COMP 598: Final Project Report

Pingsheng Li 260906286 (pingsheng.li@mail.mcgill.ca)

Hao Lee 260855685 (hao.lee3@mail.mcgill.ca)

Ying Wang 260778562 (ying.wang14@mail.mcgill.ca)

1 Overview

Two weeks after the 2020 US election, the election result was still a popular topic on Reddit, but the doubt about the election legitimacy gradually faded after Georgia's recount and the lawsuit failure of Trump's team. When discussing Trump, it's hard to avoid law-related keywords like "fraud" and "legal"; whereas Biden was already perceived as the winner of the election and frequently mentioned in discussion about the "transition" and new "administration". Furthermore, liberals are more willing to get involved in all candidate-related topics than conservatives.

2 Data

We collected around 2,000 Reddit posts for r/politics and 2,500 posts for r/Conservativein on Nov 17&20&21 2020, and then randomly selected an equal amount of posts in each subreddit on each single day to form a dataset of 2000 posts. Nearly half of the 2000 posts mentioned "Trump" or "Biden".

2.1 Ranking Algorithm

We focused on the "hot" ranking because it balances time effect and popularity. Although Reddit's "hot" ranking algorithm has been changed several times over the past decade, it roughly follows the formula given below when the number of upvotes are more than that of downvotes:

$$log(|U - D|) + A/4500,$$

where *U*, *D* and *A* denote the number of upvotes, that of downvotes and the age of the post respectively (Michel Billard 2019).

By comparison, other ranking methods like “new” and “top” have their limitations. “new” refers to the latest posts without considering Redditors’ attitude towards the post. Hence, even though “new” posts are more likely to be closely related to the recent election, they don’t necessarily reflect the topics that are primarily discussed on Reddit. On the contrary, “top” simply displays the posts with highest scores, resulting in a bias towards early posts. This ranking has two disadvantages (i) It might give us early posts that are irrelevant to the recent US election; (ii) It will produce a large number of redundant posts if we collect data over a short period of time (less than a week).

2.2 Sampling

Instead of scraping over consecutive three days, we collected Reddit posts on a random weekday (Nov 17, 2020) and a weekend (Nov 20 & 21, 2020) according to a research about media sampling approach: *“Constructed Week Samples Tend to Produce More Efficient Estimates Than Consecutive Days Samples”* (Hester & Dougall, 2007). Note that more emphasis was placed on weekends because it tends to be a better time for people to read news and cast upvotes/downvotes.

To construct a more representative dataset, we collected as much as possible Reddit post titles at approximately the same time over the above mentioned three days. Due to constraints of the Reddit API, we could only get around 700 hottest posts for r/Conservative and 850 hottest posts for r/politics daily, which added up for about 2,000 posts for r/politics and 2,500 posts for r/Conservative. After removing the redundant posts, we randomly select 333, 333, 334 posts from each subreddit on each single day, resulting in a dataset of 2000 posts (referred as *original dataset*)

2.3 Filtering

To facilitate the later analysis of candidates, we constructed a *candidate dataset* from the *original dataset* by only keeping the posts mentioning “Trump” or “Biden” (case sensitive), decreasing the original into nearly a half. The filtered dataset consists of 934 candidate-related posts, where 646 posts (59%) were from r/politics and 288 posts (29%)

were from r/conservative (Figure 2.1). It's also worth noting that "Trump" was mentioned almost twice as frequently as "Biden" in both subreddits (Figure 2.2).

