COMP 598 Group Project
Group 13
Dibbyajit Dam
David Budaghyan
Abdullatif Hassan

Overview:

The US elections had a significant impact on the lives of Americans. This study focuses on the perception of the election results for each of republicans and democrats. Reddit data is collected over 3 days from 2 subreddits, r/conservative, and r/politics, with the community of each representing a majority for one of the 2 main political parties. Following that, the tweets were classified into different categories depending on topic and a tf-idf analysis was run to identify the words that prevailed over the data for each category. It was found that overall, conservatives view the election as much less legitimate than democrats do.

Data:

We scraped 1000 posts that mention either Trump or Biden from each of r/conservative and r/politics. A post mentions Trump or Biden if and only if it contains the word "Trump" or "Biden" (case-sensitive, Alpha-numeric). Collection was done along 3 days consecutively (Nov 24-26). For each day, 400 posts with titles mentioning either candidate were collected from each subreddit and 334 from those posts were then selected randomly. Thus, our initial dataset consisted of 2000 posts mentioning either candidate, coming from both subreddits equally. After combining all posts from the same reddit and removing duplicate posts, the number of posts for r/conservative and r/politics is 452 and 786 respectively. Those were the posts that were later annotated.

Date	Subreddit	Raw Data (contains 'Trump or 'Biden')
November 24, 2020	Politics	334
November 25, 2020	Politics	334
November 26, 2020	Politics	333
November 24, 2020	Conservative	334
November 25, 2020	Conservative	334
November 26, 2020	Conservative	333

Methods

Collection

Instead of gathering 2000 posts and then keeping only those that mention Trump or Biden, we directly collected 2000 posts that mention the candidates. This was done in order to get enough data, as we did not want to end up with too little posts after filtering. Indeed, we think this was crucial as future filtering reduced our dataset size significantly.

The next step was to drop duplicate posts. A post is considered duplicate if and only if the post title and author are the same. Two posts with the same post titles but with different authors were not considered duplicates. Hence, our final dataset consisted of 452 unique posts from r/conservative and 786 unique posts from r/politics, all of them mentioning either Trump or Biden. We think this discrepancy is caused by the sizes of the subreddits. r/conservative is much smaller, hence when scraping in consecutive days we might have scraped the previous days posts as well. (Note: We noticed this duplication problem later during the project. Given more time, we would have scraped more posts in order to have the same number of posts from each subreddit.)

We also experimented with not removing the duplicates. It turns out that the top tf-idf words were slightly shuffled but overall stayed the same. We found this interesting as this shows that duplicate posts did not specifically belong to a certain topic, rather they were distributed equally over all the topics.

Topic selection

In order to come up with unbiased topics (with respect to candidates and to the day scraped) we selected 100 random posts across all days from each candidate. We then permuted these 200 posts and read them. We came up with 7 topics which we present in the next section. The

annotation of the rest of the dataset was done by assigning each group member to annotate the posts from a given day.

tf-idf computation

We computed our tf-idf using all (452+786) posts. In other words, we defined a document to be a (subreddit, topic) pair. Hence, we had 14 documents. This is consistent with our objective, since this means we attribute the highest tf-idf scores to words that are specific to the topic as well as the subreddit.

Results:

Typology

Since the main goal was to study the perception of election legitimacy, 3-4 of our topics are directly related to the latter. However, in reading the 2000 posts collected, we noticed that even the posts which talked about covid19, racism, or any other more general topic not directly mentioning the election or its legitimacy, were mostly still about the election process. Hence, we decided to add these topics to our typology, since we thought they would contribute to our understanding of the people's perception from both subreddits regarding the election. The following is our typology:

Note: in addition to the description given, it is assumed that the post has some connection to the election, and as just explained, it almost always did for topics excluding N/A:

- **Transition process and Concessions:** If the post mentions a statement, action or opinion regarding the transition of the presidency, Biden appointments or Trump concessions.

Example: " 'A cabinet that looks like America': Harris hails Biden's diverse picks. "

- **Lawsuits/legitimacy:** If the post mentions any legal process, the statements of a lawyer or a judge, a law about the election process, the court or court decisions.

    Example: "Biden's win in Nevada certified by Nevada Supreme Court."

- **Foreign Relations and Intervention:** If the post mentions anything about another country, another president or some element of the US that has to do with other countries. Also include any post talking about foreign intervention in US politics.

    Example: "Putin Says He's Not Ready to Recognize Biden as U.S. President."

- **Voting/vote count:** If the post mentions any statistic about the number of votes (e.g. demographic, regional, time) and anything about the vote counting process.

