COMP 598 Final Project – Data Science Project

By: Brian Hu, Dawei Zhou, Logan Ralston (Group 4)

Overview

After collecting Reddit data over a three day period following the 2020 US election from a primarily liberal and a primarily conservative subreddit, several findings were made regarding the discussion surrounding each candidate and the level of engagement for different topics within both subreddits. Both communities had a similarly extreme number of posts directly challenging the outcome of the election – approximately 40% of their overall posts made on either candidate; 5 times as many as the number posts discussing the election results in a neutral manner (~8% of each group's posts). The conservative community generated more discussion around Covid, foreign affairs, and policies (~30% overall) than their liberal counterpart (~20% overall). Finally, there were more instances of personal attacks and discussion from the liberal community (~8% more than the conservatives), while the conservative community involved more instances of outrage, threats and protests (~3% more than the liberals).

Data

In order to investigate the salient topics discussed around each candidate and the relative engagement with those topics among liberals and conservatives, we gathered the hottest posts from two Reddit communities: /r/politics and /r/conservatives, which are primarily pro-Biden and pro-Trump respectively. The "hottest" posts are ones that have been upvoted the most by members of their community and often generate the most discussion. By collecting these posts from subreddits associated with their respective political stance, each post can be assigned a topic of discussion to be characterized in order to analyze the engagement of these topics from each subreddit.

Each datapoint includes: *name* – a unique identifier that distinguishes one post from another, *title* – a brief overview of the post written by the original poster, and *subreddit* – the originating subreddit of the post (either /r/politics or /r/conservatives). The data collection was done every 24 hours over the course of three days from November 20 to November 22, 2020.

In order to collect a total of 1000 “hottest” posts, we began by collecting 400 posts the first day, 300 posts the second day and 300 posts the third day, from each subreddit. We then filtered through these posts to collect the ones that had the keyword ‘Trump’ or ‘Biden’ in the post title, and finished with a combined total of 964 posts – 689 posts (71%) from /r/politics and 275 posts (29%) from /r/conservatives. However, there were still posts with the same ID within our dataset – the posts that were “hottest” for more than one day, so we decided to keep only posts that had a unique ID. This was to ensure that there would be no repetition in the annotation phase that would affect our analysis of the results, given that duplicate posts would result in duplicate annotations, thus skewing the results and allowing for an overrepresentation of annotations. After filtering out duplicate posts, we ended with a total of 902 unique posts ready for annotation – 672 unique posts (75%) from /r/politics and 230 unique posts (25%) from /r/conservatives.

Methods

First, we needed to prepare the posts for open coding. To make the data as representative as possible, we selected 200 random posts from the whole population. We initially tried to use the major issues in the US election as the topics – economy, health care, supreme court appointment, Covid, crimes, and foreign policy. During the open coding, we found that the topics we prepared did not categorize our data well, as many posts were not able to be categorized with these topics. Therefore, we decided to develop a more comprehensive

typology as we annotated the sample posts. After open coding, we had the following topics with brief descriptions:

LEGAL - Posts discussing legal challenges to election results and their outcomes

STEAL - Posts alleging other side is trying to steal the election (while not specifically referring to legal challenges)

VOTE COUNT - Posts reporting statements on vote tallying, statistics or outcomes, without alleging fraud on either side

POLITICS - Posts talking about (whether praising, neutral or criticizing) either the president or president elect going about their job, carrying out policy, etc. (without alleging fraud)

COVID - Posts relating to Covid

MEDIA - Posts relating to how Trump and Bidens media coverage (how companies like CNN, Fox, Twitter and Facebook are treating either candidate)

DENIAL: posts relating to Trump's denial towards the result

DISCONTENT: posts relating to people's discontent towards the result

PERSONAL: posts relating to people's personal opinion

TRANSITION: posts relating to the presidential transition

VIOLENCE: posts relating to violence

After a sanity check, we realized some problems existed with this typology. First, we found that too many posts fell into "POLITICS", and it would make more sense if we divided them into "DOMESTIC" (domestic politics) and "FOREIGN" (foreign politics). There was some overlap among "DENIAL", "DISCONTENT", "STEAL", and "VIOLENCE". To ensure our typology was well-defined, we kept "STEAL" and recategorized the others into "DOMESTIC" or "STEAL". Similarly, the topic "PERSONAL" seemed unreasonable because most posts are personal opinion to some degree, so we decided to drop this topic and recategorize the posts. In addition, there are few posts with "TRANSITION", so we decided to merge them into "DOMESTIC". Finally, we merged the posts from /r/politics and /r/conservative and removed the duplicate

posts, finishing with 902 annotated posts. After that, we divided posts into 3 parts and each one of us do single annotation on one part.