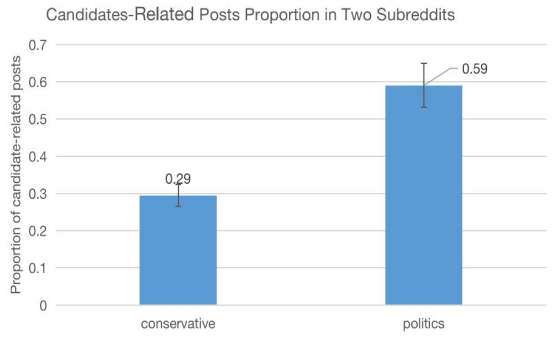


Figure 2.1

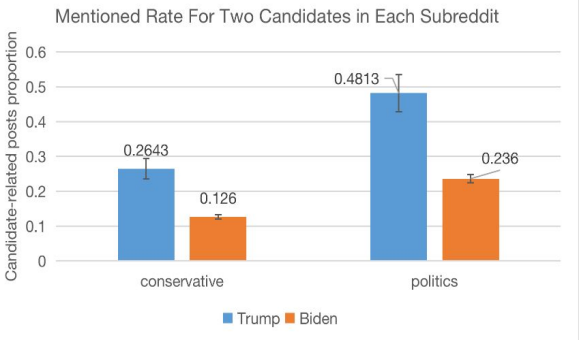


Figure 2.2

3 Methods

3.1 Open Coding & Annotation

We used the following open coding procedure to develop a profound and comprehensive typology for candidate-related posts. First, we took a sample from the *candidate dataset* by randomly picking 200 posts (66:66:67 for three days) from each subreddit, resulting in a subset of 400 selected posts. Next, each team member went through the subset and came up with potential topics individually. After everyone had formed ideas on the categorization, all three team members exchanged opinions and brainstormed on the typology. Finally, we developed the *Typology Design* document by reviewing each topic, removing objectivity from definitions and providing positive/negative/edge examples. (See *4.1 Typology* for the results)

Due to the time constraints, we adopted the single annotation approach and assigned one third of the dataset to each team member. To reduce the randomness of single annotation, we discussed ambiguous cases together and randomly selected several relevant posts that were annotated by others to justify. In the meantime, we also refined our *Typology Design* through the discussion of edge cases.

3.2 tf-idf Calculations

Tf-idf (Term Frequency-Inverse Document Frequency) was used to spot the key words for different groups of Reddit posts. To produce more informative results, we first

discarded stop words using the NLTK library in Python and candidate names including "Trump", "Donald", "Biden" and "Joe", and then selected the top 10 alphabetical words with the highest tf-idf scores in each group. These techniques led to good performance in practice (see *4.2 Topic Characterization* and *5 Discussion*) and the two versions of tf-idf are as follows:

(i) In the interest of **topic** characterization, we used the tf-idf scores defined by the product of the following two statistics:

$$tf(w,\ T)\ =\ the\ raw\ count\ of\ the\ word\ w\ in\ the\ topic\ T \quad \text{and} \quad idf(w)\ =\ log(\ N/N_m\ ), \qquad (\Delta)$$

where N and N_m denote the number of topics (7 in our case) and the number of topics in which the word *w* is used (i.e, N_m is dependent of the word w).

(ii) To characterize **subreddits** or **candidates**, we should change our definitions for the tf-idf because we only have two groups (politics & conservative, Biden & Trump) now. Otherwise, most of the words may have identical idf values, degrading the tfidf calculations to a pure counting problem. Therefore, the following tf-idf definition is preferable

$$tfidf(w,\ C,\ D)\ =\ tf(w,\ C)\ \times\ idf(w,\ D)\,,$$

where $tf(w,\ C)$ means the raw count of the word *w* in the category C, and $idf(w,\ D)$ is defined by

$$idf(w,\ D)\ =\ log\ (N^*/N_m^*),$$

where N^* and N_m^* denote the number of all alphabetical words in the dataset consisting of 2000 posts and the number of occurrences of the word w in the dataset respectively.

4 Results

4.1 Typology & Topic Characterization

The definitions and characterizations of our 6 topics are summarized in the table below (see the *Typology Design* in the appendix for more details). The characterizations of each topic are depicted by the top 10 words with the highest tf-idf scores (version i).

