COMP598 Final Project

Candidate Discussion in the Days Following the US Election – Group 49

David Garfinkle (260515841), Jean-Christophe Boivin (260789717), Anton Gladyr (260892882)

Overview

This project is a study to understand candidate discussion in the days following the US election, as well as perceptions of election legitimacy. Specifically, the goal was to answer the following questions: What are the salient topics discussed around each candidate and what each topic is primarily concerned with? What is the relative engagement with those topics among liberals and conservatives? The analysis was based on the Reddit posts from two subreddits: /r/politics and /r/conservative. According to the project requirements, we assume that these subreddits roughly line up with liberal and conservative US communities.

We annotate this data with a typology of eight topics and compare relative frequency of topics between subreddits. Given only posts that mention either Biden or Trump, these subreddits were most focused on the election process (29,8%), with liberal perspectives centering on election results and dismissal of fraud claims, and conservative perspectives revolving around ballot legitimacy and lawsuits in battleground states.

Data

The dataset used to produce the current report has been scrapped from the reddit API. More specifically, data was taken from the “hot” listing between November 29th and December 3rd. The hot listing was chosen because posts labelled as hot have a higher level of engagement from reddit users, ensuring that the collected posts were highly discussed at the moment of collection. Collection occurred every 24h so we could see “hot” topics evolve over time. In total, 4000 posts were collectected in the politics subreddit and 4000 in the conservative subreddits. We collected posts from these two reddits as they reflect opinions coming from

left-leaning and right-leaning communities in the United States. Duplicate posts arose because some reddit posts may classify as “hot” for more than one collection day. Keeping all duplicate posts in the analysis retains this trending information, but may skew the results and overweight duplicate posts. We filtered out as many duplicate posts as possible to cast a wider net of topics discussed in each community, and to prevent inflation of subtopics presented by duplicate posts. Duplicate posts were defined as two posts sharing an identical unique identifier.

From this dataset, we isolated posts that mentioned a presidential candidate. Any post that did not contain “Trump” or “Biden” was rejected. We also rejected posts that contained either candidate’s name if they were not referring to candidates themselves. For example, posts centered around Ivanka Trump or Hunter Biden were rejected. In total, after filtering, 745 posts were mentioning either or both presidential candidates. 223 posts were coming from the conservative subreddit, while 522 posts were coming from the politics subreddit. Since left-leaning communities were overrepresented in this dataset, we kept both datasets separate for content analysis.

Methodology

Posts were collected from the hot listing in both the politics and the conservative subreddits to see what topics were buzzing in the United States’ liberal and conservative communities. After removing duplicate posts and filtering for presidential candidates mentions, i.e. posts containing the words Trump and/or Biden isolated by non-alphanumeric characters, we proceeded to an open coding task to develop topics conducted by a single coder. For this task, 200 posts were used. Selection of these posts was made at random from a pool of candidate mentioning posts, leading to the open coding of 135 posts coming from the politics subreddit and 65 posts from the conservative subreddit.

After revision by two other coders, eight topics were approved as well-defined and comprehensive over the filtered dataset. All the coders participated in the subsequent topic

annotation. Each coder reviewed posts they had not already coded as to further define the topics, to maximize objectivity and to minimize intercoder variations. Annotations about the candidate mentioned and the subreddit of origin were also made in order to further analyze the properties of the discussions within a topic.

After annotations, the term frequency inverse document frequency (TF-IDF) was computed in order to see important terms within each topic. We used Eq. (1), Eq. (2), and Eq. (3), which were proposed by Prof. Derek Ruths:

$$TF(term \mid topic) = \text{the number of times the term appears in titles assigned to the topic} \quad (1)$$

$$IDF(term) = log\ (\ [\#\ of\ topics] / [\#\ of\ topics\ that\ term\ is\ used\ in]\ ) \quad (2)$$

$$TF\text{-}IDF = TF * IDF, \quad (3)$$

where the term frequency was defined as the number of occurence of a term within a given topic; the inverse document frequency was calculated using the log of the number of topics analyzed over the number of topics that had at least a mention of the term. Defining a document as a topic rather than individual posts enables to get a better picture of the core of each topic. As to further clarify the center of discussion within a topic, all the stop words were removed from the posts prior to calculation.