After annotation, we calculated the TF-IDF scores for each topic and selected the words with the highest value. The formula we used was:

$$TF - IDF\ score\ =\ TF(term|topic)\ *\ log(number\ of\ topics/number\ of\ topics\ that\ term\ is\ used\ in)$$

Finally, we found that “MEDIA” and “DOMESTIC” did not show meaningful results, as 42% of posts were coded as “DOMESTIC”, and “MEDIA” itself did not seem like an actual topic of discussion. We replaced them with “POLICY”, “PERSONAL”, and “MOB”, thus finalizing our topics to characterize the data with.

Results

By conducting an open coding of 200 posts from our dataset we developed a topology for categorizing the posts into 8 topics that are comprehensive, well-defined, objective and provide insight into what the online reddit base of either candidate was concerned with over the November 20th to 22nd period for which data was collected. The topics developed and then used to label all posts are presented in Table 1. Posts were categorized based on their title.

Table 1: Topic Definitions

Topic	**Definition**
LEGAL	Posts which directly reference legal challenges to election results, whether it be speculation on the details of the case or discussing their outcomes. Positive examples: ● “Why a federal judge threw out Trump's Pennsylvania election lawsuit” ● “Rudy Giuliani says election case losses help Trump campaign’s strategy to get ‘expeditiously’ to Supreme Court” ● “Pat Toomey: Trump Campaign Has Exhausted Legal Options and Biden Won” Negative examples: ● “Trump Lawyer in PA Placed Under Protection After ‘Threats of Harm’”

	(*though it references a lawyer it doesn't reference a legal case, instead dealing with threats of violences which means that this should be categorized as MOB - see definition below*)
STEAL	Posts which accuse either candidate of fraud or outright stealing the election (without referencing a legal case - if it references legal case topic should be LEGAL) Positive examples: ● "Trump Warns 'Massive Fraud' to Be Uncovered in Michigan Vote" ● "Donald Trump's Effort To Steal The Election Is Comically Stupid — And Extremely Dangerous" ● "Mitt Romney: Trump's efforts to overturn election result are 'undemocratic'" Negative examples: ● "Trump digs deeper into debunked conspiracy theories instead of embracing reality" (*doesn't say whether the conspiracy theories that Trump is purporting are related to election fraud, so this is just a personal attack by against Trump instead of accusation of election fraud, thus should be categorized as PERSONAL*)
VOTE COUNT	Posts which talk about statistics, results from election counts or recounts without alleging fraud by either party Positive examples: ● "Michigan, 6:31 AM, 150,000 votes, 96% for Biden...nothing to see here" ● "Georgia recount results: Biden still ahead by 12,000 — but Trump has one last roll of the dice" Negative examples: ● "Trump campaign adviser Steve Cortes: Sudden Stop In Mail-In Vote Count 'Suspicious' -- Particularly since it was followed by a "deluge" of votes for Joe Biden in key battleground states." (*though it's dealing with the count of votes, there is an accusation of fraud thus it goes in STEAL*)
COVID	Posts about the policy of either candidate specifically related to Covid (not related to election results or challenges though as that would go in one of LEGAL, STEAL or VOTE COUNT) Positive examples: ● "Trump Continues To Falsely Downplay COVID After US Death Toll Reaches 250k" ● "Biden vows there will be no national shutdown" Negative examples: ● "Trump Announces New Regulations Lowering Prescription Drug Costs" (*though this will eventually affect future vaccine costs, it does not directly reference Covid so it goes in POLICY*)
FOREIGN	Posts referencing foreign policy or foreign countries commenting on