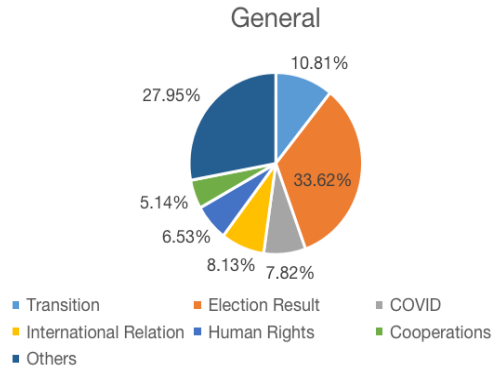
Topics	*Definitions*	*Characterization** *(Keywords in descending order)*
Transition (101 posts)	Posts that focus on the administration transition	cabinet(12/0), picks(9/0), transition(26/7), blocking(5/0), speech(1/5), concede(9/0), process(4/0), merrick(4/1), veterans(4/0), garland(4/1)
Election Result (314 posts)	Posts that focus on anything that is about, has influenced or will possibly influence the election results (the 2020 US presidential election)	georgia(29/20), recount(17/10), votes(8/13), overturn(12/3), wisconsin(10/3), certify(12/4), certifies(11/3), pennsylvania(15/4), lawsuit(15/5), election(89/30)
COVID (73 posts)	Posts that focus on COVID-19/coronavirus/the pandemic	covid(26/14), coronavirus(12/3), tests(3/3), positive(3/3), vaccine(3/6), attended(3/0), monumental(3/0), retreats(3/0), urged(4/0), fda(2/2)
International Relations (76 posts)	Posts that focus on foreign countries, international forums as well as global cooperation	israel(4/3), ties(1/4), iran(10/0), troop(2/4), foreign(1/4), china(3/5), loudest(4/0), middle(3/2), hunter(0/5), session(3/0)
Human Rights (61 posts)	Posts that focus on anything related to human rights, including issues around refugees, race, LGBTQ+, climate change, gender, drug price and etc	prescription(1/6), climate(6/2), fix(4/3), costs(3/1), black(5/2), latino(3/2), extremist(2/0), closely(2/1), immigrants(2/3), aoc(2/1)
Corporations (48 posts)	Posts that focus on the action or announcement from companies, including media, with an emphasis on the influence on candidates	account(9/0), twitter(9/0), media(2/8), obliterates(0/4), peaceful(0/3), covid(4/0), accounts(3/0), tv(3/0), hypocrisy(1/2), cnn(0/3)
Others	Any posts that don't belong to any of the above categories.	

Table 4.1 Topics and Keywords

**Note that the bracket following the keywords represents (#liberal posts/ #conservative posts)*

4.3 Topic Engagement

The pie chart in Figure 4.2 illustrates the most popular topic on Reddit is **Election Result** (33.62%), taking up one third of the discussion around the candidates. The proportions of the rest five topics are in the range of 5%-11%, indicating Redditors express a same degree of interest towards the influence of the election on other essential topics.



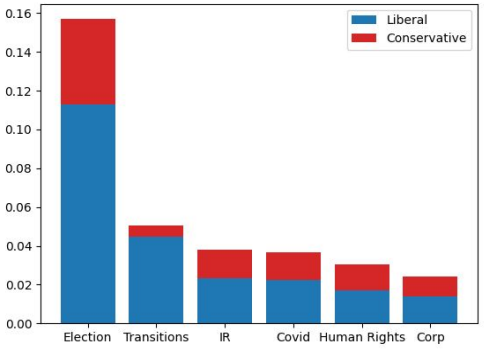


Figure 4.2

Figure 4.3

Liberal vs Conservative: The stacked bar chart in Figure 4.3 reveals the different focuses of the two subreddits, r/politics(representing liberals) and r/conservative (representing conservatives). Not surprisingly, r/politics contributes more in all topics because it has about as twice as many candidate posts in r/conservative(Figure 2.1). Even considering the different proportions of candidate posts, we can still observe an overwhelming enthusiasm towards **Transition** in r/politics compared to r/conservative, suggesting liberals' interest in the ongoing presidential transition of Biden, the democratic candidate. Nonetheless, both subreddits pay most attention to **Election Result** and consider the rest four topics almost equally important.

Biden vs Trump: The pie charts in Figure 4.4&4.5 show the different topic engagement in Reddit posts mentioning "Biden" and "Trump" in their titles. A great proportion of discussion on Trump was placed on **Election Result** (38.09%) while Biden are mentioned frequently in both **Election Result** (21.74%) and **Transition** (20.55%).

4.4 Outliers

It's worth mentioning that about 28% of posts are grouped to **Others** because they are irrelevant, ambiguous or hard to group. Some of the posts were indeed about Ivanka Trump instead of the presidential candidate we were interested in. Other posts focused on different aspects of Trump's and Biden's personal life, reputation, relationship with other politicians that were not directly related to the recent US election and were hard to group. There were also a number of post titles that were too ambiguous to infer its meaning.

