trump plans to run for re-election in 2024 (Incorrect category: This is not about the 2020 general election. This should be placed in "future of the United States")

- Report: Gun sales have increased 460% after Biden elected winner. (Incorrect category: The main piece is the gun sales, not the election. This is a "current situation in the United States")

Coronavirus

Def: Titles which make reference to the coronavirus or any of its terms including "coronavirus", "the virus", "covid", "covid-19", "pandemic" etc.

Examples:

- Covid cases reach record high for third straight day under Trump administration
- Biden: "Coronavirus is very real"
- Americans are sick of Donald Trump (Incorrect category: The term sick does not refer to the coronavirus. This is a political opinion)

Note that these definitions do not cover all situations in a satisfactory way. For example, it can be argued that the following post could belong to several of the categories mentioned:

- "We should have started planning for this years ago!" Biden slams Trump on his handling of the coronavirus

This is the main reason why our original intention was to encode posts as belonging to more than one category. In such cases, we were forced to discuss amongst each other to determine the final choice of category. Perhaps the best solution would have been to provide more explicit definitions for each of the categories in order to avoid ambiguity.

Below we present several figures containing results obtained from our data analysis. The results will be discussed in further detail in the following section.

Figure 1

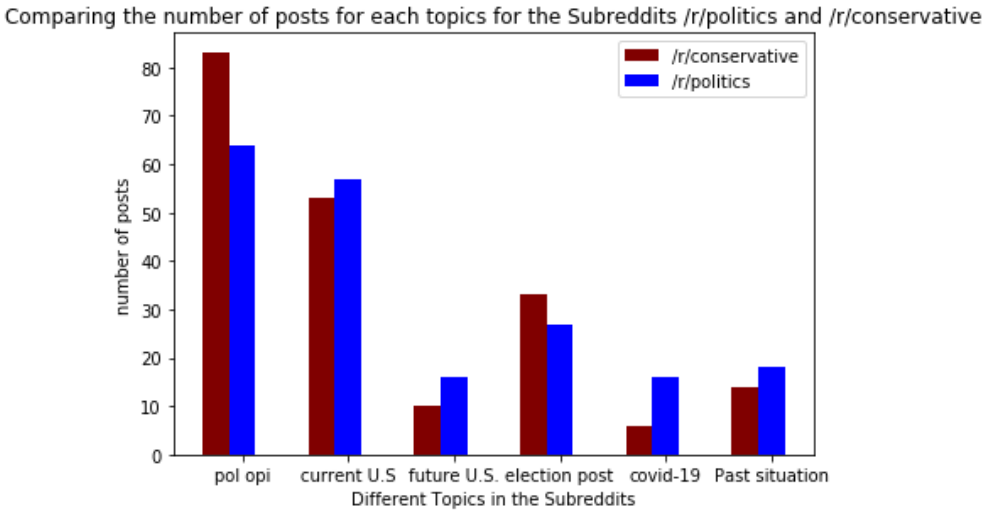


Figure 2

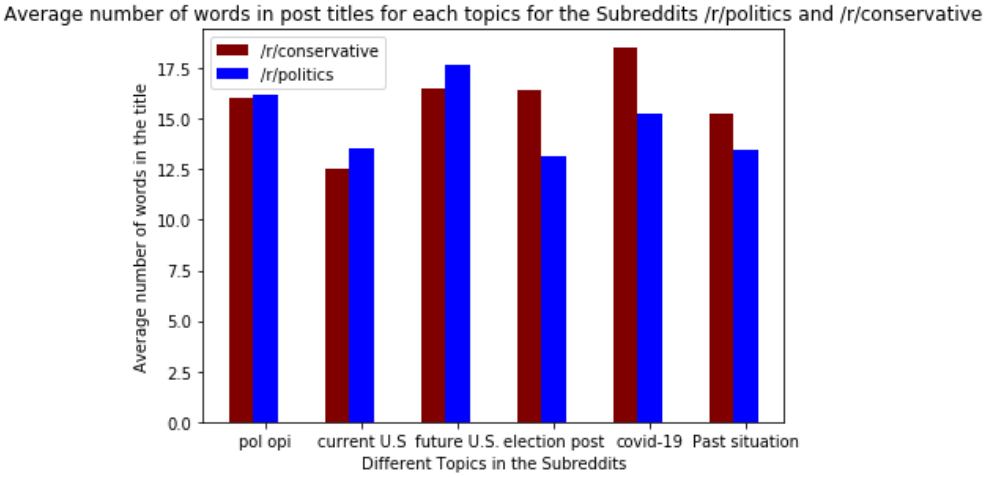


Figure 3

Top 10 words by TF-IDF in each category of /r/conservative

Topics	Top 10 words by TF-IDF score