TF-IDF weights were calculated three times on this corpus. It was calculated over the entire data set, over just the politics subreddit, and over just the conervative subreddit. These calculations enabled comparison between liberal and conservative communities through the lens of top topic words. In order to not inflate the value of rare occurrences, terms had to occur at least three times to be considered. Statistical significance was computed using the Chi-Squared test and Fisher's exact test, with a bonferroni correction for multiple comparison (8), as these tests were most appropriate to the type of data we were dealing with.

Results

Political topics that were most discussed in the US after the 2020 elections are shown in Figure 1. Figure 2 and Figure 3 represent the distribution of the discussed topics in the /r/politics and /r/conservative subreddits respectively.

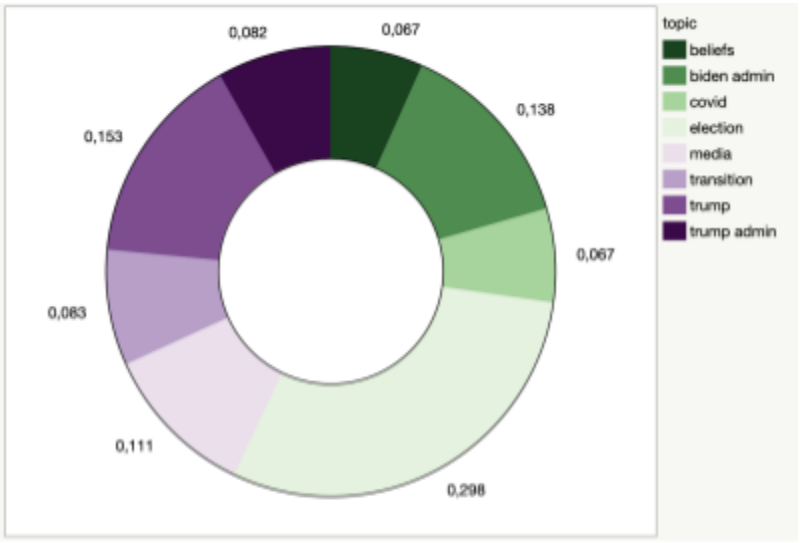

Figure 1. Distribution of discussed topics in both subreddits

First, the elections themselves were extensively discussed, representing the most discussed topic, with 29.8% (28% conservative, 30% liberal) of the collected posts discussing the topic. A post was classified as an election post if it made a direct reference to any part of the electoral process, including but not limited to the act of voting, the turnout, the number of votes received by each candidate, the (re)counting process, the lawsuits for electoral fraud and elections to come (e.g. the Senate runoff). In general, the discussions around this topic