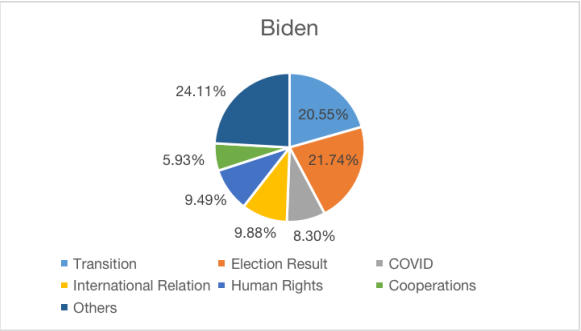


Figure 4.4

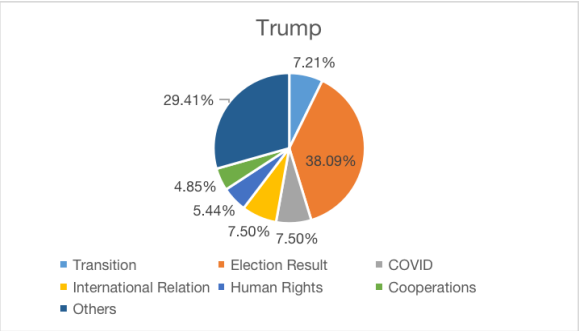


Figure 4.5

5 Discussion

5.1 Biden vs Trump

As discussed in *4.3 Topic Engagement*, the most important topic discussed around Trump is the **Election Result**, especially the result in Georgia and relevant lawsuits(Table 4.1), while the most important topics discussed around Biden are the **Election Result** and **Transition**. For other essential topics like **Human Rights**, **International Relation**, **COVID**, and **Corporations**, Biden was discussed more than Trump, which implied that Biden was perceived by the majority of Redditors as the actual winner of the election and the future decision maker around these topics.

This conclusion is also supported by the top 10 words with the highest tf-idf scores for each candidate (formula ii), as shown in the table below. These two bar charts clearly illustrate that the discussion of Trump was around the election election legitimacy with keywords like “legal”, “fraud” and “lawyer”, whereas Biden was perceived as the winner of the election with keywords like “win”, “certifies” and “victory”. The keywords related to transition (“transition”, “staff”, “administration”, (white) “house”) frequently appeared in posts mentioning Biden, which reflected Biden’s move into the white house.

Candiate	*Keywords*
Trump	election, gop, michigan, president, legal, campaign, claims, fraud, results, lawyers
Biden	transition, georgia, win, election, white, staff, administration, house, certifies, victory

Table 5.1 Candidate Characterization

5.2 Liberals vs Conservatives[1]

By analyzing the posts where the keywords appear and counting the number of posts about this word posted in r/politics and r/Conservative respectively (Table 4.1), we are able to infer the perception of liberals and conservatives on each candidate. In summary, despite a higher mentioned rate of Trump among posts from both subreddits, discussion of all essential topics for U.S. society has been generally shifted to Biden, which indicates both conservatives and liberals have assumed Biden is the winner of the 2020 election, and liberals take leads in the participation among all selected topics.

Election Results: Generally, the majority of both subreddits admitted the certification of election results despite very few disagreements in conservatives. In particular, liberals overwhelmingly participated in topics about the certification of Biden's victory after the recount in Georgia and the failure of Trump's attempt to overturn the result in Wisconsin and Pennsylvania. Nevertheless, conservatives discussed more about whether the number of votes was incorrectly counted whereas liberals emphasized that Biden had obtained the highest number of votes in history, confirming Biden's winning of the election.

Transition: This topic was dominated by liberals. They were primarily concerned about cabinet picks in the process of presidential transition, and especially interested in the appointment of attorney general: Merrick Garland. Liberals also concentrated on criticizing Trump for his refusal to concede, which blocked the normal power transition to Biden's administration.

Corporations: Most of the posts were about the media, however different subreddits focused on different aspects. Liberals were concerned about the transfer of the government official twitter account from Trump to the Biden administration, indicating a gradual peaceful power transition. By comparison, conservatives generally expressed more negative sentiments towards the media. Many of them criticized the media, like CNN and TV, and supported to obliterate democrats and media for their hypocrisy, and some conservatives

[1] Keywords obtained from tf-idf (see Table 4.1) are in orange

claimed that the unfair media was destroying Trump while covering Biden, even maybe intentionally delaying the news of COVID vaccine until after the election.