	election (accusations of foreign interference would go in STEAL, foreign affairs involving Covid would go in COVID) Positive examples: ● "Trump administration in 'staggering' isolation at UN on health issues" ● "Ilhan Omar Urges Biden To 'Reverse' President Trump's Middle East Agreements" Negative examples: ● "G20 leaders meet to discuss help for poorest nations in post-Covid world, Trump golfs" (*though it involves foreign countries, it focuses on post Covid efforts so it goes in COVID*)
POLICY	Posts talking about (whether praising, neutral or criticizing) either the president's or president-elect's policy position or performing an act to promote or inhibit a policy (policy must not be related to Covid or foreign policy or it would go in COVID or FOREIGN topic respectively) Positive examples: ● "Biden Wants Billions in Taxes from Firearms Owners" ● "Trump Announces New Regulations Lowering Prescription Drug Costs" Negative examples: ● "Biden vows there will be no national shutdown" (*though this is a policy invoked by the president-elect, it is directly related to Covid so it goes in COVID*)
PERSONAL	Post's title insults, praises or makes statements related to the President, President-elect or a member of their teams. Positive examples: ● "Atlanta Mayor Suggests Trump Would 'Eat His Own Children' To Advance Agenda" ● "Business Leaders to Trump: Give It Up" Negative examples: ● "Mitt Romney issues scathing statement condemning President Trump's voter fraud claims" (*though it's a personal attack on the President, it involves allegations of fraud so it goes in STEAL*)
MOB	Posts relating to groups of one candidates' supporters expressing outrage, protesting or threatening violence or someone vilifying or praising a group of either candidates' supporters. (Our definition of group of supporter does not include groups of politicians, only groups of non-government supporters) Positive examples: ● "DNC Member Rants: 'Deprogram' 75 Million Trump Supporters" ● "Michigan State senator met by protestors as he arrives for meeting with Trump" Negative examples:

	● "Arizona's secretary of state is the latest election official to receive death threats, and she's ripping Trump and Republican leaders for their baseless claims of fraud" (*though there is outrage towards Arizona's secretary of state, the mention of fraud categorizes this post as STEAL*)

Table 2: Posts of Topic Frequency by Subreddit

Topic	**# of Posts from /r/conservative**	**# of Posts from /r/politics**	**% of Posts from /r/conservative**	**% of Posts from /r/politics**
LEGAL	42	132	17.721%	19.880%
STEAL	49	156	20.675%	23.494%
VOTE COUNT	19	58	8.017%	8.735%
COVID	22	43	9.283%	6.476%
FOREIGN	21	27	8.860%	4.066%
POLICY	32	69	13.502%	10.391%
PERSONAL	36	159	15.190%	23.946%
MOB	16	20	6.751%	3.012%

After labelling the posts with their topics, we characterized the posts by computing the 10 words in each category with their highest TF-IDF score.

Table 3: Topics Characterized by TF-IDF Scores

Topic	**Words with Highest TF-IDF Scores (Top 10)**
LEGAL	'judge', 'pennsylvania', 'campaign', 'lawsuit', 'lawyer', 'federal', 'legal', 'certification', 'lawyers', 'team'
STEAL	'fraud', 'overturn', 'romney', 'coup', 'claims', 'election', 'undemocratic', 'results', 'efforts', 'campaign'
VOTE COUNT	'georgia', 'recount', 'certifies', 'win', 'secretary', 'presidential', 'race', 'certify', 'results', 'officials'
COVID	'covid', 'coronavirus', 'tests', 'pandemic', 'carson', 'golf', 'vaccine', 'positive', 'at', 'meeting'

FOREIGN	'middle', 'israel', 'east', 'world', 'administration', 'and', 'not', 'that', 'president', 'on'
POLICY	'cabinet', 'top', 'picks', 'administration', 'chief', 'first', 'senate', 'tuesday', 'could', 'staff'
PERSONAL	'transition', 'team', 'christie', 'about', 'twitter', 'chris', 'elect', 'calls', 'national', 'hogan'
MOB	'supporters', 'he', 'voters', 'that', 'in', 'election', 'a', 'as', 's', 'biden'

Discussion

The first key point of interest revolves around the occurrences of each topic by subreddit. An effective form of comparison is to directly compare the percentage of posts with their designated topics from each subreddit; this allows a comparison to be drawn regarding the topics each community is more engaged with.

From looking at the number of posts in Table 2, it is evident that there are more posts regarding US elections in /r/politics. Since we have filtered out the posts that do not mention ‘Trump’ or ‘Biden’ along with duplicate posts during the data collection phase, this difference in number of posts is most likely caused by the difference in members. There are 576k members in /r/conservative, as opposed to the 7million in /r/politics. Therefore, the discussion in /r/conservative is more concentrated on some posts and there are less new hot posts coming up.

Upon closer look at the popularity of topics, the portions of each topic were similar between the /r/conservative and /r/politics communities. The topics directly concerned with challenging the outcome of the election, “LEGAL” and “STEAL”, made up 38.4% of the /r/conservative posts and 43.4% of the /r/politics posts. These are staggering numbers, showing extraordinary engagement in both communities with posts on the other side trying to interfere with the election results. ~40% of posts about Trump or Biden cast doubt on them being good

faith actors in ensuring fair democratic elections, this is symbolic of a truly extraordinary lack of trust in the election's legitimacy.