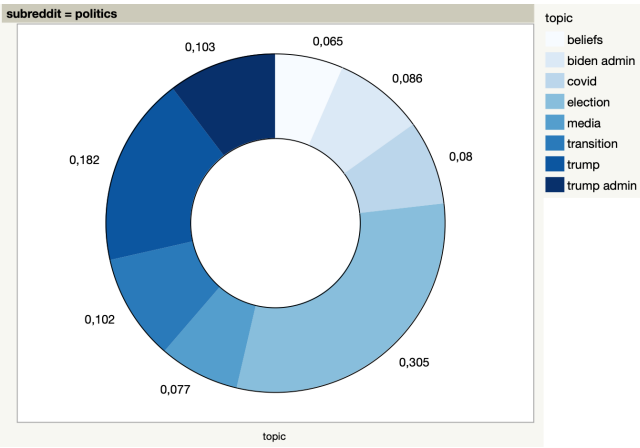

Figure 2. Distribution of the discussed topics in the /r/politics subreddit

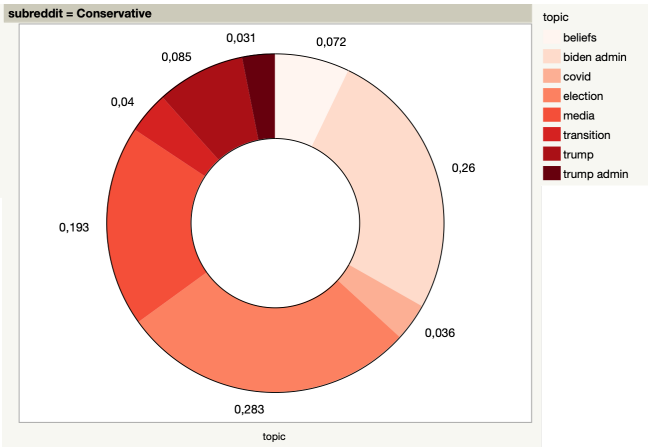

Figure 3. Distribution of the discussed topics in the /r/conservative subreddit

centered around the following terms, in order of importance : Wisconsin, ballots, vote, Pennsylvania, Arizona, results, certifies, fraud, assault, recount. Conservatives mostly centered their discussion around terms like “ballots”, “lawsuit”, “Winsconsin”, “Pennsylvania” and “fraud”, while liberals’ discussion was centered towards “wisconsin”, ”trump”,”results”, “recount”, “victory” and “Georgia” .

The second most discussed topic was Trump himself; posts entering this category had to refer directly to Trump’s action as a public figure rather than as a politician and to his traits of character. 15,3% of the posts entered this category, with a significantly higher proportion of involvement within the liberal community ($p = 0.0006$, 8.52% vs 18.2%), with the conversation involving terms such as: pardon, see, discussed, grift, Donald, fund, run, four, threats, and christmas. Liberals stirred their conversation around “pardon”, “Trump”, “possible”, “violence” and “dishonest”, while conservative used terms such as “hints”, “four”, “years”, “Donald” and “presidential”.

Third most important topic (13,83%) was the Biden administration; posts about the biden administration had to mention executive decisions made by Biden in regards to his presidency, make a direct reference to Biden’s decision as president-elect or to actions made by members of his cabinet. Conservatives were significantly more involved in the discussions about the Biden administration, with 26.01% of conservative posts making mention of the Biden administration versus 8.62% of liberal’s posts ($p < 0,0001$). Most of the discussion around this topic involved terms such as: “pick”, “Biden”, “OMB”, “Tanden”, “Neera”, “deletes”, “tweets”, “bathroom”, “lockers”, “room”. While liberals focused on terms such as : “economic”, “neera”, “relations”, “taps” & “tanden”, conservatives mentioned: “pick”, “Biden”, “omb”, “China”, “tweets” & “deletes”.

With 11,14% of posts discussing the topic, media coverage came fourth in terms of importance. Media posts had to directly refer to mediatic outings, mediatic representation of candidates, censorship, behaviour of a mediatic entity (e.g. a news reporter, a news channel,

etc…), social media as an entity or refer to elements that affects mediatic empires. Conservatives were relatively more involved in this topic (19.28%) than liberals (7.66%), a difference that was significant ($p < 0,0001$). Conservatives focused on “cnn”, “hunter”, “press”, “story” and “suppression”, while liberal discussions were mostly concerned with “veto”, “defense”, “fox”, “media” and “tech”.

Fifth in terms of importance, the transition from Trump to Biden as president was discussed in 8,32% of the collected posts. Transition posts have to make reference to a potential change of president. These posts are more focused on the outcome of the presidential election than the electoral process. They may also include a shift of administration and its consequences. Any contrast between administration and party behaviour, e.g. mentioning that the Republican party will act with hostility towards the Biden Administration, is sufficient to be classified as a transition post. There wasn’t a significant difference between liberals’ involvement in the discussion (10.15%) and conservatives’ (4.04%), though liberals were relatively more involved in the topic. Discussion surrounding this topic involved terms such as transition, nato, Fauci, Biden, ambassador, ajit, pai, poland, scraps, fort, without reddit-specific salience of specific terms.

Sixth most important topic was the Trump administration, defined as any post making a direct reference to Trump’s decision as president, extending to decisions made by the US government under Trump’s presidency. 8.19% of the posts were classified as belonging to this topic. Discussions over the Trump administration were mostly coming from the politics subreddit (10.34% versus 3.14% of the conservative subreddit’s post; $p = 0,0004$). Liberal discussions in this topic frequently used terms like “veto”, “arctic”, “refuge”, “defense”, and “pardon”, while conservatives were posting about “census”, “remove”, “legacy”, and “groundwork”.