Human Rights: Conservatives heavily praised Trump's contribution towards lowering the costs of prescription drugs, whereas liberals strongly emphasized Biden's plans on fixing the climate issue with the support from AOC. Besides, conservatives argued that there is a great surge of support for Trump in the Latino community since they consider themselves as working class and Trump spoke for them, while liberals stated that the black community and BLM supporters played a major role in the winning of Biden because they consider Trump as a racist, and liberals argued that anti-immigrants extremist were closely related to the trump administration.

International Relation: Both conservatives and liberals extensively participate in the discussion about Israel-US relations and wars in the middle-east. Besides, liberals focused on the attacks on Iran's nuclear sites planned by Trump's administration. However, conservatives proposed that Trump decided to reduce the troops in Afghanistan and Syria, bringing soldiers home, denouncing Biden for his ties to foreign political power like Chinese and Russian government, suggesting Biden would pivot back to foreign policy failure. Liberals challenged conservatives by arguing the loudest attacks on US votes are from Trump, not Russia, and blaming Trump's indifference to the global crisis for skipping the G-20 COVID-19 pandemic session.

COVID: Liberals were overwhelmingly concerned about how Biden will fight against the pandemic without the help from Trump, what treatments are authorized by FDA, and how vaccines will be distributed, whereas conservatives focused on how Biden will damage the economy by national lockdown, which Biden has vowed not to do. Liberals also excoriate Trump's monumental sulk and indifference when he retreated from the public as COVID still ramplages in the U.S.

6 Group member contribution

Pingsheng Li: Data collection, Annotation, Visualization, Interpretation of the tf-idf results

Hao Lee: Data cleaning, Annotation, Calculations of tf-idf Scores, Visualization

Ying Wang: Typology design, Annotation, Result analysis, Finalization of the report

7 References

- Saliherfendic, A. (2015, Dec 9). *How Reddit ranking algorithms work.* Retrived from https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9
- Michel Billard (2019, Sep 9). *An introduction to ranking algorithms seen on social news aggregators.* Retrived from https://coderwall.com/p/cacyhw/an-introduction-to-ranking-algorithms-seen-on-social-news-aggregators
- Hester, J. B., & Dougall, E. (2007). The efficiency of constructed week sampling for content analysis of online news. *Journalism & Mass Communication Quarterly*, 84, 811–824. doi:10.1177/107769900708400410

