COMP – 598: Introduction to Data Science

GROUP – 50: Project Report

Members:

Grover, Harmanpreet Singh **harmanpreet.grover@mail.mcgill.ca**

Simas, Tristan **tristan.simas@mail.mcgill.ca**

Miao, Yan **yan.miao@mail.mcgill.ca**

1. Overview

In this report, we investigate the perception of election legitimacy by analysis of reddit posts collected from the "/r/politics" and "/r/conservative" subreddits. These posts were manually annotated for topics related to election legitimacy perception and for candidate preference. These annotations combined with various analysis methods have led us to the following findings: Posts concerning Trump are highly associated to discussions raising suspicion of election illegitimacy as well as evidence of his interference with the election process. Furthermore, we found that posts concerning Biden were associated to trust in the election's legitimacy as well as discussion about topics unrelated to election legitimacy. We propose that the liberal sentiment bias from our post dataset combined the two weeks of time passed between our data collection and Biden's announcement as President elect are major contributors to the strong association between Biden posts and perception of election legitimacy.

2. Data

As we are trying to understand people's perception of election legitimacy, we collected a total of 2000 Reddit posts from two subreddits: "/r/politics" and "/r/conservative" over a 3-day period from 21st to 23rd November 2020. On the first date, we collected 400 posts from each subreddit. On the second date, we collected 300 posts from each subreddit. Similarly, 300 posts were collected from each subreddit on 23rd. We believe that we collected maximum number of posts

on the first date since there is likely more election related posts on a date closer to the election result declaration date. As we are concerned only about election related posts, we removed all posts without mention of Trump or Biden. We defined mention as the occurrence of case-insensitive Trump and/or Biden substring anywhere in the post title. After filtering our 2000 posts, we were left with 966 posts.

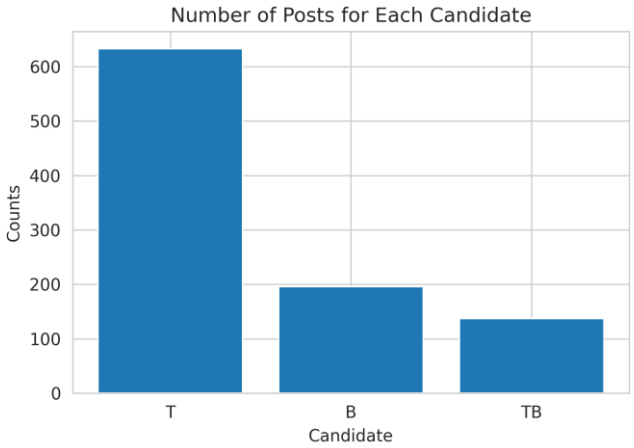


Figure 1: Plot showing the count of posts having **Trump (T)** mention, **Biden (B)** mention and **Trump & Biden (TB)** mention

3. Methods

Methods in this project can be summarized as the following: main contents of each topic were characterized identifying the top 10 keywords ranked TF-IDF for each topic, topic trends were analyzed over the three-day period with time series and word-topic correlation were analyzed using named entity analysis.

TF-IDF

TF-IDF for words per topic is defined as follows:

$$\begin{aligned} & TF-IDF \\ &= TF(term|topic) * IDF(term) \\ &= (\#\text{ of times the term appears in titles assigned to the topic}) * log(\frac{\#\text{ of topics}}{\#\text{ of topics that term is used in}}) \end{aligned}$$

The resulting top words ranked by TF-IDF scores revealed what discussion around each topic primarily concerned. This was done for each mentioned candidate and for the whole corpus of post titles.

Time Series

A time series is a collection of quantitative observations that are evenly spaced in time and measured successively [1]. Since our data contains posts over a span of three-day period, we were interested to know if there exists certain underlying trend in it. On a day-to-day basis, we explored the variation of number of posts belong to each topic to obtain their temporal popularity.