Political opinion	1- “democrats”, 2- “do”, 3- “evidence”, 4-”even”,5- “reportedly”, 6- ”obama”, 7- “believe”,8-”tucker”, 9- “powell”, 10- “are”
Current Situation in The U.S.	1-"wisconsin",2-"counties",3-"campaign", 4-"recounts", 5- "wayne",6-"head",7-"path", 8-"administration",9-"file",10- "fires"
Future of the U.S.	1-"gun",2-"billion",3-"proposed", 4-"reparations",5-"argued",6- "vows",7-"tax", 8-"owners", 9-"full",10-"report”
General election posts	1-"votes",2-"finds",3-"second",4-"county", 5-"ballots",6-"again",7-"poll",8-"uncounted, 9-"thousands",10-"georgia"
Coronavirus	1- "covid",2-"vaccine",3-"back", 4-"warns",5-"follow",6-"holidays" 7-"cuomo",8-"blasts",9-"makers", 10-"rushing"
Past situation in the U.S.	1-"blm",2-"child",3-"schools",4- "take", 5-"your",6-"threats",7-"thug,8-"sucked", 9-"punched",10- "supporter"

Figure 4

Top 10 words by TF-IDF in each category of /r/politics

Topics	Top 10 words by TF-IDF score
Political opinion	1-"about",2- "americans,3-"should",4- "we", 5-"calls",6-"never",7-"an",8-"like",9-"call", 10- "republican"
Current Situation in The U.S.	1-"officials",2- "judge",3-"fires",4-"blocks", 5-"confirmation, 6-"block", 7-"unaccompanied", 8-"children", 9-"democracy",10-"wall"
Future of the U.S.	1-"lose",2-"nominees",3-"plans", 4-"inauguration",5-"before",6-"january", 7-"crimes",8-"covid",9-"twitter", 10-"privileges"
General election posts	1-"million,2-"shows",3-"fraud",4-"claims", 5-"still",6-"voter",7-"recount", 8-"concludes",9-"count",10-"ahead"
Coronavirus	1-"vaccine",2-"coronavirus",3-"covid", 4-"death",5-"toll",6-"distribution", 7-"weeks",8-"months",9-"put",10-"behind"
Past situation in the U.S.	1-"thanksgiving",2-"mar",3-"lago", 4-"fired", 5-"officials",6-"secretary",7-"did",8-"called", 9-"judge",10-"czar"

Figure 5

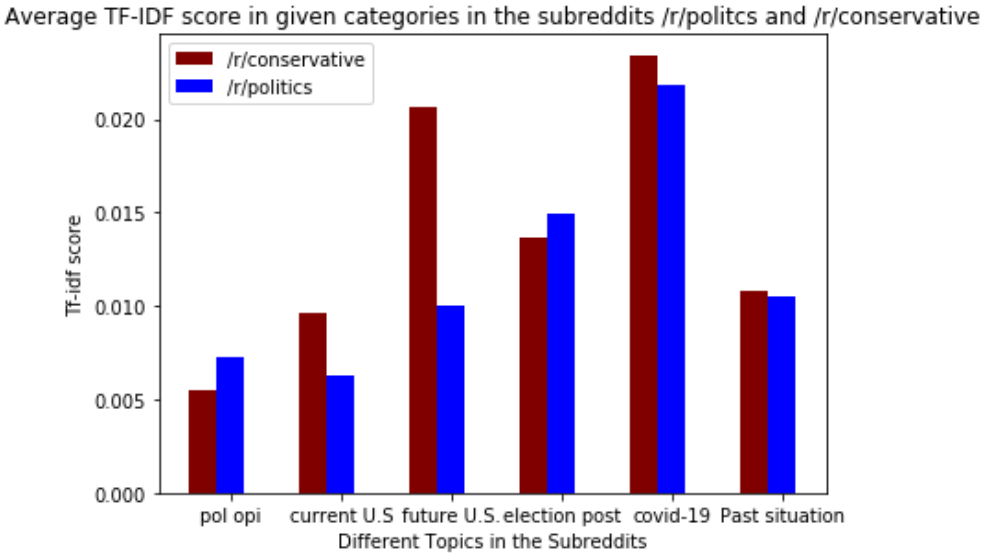
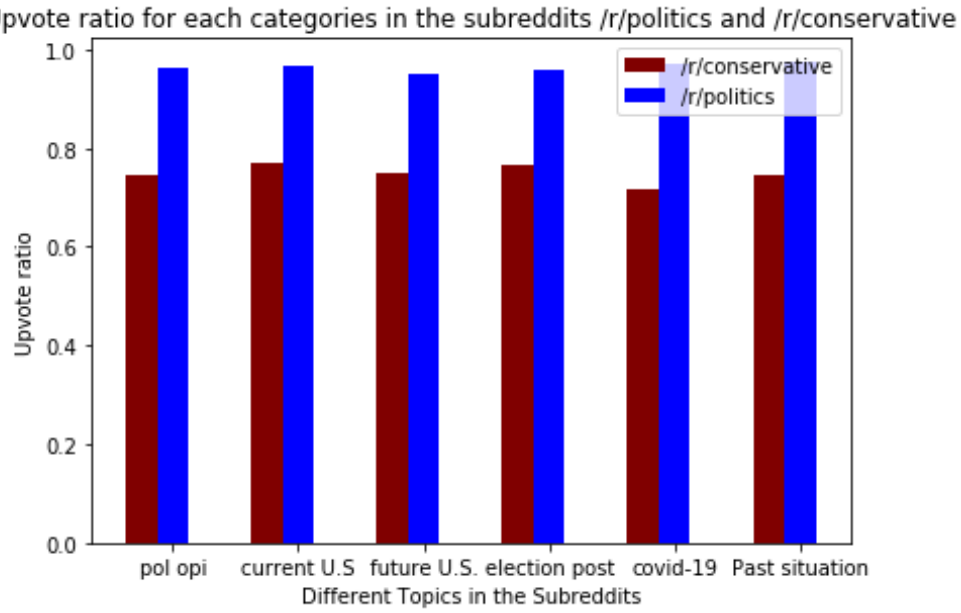