Sharing the last spot on this ranking, the coronavirus pandemic and political beliefs managed to gather 6,71% of the public’s involvement in discussions surrounding presidential candidates. To belong to either topic, a post had to respectively make reference to the ongoing

pandemic, or refer to political beliefs, attitudes of a political entity such as party members or trump supporters, or speculations about a political entity's behaviour. Conservatives and liberals were equally involved in political beliefs topics, and liberals were more active discussing COVID-19. The liberal r/politics subreddit often used terms like "covid", "response", "scott", while r/conservative leaned more into terms like "plans", "undocumented", and "vaccination" to discuss the pandemic. To discuss political beliefs, r/conservative subreddit used terms such as "freeze", "box", "democrats", "fear" and "socialism" while r/politics centered its beliefs' discussion around "fooled", "trump", "conspiracy", "right", "home" and "theories".

Discussion

What are the prevailing perspectives on election legitimacy? The salient subjects drawn from the election topics include: ballot counting, battleground states, mail-in ballots, election victory, and the lawsuits to contest that victory. Liberals and conservatives alike are engaged with all of these topics to the highest relative degree, but topic extraction illustrates differing focal points. Over the five-day collection period, the r/politics subreddit was most concerned with the process in Wisconsin as it recounted and finally certified its election result. On the other hand, the top five topic words in the r/conservative subreddit focus on mail-in ballots and their validity, in addition to a greater focus Wisconsin, and Pennsylvania where lawsuits are still continuing. The conservative subreddit contests the voting process more than it discusses the reported result. It's hard to say from this one statistic whether one camp views the election with more legitimacy than the other, but this measurement does suggest that the liberal viewpoint is focused on the reported (re)counts and how these counts solidified Biden's victory, while the conservative viewpoint is more about the fraud involved in the counting process, and how the ballots received by mail that were majoritarily for Biden in Pennsylvania act as a proof for such claim.

We see this trend again in the discussion surrounding political beliefs. The salient terms for this topic really shows this sentiment of being “fooled” by the opposite camp. By analyzing the content of liberals’ reddit discussion, we see that they make the claims about election fraud is a part of a delusional conspiracy theory, while conservatives claim that we should “fear” the socialism brought by the democrats’ win.

The media topic has a theme on misinformation. On one hand, the liberal subreddit collects posts about Trump broadcasting misinformation. The top words in this subreddit reference Article 230, the censorship of Trump’s fallacious claims on social media presented as positive, and Trump’s intent to revert the Article. On the other hand, the conservative subreddit is more about media misinformation via *suppression*, i.e. by selectively choosing topics that misrepresent the truth. It’s interesting that through this topic, we see both subreddits engage in the idea of misinformation, but with opposite perspectives. It also shows that conservatives are more likely to believe the allegation of Trump, whereas the liberals are worried about the threat that he represents for objective, neutral and factually accurate media. This idea of Trump posing a threat is actually widespread in the liberal subreddit; on the topic of Trump as a public figure, liberals readily qualify him as a dishonest grifter that uses false claims about the election to gain both money and popularity amongst conservatives. They also depict him as a someone that will abuse his powers, discussing preemptive pardons for his family and himself and enticing violence. À contrario, conservative more readily trust the man, as their discussion gravitates around the man presenting himself again at the 2024 election.

Interestingly, both political parties were actively discussing Biden administration’s picks. However, conservatives were mainly concentrated on Neera Tanden’s persona – Biden’s pick for budget chief. Namely, they were concerned with Tanden’s tweets criticizing Republican Party leaders. Furthermore, another concern of the conservatives towards Biden’s administration was translated by the overrepresentation of “China” in their discourse. More specifically, as China represents a greater evil for them (Gries & Crawson, 2010), conservatives were readily claiming

that a Biden Administration be under chinese influence as a way to undermine its authority. At the same time, liberals were involved in discussions of the Trump administration's attempts to sell leases in the Arctic National Wildlife Refuge. The results of TF-IDF computations corroborate these conclusions having "biden", "pick", "omb", "arctic", "refuge", "sell" words as the most significant in the "biden admin" and "trump admin" topics.

It is clearly seen that liberals are more concerned in topics related to COVID-19 and pandemic itself. They use words such as "covid", "coronavirus", "pandemic", "vaccine", etc., more frequently than conservatives. But at the same time they tend to discuss or criticize the political opponents in frameworks of this topic. For example, TF-IDF computations show us that the most important words in the "covid" topic are related to Trump's ex-adviser – Dr. Scott Atlas who resigned in the beginning of December.