Named Entity Analysis

Named entity recognition (NER) is formally defined by (Grishman and Sundheim 1996; Chinchor 1997; Sang and Meulder 2003), aiming to extract named entities from free text and classify the extracted named entities into certain categories [2]. In this project, we only focused on analyzing the entities we extracted from the posts. During the process of annotating data, we noticed a few named entities keep showing up under certain context. For instance, “Chris Christie” always accuses Trump for being a “national embarrassment”; “Toomey” congratulates Biden many times. Out of curiosity, we performed named entity analysis on the posts to obtain deeper insights. Particularly, we put our focus on human names. Due to the appearance of various names, we decided to filter them based on the total counts in all the posts and the chance of co-occurring with Trump/Biden, as these are the factors that contribute to the significance of a specific named entity. In order to calculate these correlations, certain techniques need to be applied to convert the appearance of names of our interest and the topic for each post into two lists to make comparisons possible. We tackled this with **one-hot encoding**, which is represented as follows:

given a named-entity *N*, if *N* exists in post *P*, **one-hot**(P_N) = 1, **one-hot**(P_N) = 0 otherwise;

given a post *P*, if *P* is categorized as topic *T*, **one-hot**(P_T) = 1, **one-hot**(P_T) = 0 otherwise.

After obtaining two lists for each (named-entity, topic) pair, the correlation between the lists was computed using **Pearson correlation coefficient** as shown below:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

where x_i, and y_i are the elements in named-entity list and topic list respectively; x bar and y bar are the averages for each list.

The sorted list of topics according to *r* under each named entity provided valuable insights for us to testify the effectiveness of our annotations and at the same time, gave us an alternative interpretation for each topic.

4. Results

We first begin by manually annotating posts from our dataset using the four following topics related to perception of election legitimacy:

- "Irrelevant" (I), defined as the posts that are not related to the election process and/or results.
- "Suspicious" (S), defined as posts containing claims of interference or attempted interference with election process.
- "Legitimate" (L), defined as posts that do not contain facts or suspicions suggesting the election is illegitimate
- "Not Legitimate" (N), defined as posts that containing facts that are evidence for interference with the election process

Name	title	coding
t3_jyjecp	Wisconsin officials: Trump observers obstructing recount	N
t3_jyhv9o	Trump was pictured on his Virginia golf course during a coronavirus meeting for G20 leaders	I
t3_jyj5z6	Detroit voters sue Donald Trump over unsubstantiated claims of voter fraud	S
t3_jyi4xi	Impeach the president again — Removal from office wouldn't be the goal this time. It would be to...	I
t3_jyih12	Yes, Trump Is (Still) Engaged in an Attempted Coup; and Yes, It Might Lead to a Constitutional C...	S
...	...	...
t3_jyby0o	Trump Announces Groundbreaking Rules to Lower Prescription Drug Prices	I
t3_jyr204	Trump campaign files for new recount in Georgia	N
t3_jynua5	Release: Toomey Statement on PA Federal Court Decision, Congratulates President-elect Biden | U....	L

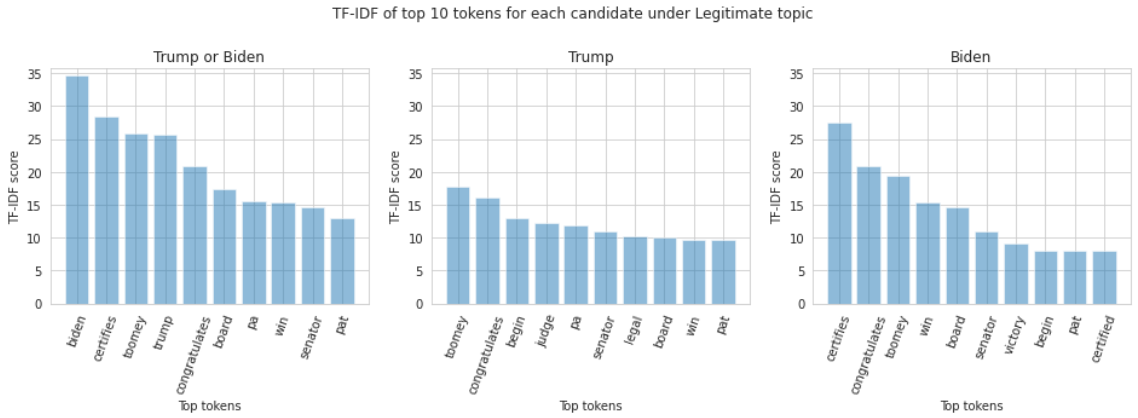
Figure 2: Few examples of posts with coding