Appendix

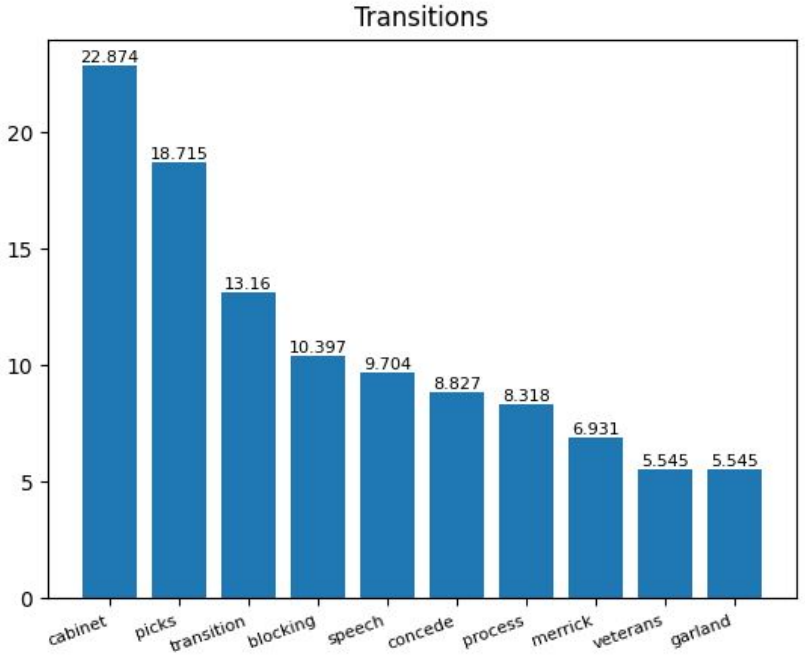
1. Typology Design

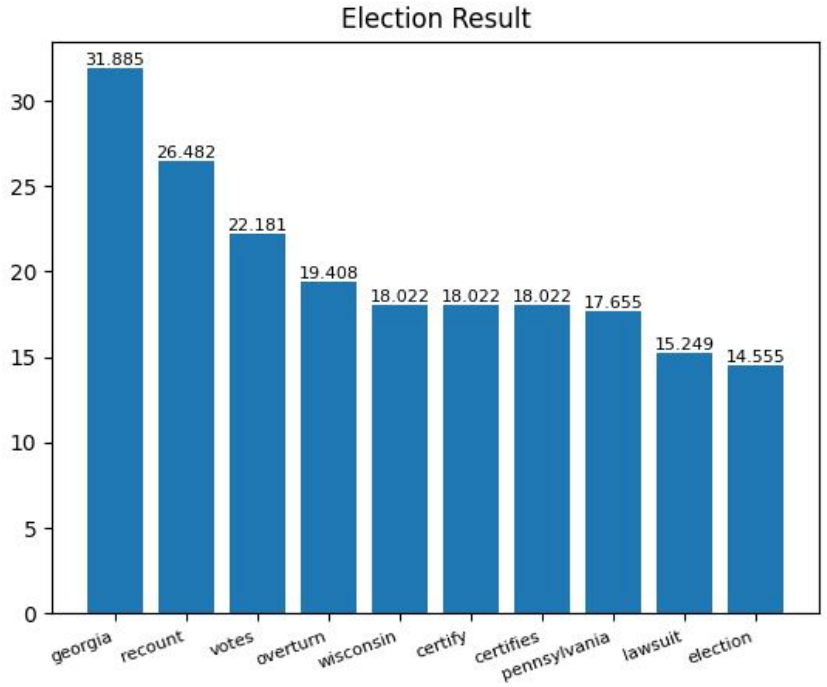
(a) Motivation
This

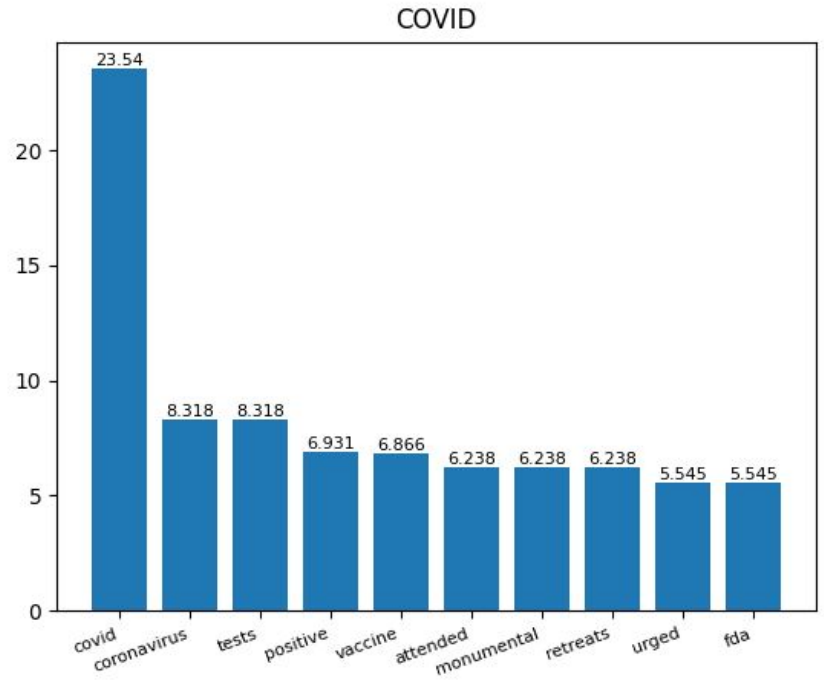
(b) Overview

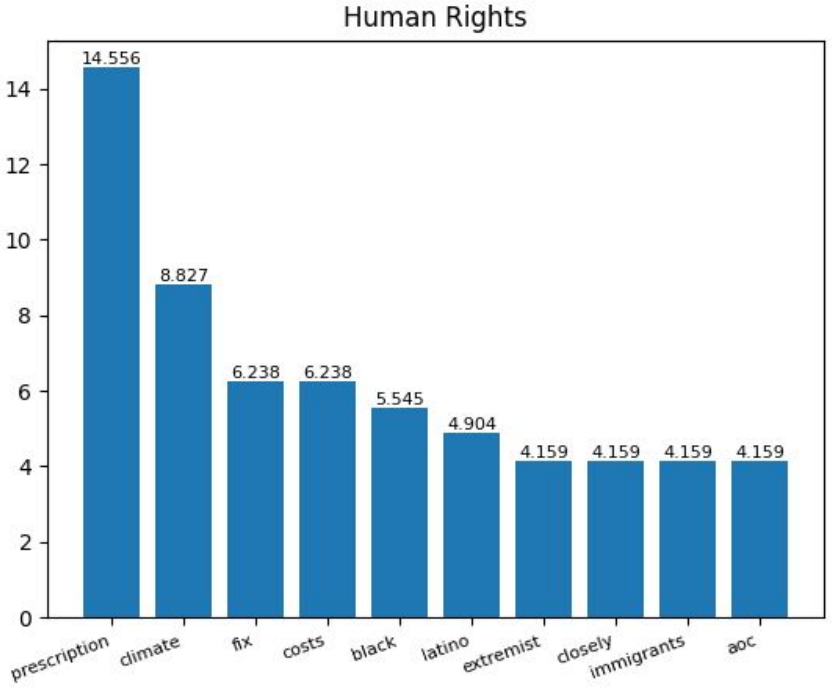
Topics	Positive example	Negative example	Edge case
Transition	Trump Is Trying to Lease Arctic Land for Drilling Before Biden Takes Over	Biden to Impose $200 Gun Tax | Americans for Tax Reform (The topic **Transitions** is only about administration transitions.)	Obama to Trump on Conceding Election: 'Think beyond your own ego' (Obama is pushing the administration transition.)
Election Result	Conservative Radio Host Got Less Than 24 Hours to Win Trump Campaign Lawsuit That He Knows Won't Change Election Outcome	Trump Team Making False Argument about his 2016 election (We only focus on the 2020 US presidential election.)	'Everything for Our Whole Family Hinged on This Election': International Students and Graduates Respond to a Biden Win (It's the attitude towards the election from international students' perspectives.)
COVID	Biden wades into coronavirus relief fight		Biden transition team prepares to fight covid with no help from Trump (The post focuses on how to fight COVID instead of the administration transition.)
IR	Pompeo touts Iran policy in Gulf ahead of Biden presidency	By the Way, Donald Trump Could Still Launch Nuclear Weapons at Any Time: The president's responsibility for the US nuclear arsenal is a Cold War anachronism. The Trump era shows why it needs reform. (The post seems to be related to Russia as it mentions the Cold War, but it's in fact a media critique.)	Trump Skips G20 Pandemic Preparedness Meeting as Covid-19 Cases Surpass 12 Million in US (The post concentrates on the G20 meeting which is an international event.)
Human rights	AOC and Cori Bush Join Protest Outside DNC to Push Biden on Climate Action (States have a human rights obligation to prevent the foreseeable adverse effects of climate change and ensure that those affected by it.)		How Members of Anti-Immigrant Extremist Groups May be against the Biden Administration (The post mentions "the Biden Administration, but it mainly focuses on the human rights concerning immigration.)
Corp	Twitter will transfer presidential accounts to Joe Biden on Inauguration Day	'I'm Joe Biden and Why Am I on This Magazine?' (The post doesn't seem to be highly related to any specific media.)	Silicon Valley eagers for Biden to reverse Trump visa rules, hire more immigrants (Although the post is related to the human rights about immigration, it