While both sides had extremely high engagement with posts concerned with challenges to the election's result and accusations of fraud, /r/politics had 5% more posts from these categories than /r/conservative. This supports the theory that the liberal community is more concerned with Trump trying to steal back an election he lost than the conservative community is concerned with the election being stolen from Trump.

Both sides had approximately 8% of their posts regarding the post-election discussion and analysis of vote totals, gains and loss in states and demographic groups (without alleging that the other party is interfering with those results). Among these posts directly discussing the election results, 2.66 times as many posts alleged fraud than didn't (205 STEAL vs 77 VOTE COUNT posts), again reinforcing the extreme level of worry about legitimacy of the outcome of the election.

In addition, we found that the conservative communities discussed more about "COVID", "FOREIGN" and "POLICY", compared to the liberal communities. These three topics are closely related to normal governing politics. The higher engagement of discussion in /r/conservative on these topics shows that the conservative communities concern more about the governing after the election, while the liberal communities concern more about the legitimacy of vote result.

Another observation is that there are notably more posts in the "PERSONAL" category from /r/politics than /r/conservatives - approximately 24% as opposed to 15%. Looking at the top words in this category, it's clear the main topic of discussion surrounds the transition of the presidency itself. A higher percentage of posts made by /r/politics potentially indicates a higher level of concern in this area. Again, it would make sense that liberal communities are more concerned with the transition process given that President Trump is unwilling to concede and instead alleges election fraud, as indicated by the "LEGAL" and "STEAL" figures. On the other hand, /r/conservatives had twice the number of posts in "MOB" then /r/politics, suggesting that

there are more instances of outrage/protests/threats involving the former than the latter. These findings also integrate well with our prior knowledge.

The TF-IDF results appear to be focused and most words have a strong correlation to the topic they belong to. The only notable discrepancies are the stopwords that are found in “FOREIGN” and “MOB”, as they are the only two categories to contain multiple stopwords. This could be due to the fact that only certain words were repetitive enough to be considered with all other words being infrequently used and more unique, resulting in the stopwords having a higher TF-IDF score.

Given the time constraints, we only did single annotation on posts. This can bring bias during the annotation phase due to coders’ knowledge about US elections. There is also more potential manual error produced during human coding in single annotation. To minimize the bias, we did some research about US presidential elections and we strictly followed the definition we made for each topic. If time permits, doing a double annotation would produce a more convincing result. However, the single annotation gives us enough confidence about the quality of annotation.

When attempting to extrapolate from a sample of Reddit supporters of the candidates as a whole, it is important to note that Reddit does not provide a good representation of the general public. According to the latest Reddit statistics, this site is most popular among users in the 25 to 29 age group (Marketing Charts, 2019). As many as 23% of US adults in this age range use Reddit, with 21 % in the 18 to 24 range, showing that Reddit is clearly more popular among young adults. From this point onwards, there is an inverse correlation between usage and age. Statistics show that 14% of US adults between the ages of 30 and 49 use Reddit, 6% for those from 50 to 64 and a mere 1% among US adults above 65 (Lin,2019). This indicates that there exists a significant bias in Reddit user’s age. As of February 2019, 15% of male respondents stated that they used Reddit. Whereas only 8% of females stated they use the website. Likewise, it’s safe to say that Reddit is popular with males in the US (Foundationinc,2020). This

means that the data collected from Reddit also has bias on gender. Therefore, Reddit posts provide an interesting angle to view the two communities, but they are not representative enough given these biases.

Group member contribution

Brian Hu:

- Filtering and trimming data
- Annotated 387 posts
- Data and Overview of report

Dawei Zhou:

- Prepare data for annotation
- Annotated 258 posts
- Calculate tf-idf and select top words
- Methods part of report

Logan Ralston:

- Collect reddit posts
- Annotated 258 posts and reannotated posts after revising typology
- Definition of topics
- Results part of report

The rest of the work was done together.

References

Lin, Ying. “10 Reddit Statistics You Should Know in 2020 [Infographic].” *Oberlo*, Oberlo, 19 Nov. 2020, www.oberlo.ca/blog/reddit-statistics.

Team, Foundation. “Reddit Statistics For 2020: Eye-Opening Usage & Traffic Data.” *Foundation Marketing*, 23 Nov. 2020, foundationinc.co/lab/reddit-statistics/.