Figure 6



5 Discussion

An important thing to note about Reddit is that the content of posts are often not independent of each other. For example, shortly after the election was called by the major news networks, almost the entirety of "hot" in /r/politics was composed of titles of the form "Goodbye [republican politician]", an unusual pattern compared to the other days the subreddit was visited. Patterns such as this could influence the results of an analysis in such a way to give misleading interpretations of the political discussion.

Figure 1 depicts the number of times a post was placed in a certain category. Unfortunately, the posts were not as evenly distributed as desired: 152 and 110 in total were placed in the "political opinion" and "current situation in the United States" categories while only 26 and 16 posts were placed in the "future of the United States" and "coronavirus" categories respectively. Since the TF-IDF scores require computing frequencies across various categories, there is potential for bias in the results. For example, say a word appears twice overall in the /r/politics data. If it appears once in the "coronavirus" category (6 posts form /r/politics were placed in this category), the probability of it being found in another category (decreasing its IDF score) would be $194/200 = 0.97$. If the word appears in "political opinion" category (88 posts from /r/politics were placed

in this category), the probability of appearing in another category is 112/200 = 0.56. Perhaps a solution to this issue would have been to break up the two larger categories into smaller ones, or combine smaller categories into larger ones.

Figure 2 depicts the average number of words found per post. There does not appear to be a large discrepancy in post length between/r/politics and /r/conservative. There is also very little variance across categories (the range is 12.547-18.5). This may suggest that the structure of the titles are not influenced by its category

In regards to the upvote_ratio, we see from figure 6 that posts on average in /r/politics have an upvote_ratio of over 0.950 in all categories. In /r/conservative the ratio range from 0.72 to 0.77. There are many possibletw reasons for this discrepancy. The simplest explanation is that users of /r/politics are more likely to agree amongst themselves compared to those who frequent /r/conservative. However, there may be underlying factors causing this difference. For example, when posts on Reddit obtain exceptionally many upvotes in a short amount of time, they may also appear on /r/all, a place where posts of any kind can appear and may be read by those with different political opinions or by those possibly not concerned with politics at all. As of December 7 2020, /r/politics has a total of 6,979,664 subscribers whereas /r/conservative has 577,307, indicating that the Reddit user-base as a whole may be biased towards liberal ideals. If a post from /r/conservative reaches the top of /r/all, it may quickly be voted down since its opinions go against the majority website consensus. This is to say that the scores of posts from /r/conservative may be affected by the overall bias of the website, and may not be indicative of the actual engagement with the material.

We obtained mixed results in regards to our TF-IDF analysis. For the category "political opinion" in /r/politics, some of the words with the highest TF-IDF scores were "about", "should", "we" and "an". The structure of political opinions, with phrases such as "we should...", "this is not about [person],it's about...", "Americans are worried about...", "it's an...", means that these words show up a disproportionate amount to times despite being common words. We also obtained issues with unique words that branch out into multiple words. For example, the name "Mar-a-lago" splits into three words. The words "mar" and "lago", obtaining high TF-IDF scores in the category of "Past situation in the United States" in /r/politics, then took up two of the highest 10 scores despite referring to the same thing.