Group Member Contribution

Jean-Christophe Boivin wrote the data collection script. Anton Gladyr proceeded in collecting the data and wrote the data cleaning script. Jean-Christophe Boivin did the open coding of 200 posts to develop our topics for the sake of this analysis. Anton Gladyr and David Garfinkle reviewed the initial open coding to further define the topics and assess their validity. All team members participated in the coding of posts mentioning presidential candidates. David Garfinkle wrote the Term Frequency - Inverse Document Frequency script on the coded dataset. All three team members proceeded to the analysis of the data. Jean-Christophe Boivin generated the figures and did the statistical analysis for the topics' relative importance depending on political affiliation. Team members contributed to writing the report in the following manner: Jean-Christophe Boivin wrote the Data, Method and Group Member Contribution sections, whereas David Garfinkle and Anton Gladyr wrote Overview and Discussion sections. All team members reviewed each section and made corrections and additions when necessary.

Reference

Gries, P. H., & Crowson, H. M. (2010). Political orientation, party affiliation, and American attitudes towards China. *Journal of Chinese Political Science*, *15*(3), 219-244.

Appendix

Calculated TF-IDF results for each topic:

==========

trump admin

veto 0.007039363791464622
arctic 0.00514171117941393
refuge 0.00514171117941393
defense 0.004427856094784946
census 0.003519681895732311
groups 0.003519681895732311
uae 0.003519681895732311
sell 0.003519681895732311
groundwork 0.00342780745294262
firing 0.00342780745294262

==========

others

police 0.0033166330516253314
mayor 0.0023930735343617284
covid 0.0018195643722216906
stimulus 0.0017181040759520103
bipartisan 0.0016567432160965814
residents 0.0016134971602501612
cnn 0.0015953823562411525
lawmakers 0.0015340214963857238
dominion 0.0015340214963857238
voting 0.0015238584291251522

==========

biden admin

biden 0.005810927755343397
pick 0.004616301041079671
omb 0.004281055209040628
neera 0.004169306598360948
tanden 0.004169306598360948
deletes 0.004169306598360948
tweets 0.004169306598360948
bathrooms 0.004169306598360948
locker 0.004169306598360948
rooms 0.004169306598360948

==========

beliefs

box 0.004694924310547478
fooled 0.004694924310547478
freeze 0.0032138405914022956
socialism 0.0032138405914022956
green 0.0032138405914022956
home 0.0025991353083856692
lies 0.0025991353083856692
right 0.0025991353083856692
supporters 0.0025119088243680306
democrats 0.0025119088243680306

==========

transition

transition 0.009977296164353394
nato 0.005465732779443332
fauci 0.004988648082176697
biden 0.004394889390163561
ambassador 0.003741486061632523
ajit 0.003741486061632523
pai 0.003741486061632523
poland 0.0036438218529622215
scraps 0.0036438218529622215
fort 0.003643821852962215

==========

media

veto 0.009212055325467157
cnn 0.007537136175382221
unless 0.006699676600339751
hunter 0.006116994925768985 5
defense 0.005418241979173689
media 0.005236406814272775
story 0.004893595940615188
section 0.004893595940615188
suppression 0.004893595940615188
editor 0.004893595940615188

==========

covid

atlas 0.014225059514626921
adviser 0.013130824167347925
scott 0.012036588820068932
resigns 0.009848118125510945
covid 0.009692393022505921
coronavirus 0.008884693603963762
dr 0.007659647430952957
response 0.006565412083673962
fail 0.004494255169650222
vaccine 0.004376941389115975

==========

trump

pardon 0.004119999987631675
see 0.004106961826796672
discussed 0.003080221370097504
grift 0.003080221370097504
donald 0.002842045150290872
fund 0.0028113596201425686
run 0.0026525754736048137
four 0.0026525754736048137
threats 0.0026525754736048137
christmas 0.00256685114174792

==========

election

wisconsin 0.006108831676312888
ballots 0.003665299005787733
vote 0.003607340819467655
pennsylvania 0.0031765924716827014
arizona 0.0030662396965475064
results 0.0030662396965475064
certifies 0.002932239204630186
fraud 0.002705505614600741
assault 0.0023417575127744 48
recount 0.0023417575127744 48