    Example: "Poll: One In Six Biden Voters Would Have Changed Their Vote If They Had Known About Scandals Suppressed By Media"

- **Covid and health:** If the post mentions COVID19 or if the post mentions the healthcare system with a very clear connection to Covid (this was clarified between team members)

    Example: "Trump campaign strategic advisor and unofficial Jewish liaison tests positive for coronavirus"

- **Racism / minorities**: If the post mentions anything regarding racism or minorities or immigrants.
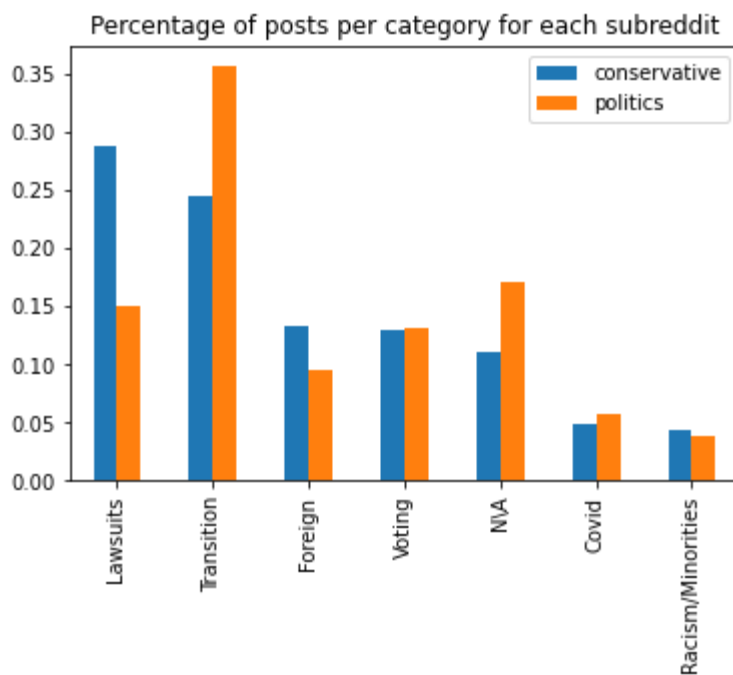
    Example: "Obama takes jab at Hispanic voters for supporting Trump."

- **N/A:** The post has absolutely nothing to do with the election OR it cannot fit in any way in the above categories.

  Example:" El Paso lawyered up to try and collect $570,000 owed to the city by the Trump campaign from a rally 2 years ago"

Topic engagement

We plot a histogram indicating the engagement for every topic in each subreddit as a percentage of posts of that subreddit.

Percentage of posts per category for each subreddit

Only the 3 topics ['Lawsuits', 'Transition', 'N/A'] differ significantly in terms of engagement between the subreddits. Indeed, it is interesting to note that the conservative subreddit talked much more about lawsuits than the politics subreddit, but a bigger percentage of the latter's posts mentioned elements about the transition process. We will interpret this and other findings in the next section. But first we display the top 10 words with respect to tf-idf for each category and for each subreddit.

Transition process and Concessions:

Politics: ['Cabinet', 'administration', 'transition', 'leave', 'Gaetz', 'himself', 'electoral', 'Which', 'college', 'signals']

Conservative: ['being', 'Blinken', 'Leave', 'Transition', 'Flynn', 'Return', 'Process', 'Secretary', 'Mattis', 'Political']

Lawsuits/legitimacy:

politics: ['Legal', 'Rudy', 'Giuliani', 'Pennsylvania', 'legal', 'Court', 'fraud', 'lawsuit', 'Lawyer', 'making']

Conservative: ['Court', 'Powell', 'Review', 'Sidney', 'ballot', 'Pennsylvania', 'legal', 'court', 'Nevada', 'lets']

Foreign relations and intervention

Politics: ['China', 'Xi', 'policy', 'foreign', 'Jinping', 'congratulates', 'deaths', 'introduces', 'Foreign', 'picks']

Conservative: ['China', 'Flynn', 'Michael', 'pardons', 'Chinese', 'Afghanistan', 'Iraq', 'security', 'Foreign', 'Staffer']

Voting/vote count:

Politics: ['certifies', 'Pennsylvania', 'presidential', 'Quaid', 'Randy', 'vote', 'million', 'winner', 'results', 'becomes']

Conservative: ['Known', 'Votes', 'Media', 'Had', 'Voters', 'Unless', 'Nearly', 'Scranton', 'Tenth', 'honor']

Covid and health:

Politics: ['pandemic', 'coronavirus', 'COVID', 'response', 'tests', 'positive', 'address', 'Covid', 'Mnuchin', 'Gather']

Conservative: ['COVID', 'Emergency', 'Antibody', 'FDA', 'Cocktail', 'vaccine', 'Grants', 'Guard', 'Use', 'By', 'Dominion']

Racism and minorities:

Politics: ['Black', 'racist', 'Immigrant', 'Urged', 'race', 'Sovereignty', 'Pipelines', 'Halt', 'repent', 'girls']

Conservative: ['Amnesty', 'Aliens', 'Illegals', 'Days', 'Illegal', 'Latino', 'M', 'Racist', 'Going', 'Border', 'March']

N/A:

Politics:['Thanksgiving', 'Eve', 'Lame', 'down', 'address', 'unity', 'Night', 'Melania', 'mine', 'person']

Conservative: ['Book', 'extends', 'Bible', 'Devout', 'Intimidated', 'Catholic', 'Tall', 'Alyssa', 'Milano', 'olive', 'branch']

Discussion:

By observing the engagement plot presented above, the conservative subreddit discusses lawsuits on the election's legitimacy much more than the politics subreddit. Naturally, this can be attributed to the fact that the conservative subreddit's users hope that the lawsuits are successful in determining that the election was illegitimate. The politics subreddit on the other hand discusses this topic almost half as much. This clearly indicates that the users in the politics subreddit are not concerned with, or at the very least do not want to discuss, the legitimacy of the election. Moreover, it is worth noting that the lawsuit topic is the most popular topic in r/conservative, whereas it is tied for fourth in r/politics. This further confirms the above interpretation.