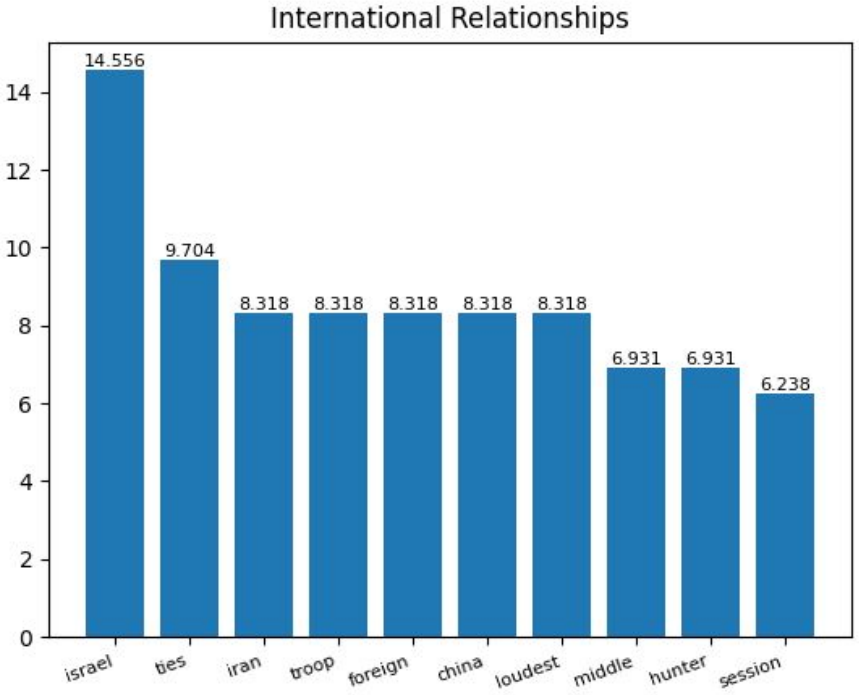
			focuses on companies' human resource planning.)

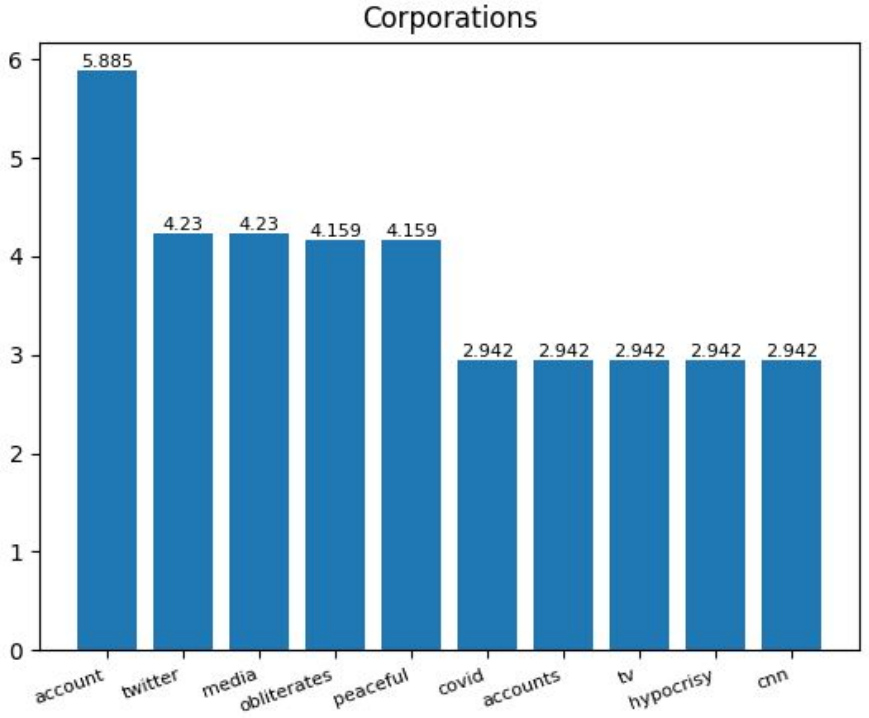





COVID
23.54
8.318
8.318
6.931
6.866
6.238
6.238
6.238
5.545
5.545
0
5
10
15
20
covid
coronavirus
tests
positive
vaccine
attended
monumental
retreats
urged
fda


Human Rights
14.556
8.827
6.238
6.238
5.545
4.904
4.159
4.159
4.159
4.159
0
2
4
6
8
10
12
14
prescription
climate
fix
costs
black
latino
extremist
closely
immigrants
aoc


International Relationships
14.556
9.704
8.318
8.318
8.318
8.318
8.318
6.931
6.931
6.238
0
2
4
6
8
10
12
14
israel
ties
iran
troop
foreign
china
loudest
middle
hunter
session


Corporations
5.885
4.23
4.23
4.159
4.159
2.942
2.942
2.942
2.942
2.942
0
1
2
3
4
5
6
account
twitter
media
obliterates
peaceful
covid
accounts
tv
hypocrisy
cnn