We chose these topics as they are not inherently related to candidate preference, minimizing any potential annotation biases. An “S” topic post could raise suspicions of Trump’s interference or Biden’s interference. We perform a secondary annotation for candidate preference, which we describe at the end of this section. The “Irrelevant” (I) posts were the most abundant with a count of 535 posts. We will not look at the results from this topic in detail as it they are mostly irrelevant to the question of election legitimacy perception. The second most popular topic was the “Suspicious” (S) topic with a count of 312 posts. The “Legitimate” (L) topic came in at the third place with 183 posts. Lastly, the “Not Legitimate” (N) topic came in last place with 35 posts. To characterize the content of discussion around each topic, the TF-IDF of the top 10 words for each category were computed. This was done for 3 different post groups: for the 966 posts including Trump or Biden (ToB), which is essentially all collected posts, for the 633 posts including Trump (T), for the 196 posts including Biden (B). We also performed this analysis on the Trump and Biden post group but omitted them from our report as they did not provide much insight.

The top ranked tokens by TF-IDF from posts of L topic in ToB group include both “Trump” and “Biden” (Fig 3 Left). We also find “certifies”, “win”, and “congratulates”. These three words are also ranked top 10 in posts of “L” topic for the B and T groups (Fig 3 Center and Right). This is

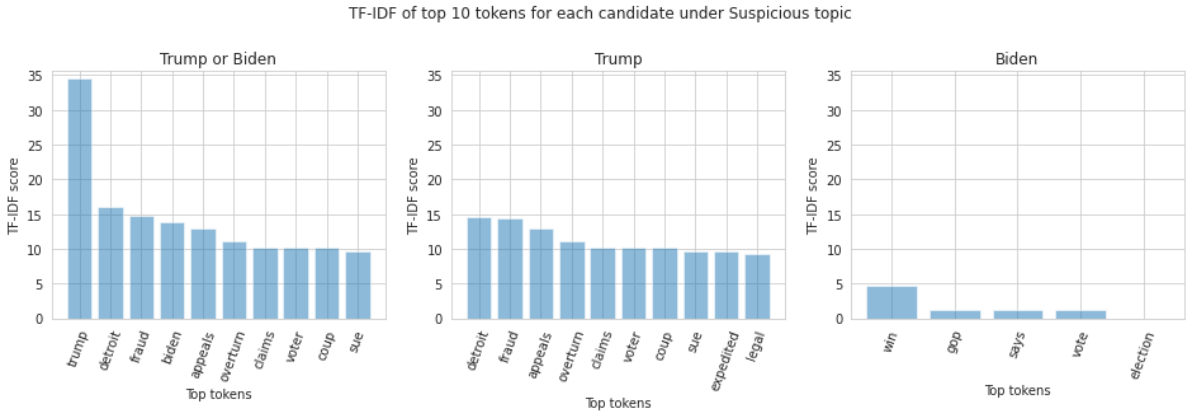
also the case for “Toomey”. While the appearance of “Toomey” may not be immediately obvious, we provide an insightful interpretation in the Discussion section.

Figure 3:



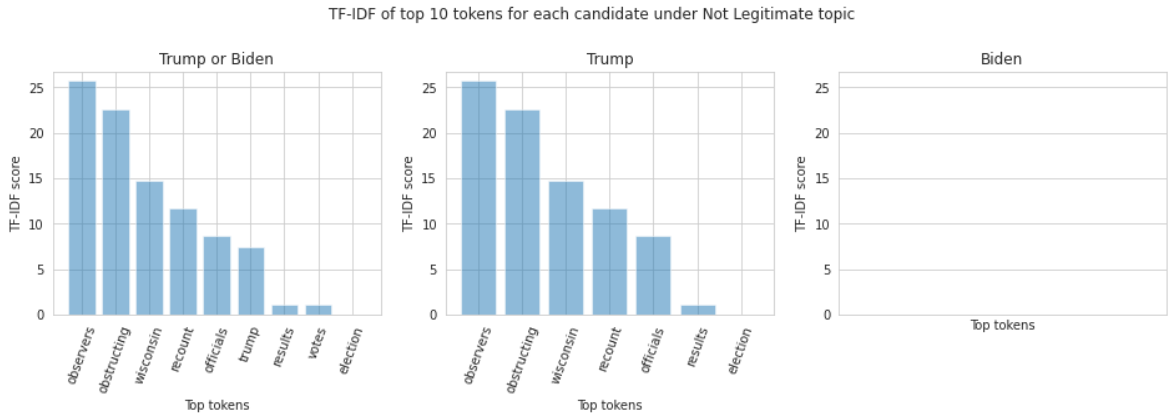
For the top ranked tokens found in posts of S topic in the ToB group, “Trump” and “Biden” both appear again (Fig 4 Left). However, the TF-IDF score for “Trump” is over twice a high than that of “Biden” (34 vs 13). This contrasts with the difference in score for these tokens in posts of “L” topic where “Biden”’s score of 34 is only 30% larger than “Trump”’s score of 26. Other words related to suspicion and election interference, such as “fraud”, “overturn”, “claims”, “coup” and “sue”, appear in the top tokens for posts of “S” topic in the ToB group. These same words also appear in the top tokens for “S” topic posts in the T group (Fig 4 Center). However, none of these words appear in the top tokens of the same topic in the B group (Fig 4 Right). Furthermore, this top 10 token list only includes 5 words due to the small number of posts in the group compared to the number of posts of the same topic in the T group (16 vs 180 respectively).

Figure 4:



A similar trend in token ranking across the ToB, T and B group is seen for posts of the “N” topic. Despite the low total “N” topic post count of 35, we obtain 10 top posts ranked by TF-IDF in ToB group. These tokens (Fig 5 Left), such as “observers”, “recount”, “obstructing” and “officials” are all relevant to the vote count process and/or interference with it. We also see the appearance of “Wisconsin”, which we will develop on further in the discussion. When looking at the same topic in the T group, we only get 7 tokens (Fig 5 Center). However, they are all present in the ToB group as well. Like with posts of the “S” topic, the number of top ranked tokens is drastically reduced in the B group. In fact, there are no tokens since no posts of the B group were labeled as topic “N” (Fig 5 Right). We offer an interpretation for the large disparity in token score and token number for posts of topic “S” and “N” between the T and B groups in the discussion section.

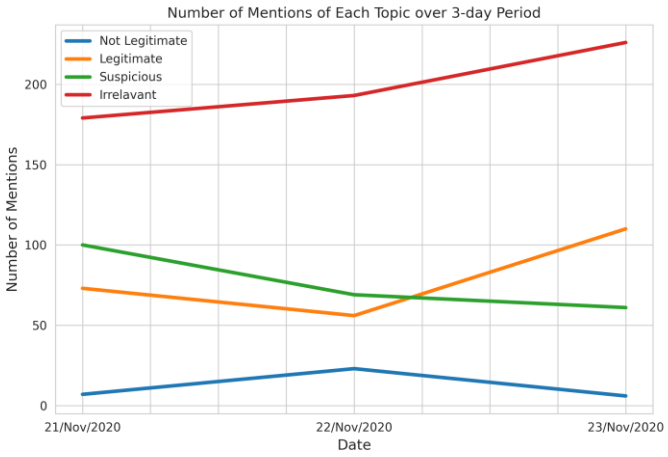
Figure 5:



Time Series

We then set out to analyze how the topic engagement among all posts changed over on a day-to-day basis over the 3-day period of data collection (Fig 6). Posts belonging to the “I” topic steadily rose every day, while the number of posts belonging to the “S” topic steadily declined. The “L” and “N” topic posts do not follow a consistent trend over the three days. However, the direction of change in between the same days are opposite, with the “L” topic posts first decreasing then increasing and the “N” topic posts first increasing then decreasing. At the 3-day end point, the number of “L” topic posts overtakes the number of “S” topic posts while “N” topic posts remain at a similar level as day 1.