On the other hand, looking at the same plot, we see that r/politics's users discuss the transition process much more than the r/conservative's users. This again has a natural interpretation, which is that those in r/conservative are less keen on the transition taking place and less engaged in the various transitional processes and nominations conducted by the Biden administration. Those in r/politics however are very engaged in these topics. This type of discrepancy cannot simply be explained by the fact that the winners are more enthusiastic for what is to come. In fact, we think that this indicates the relative refusal of the election results by the r/conservative users. Indeed, it is quite telling that the top 3 words in r/politics for this transition topic are "cabinet", "administration" and "transition", whereas only "transition" of these three words appears in the top 10 in r/conservative, in 4th place.

When it comes to the words in the "voting" topic, we can see the same type of pattern. We have the revealing words: "certifies", "presidential", "winner" and "becomes" in the politics subreddit's top words. None of these words appear in the r/conservative's top words. Once again this is evidence that the conservative subreddit avoids talking about Biden's victory, whereas the politics subreddit has it as one of its top subjects. Instead, in r/conservative, we have the words: "votes" and "media", which were part of posts criticizing the media's negative impact on the voting process. As for why the top word in r/conservative is "known", it's because there was a very popular piece of information, posted by different users, that roughly said the following: 17% of people who voted democrat say they would have voted differently had they *known* the various fraudulent things done by the democratic party with regards to the voting process. In other words, the most popular post in r/conservative that is specific to the "voting" topic is directly mentioning the election's illegitimacy. This clearly indicates that r/conservative does not perceive the election as being legitimate.

We can even see this through the "foreign" topic's top words. In r/politics, we have the words: "congratulates", referring to Xi Jingping congratulating Biden for his win, "introduces", referring to the Biden administration introducing people and his future plans, and "picks" referring to Biden picking different people for various positions. None of these words appear in r/conservative's top words. Instead we have "pardon", which refers to Trump pardoning various people, most importantly "Michael" "Flynn", also top words in r/conservative, but nowhere to be found in r/politics. We also have other words not found in r/politics like "Iraq", "Afghanistan" and "security". Once again r/conservative avoids talking about the successes and developments of Biden's win.

In the above 4 paragraphs, we have clearly demonstrated through the relative engagement plot and through the top words in the "lawsuits", "transition", "voting" and "foreign" topics, that the users in r/conservative view the election as less legitimate than do those in r/politics. We concluded this based on the fact that the r/conservative subreddit engages much less in anything related to Biden's success and future presidency and it often explicitly mentions the election's illegitimacy.

Since these 3-4 topics are central elements to the legitimacy of the election, it makes sense why we were able to see such clear differences in the perceptions and the engagement of both subreddits regarding this matter. We will now analyze the other topics: "covid", "racism/minorities" and "N/A". These topics are understandably less indicative about the perceptions of the subreddits on the election's legitimacy, but they are nonetheless indicative about general differences between the two. We mention a few interesting observations below.

Covid: "pandemic" is the top word in r/politics, yet it is not found in the top 10 of r/conservative. "Tests" is the 5th word in r/politics, and again nowhere to be found in r/conservative. This indicates

that people on the conservative side may be less engaged in testing and describing the spread of COVID as a pandemic.

racism/minorities: "black" is the top word in r/politics and "immigrant" is 3rd. Neither are found in r/conservative's top words. Instead we have "illegal", "illegals" and "aliens". This indicates the differences in the treatment and labeling of people who live in the U.S. but were born outside. It also indicates that r/politics is more engaged in topics involving black people.

N/A: We have "devout", "bible" and "catholic" in r/conservative's top words. None are in r/politics. We conclude that r/conservative users are more engaged in religious discussions and topics, particularly involving Christianity.

To conclude, the differences in perception regarding the election's legitimacy are clear. r/conservative users see the election as being less legitimate than do those in r/politics. Based on the keywords found and the observations made, we think that our typology combined with our definition of a document for the tf-idf computation was successful in extracting the perceptions on election legitimacy from both subreddits. However, we also realized that merging the "Covid" and "racism/minorities" topics with "N/A" would not have changed our analysis. Indeed "Covid" and "racism/minorities" did not offer insights on how the subreddits perceive the election's legitimacy, and they also took up less than 7% of engagement in each subreddit. Nonetheless they still offered interesting insights on differences between the two groups.

Group Member contributions:

Rather than allocating each part to individuals, we decided to work together on each section collaboratively. This was mainly because some parts depended on other parts (For example,

without the data collection phase, annotation or analysis can not be done). This approach really helped the group to finish the project in time. We believe this approach made our work more efficient.