While looking at our plot in figure 1, we can also see a difference between the number of posts belonging to topics in the two subreddits. The biggest difference is noticed in the political opinion category, with /r/conservative having 19 more posts in that category compared to /r/politics. The topic that seems to be discussed more in /r/politics is the topic of Coronavirus. There were 10 more posts than in the in /r/conservative subreddit. This observation to a certain extent can be consistent with the hypothesis that Conservatives in the United States take the Covid-19 pandemic less seriously. This is hypothesis can also arise when looking at the top tf-idf words for the coronavirus. In /r/politics, there is mentions of words such as "death","toll","months", that are often synonymous with taking the seriously. Where as in the conservative subreddit, there is mention of words like "rushing","holidays".That could be interpreted with the idea of downplaying the effects of the pandemic.

Another difference in the number of posts can be observed for the election posts category and the future of the U.S.. With exceeding the other subreddit by about the same amount,/r/politics by 6 for the future of the U.S. and /r/conservative by 7 for the election post category. This could indicate that democrats or liberals are more excited after winning this election and will thus be more willing to discuss topics around the future of the country. In /r/conservative nudge in election posts might be linked to the fact that Donald Trump, the conservative candidate and incumbent president, is contesting the election. This might be why 3 weeks after the election results, there might still be more discussion around election.

We now compare the key words (figures 3 and 4) in both subreddits by category

Political opinion

- In tf-idf top 10 words for each subreddit, most frequent words in /r/conservative are: "democrats", "evidence", "do". For /r/politics, most frequent words are: "about", "americans", "we". From these words, the political opinion are related to the truth such as reports or evidences, and their believes. However in conservative, people might talk about what the governments, politicians should do, also the suggestions in present or future.

- Interestingly, /r/conservative makes reference to "democrats" while /r/politics makes reference to "republicans". This indicates that political opinions on Reddit may be more concerned

with criticisms of the opposition than analysis of politicians with similar ideologies.

Current Situation in the U.S.

- In the top 10 words for tf-idf, in /r/politics, words "wisconsin", "counties", "campaign" are frequently appears in the posts, but in /r/conservative, words: "judge", "fires", "blocks", and "confirmation" are mostly appears in the posts. Comparing these words in two subreddits, the conservative are more about doing actions within the governments and administration in the current situation in U.S.

Future of the U.S.

- In the tf-idf top 10. We see that for the future of the U.S. words like "inauguration", "plan" appear in it /r/politics, this contrasts with the top of /r/conservative which has words like "gun","billion", "tax". It seems that the conservatives are concerned of what is going to happen in the future to them. Where as the liberals seems more optimistic.

General Election Posts

- In /r/politics there are two words that in the tf-idf that are contrasting the words "find", /r/conservative and "claim" in /r/politics. It is consistent with what has is being said by conservative leadership saying that they have found evidence of fraud, where as the Democratic leadership accused them of coming up with baseless claims.

- Secondly, there seems to be uses of words such as "uncounted", "thousand" in /r/conservative. This is consistent with the claims that there the conservative leadership wit a few claims from the conservative leadership.

- Interestingly, /r/politics contains more words directly pertaining to voter fraud. Some of the highest terms include "fraud", "claims", and "recount" whereas in /r/conservative there is only the word "uncounted". In both cases we obtain generic associated words such "votes", "ballots", "poll", "count". It perhaps would have been much more interesting for this category to collect words in tuples to hopefully get more context about these words.

Coronavirus

- The word vaccine had the 1st and 2nd highest TF-IDF rating in /r/politics and /r/conservative respectively. The discussion around the term mainly revolves around who will get credit for developing the vaccine and how it is being used for political gain. It is surprisingly not used in a context debating the science behind it.
- Example1: Cuomo Blasts Trump, Vaccine Makers For Rushing COVID Cure.(/r/conservative)
- Example2: Trump is furious because he thinks Biden will get the credit for coronavirus vaccine developments. (/r/politics)
- The conversation on the liberal side makes reference to the words "death" and "toll", while the conservative side uses words such as "holidays". This may indicate that users of /r/politics are much more concerned with the dangers of the virus that those of /r/conservative

Past Situation in the U.S.

- The words contained in the lists are very different from each other. Not a single word has a high TF-IDF score in both subreddits indicating very different perspectives on the past between liberals and conservatives.
- The word with the highest TF-IDF score on the conservative side is "blm" in reference to Black Lives Matter. The words "thug", "threads", and "punched" are also relevant, perhaps indicating that conservative Reddit users consider the Black Lives Matter protests and the surrounding violence to still be an important issue.
- The words on the democratic side reference "Mar-A-Lago", "fired", "officials", "secretary", and "judge". These words have used often in reference to the Trump administration, indicating a much different concern for liberal Reddit users compared to conservative ones.

6 Group Member Contributions

- Marek Adamowicz - Main author of report. Assisted with code multiple times by providing feedback.

- Romain Floreani - Author of python files collecthottest.py to collect Reddit posts and filter them for mentions of Trump and Biden. Contributed to report
- Sihan Wang - Author of python file "tfidf.py" to conduct TF-IDF analysis. Contributed to report