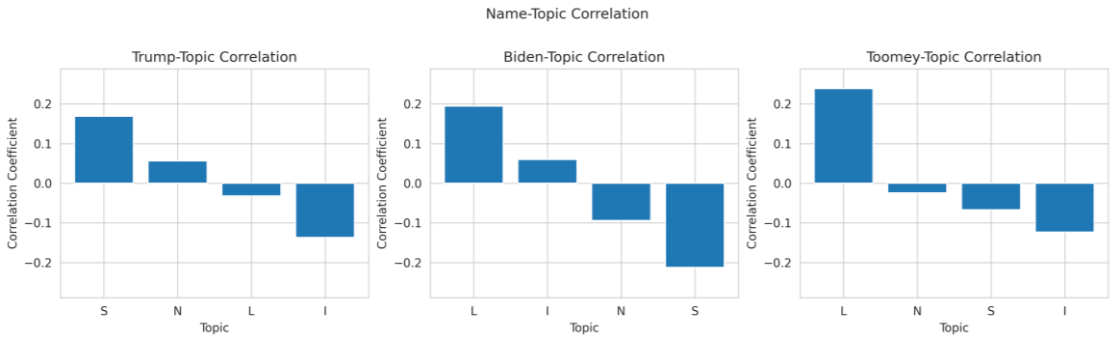
Figure 6:



Named Entity Analysis

As described in the methods section, we perform the analysis on the most frequent names to determine their correlation with each topic (Fig 7). This was done to give context to various tokens referring to people or locations that we identified in our TF-IDF analysis. We obtain “Trump”, “Biden” and “Toomey” in our most frequent names list. We also obtained “Powell”, “Giuliani”, “Hogan” and “Chris Christie”, but we will not discuss their results as they only appear as top ranked tokens in the “I” topic. We found that “Trump” had a correlation coefficient over 0.15 for “S” topic posts and just above 0.05 for “N” topic posts. Unlike “Trump”, “Biden” and “Toomey” both has a correlation coefficient of approximately 0.2 for “L” topic posts.

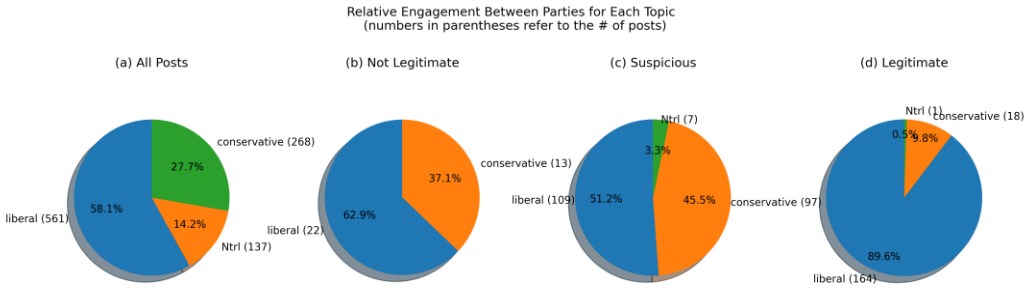
Figure 7:



Secondary Annotation

To facilitate the interpretation of the presented results, we performed a secondary annotation. While our initial dataset contains annotation for whether the post originates from the “/r/politics” and “/r/conservative” subreddits, we found that these labels did not represent whether a post had a pro-liberal or pro-conservative sentiment. We assign the conservative label to any posts that makes Trump look good or makes Biden look bad, and the liberal label for posts with the opposite sentiment. Posts that do not show support for either candidate are labeled as neutral (Ntrl). We found that most posts had a liberal sentiment (58.1%), while 27.7% of posts were conservative and 14.2% were neutral (Fig 8a). We also investigated the post sentiment percentage for each topic.

Figure 8:



5. Discussion

We use our secondary annotation to investigate party engagement per topic. We also looked at the relative topic engagement per party (see appendix). Party engagement per topic visualization shows that posts from our dataset have a bias towards liberals since these posts are twice as abundant as conservative posts (Fig 8a). Since the number of liberal posts are larger than the number of posts mentioning Biden, a considerable number of posts mentioning

Trump are from liberals. The strong liberals' sentiment among the collected posts is consistent with our following interpretation of the TF-IDF analysis results.

The reoccurrence of the words “congratulates” and “Toomey” in the “L” topic across the ToB, T and B groups originate from a series of posts about Pat Toomey, a republican senator, congratulating Biden on his victory (Fig 3). Admission of Biden’s victory from a republican senator suggests that the election must have been legitimate since even members of the losing party has agreed on the outcome. 89.6% of “L” topic posts are liberal, which is consistent with the claim that Biden won the election legitimately (Fig 8d). For the “S” topic, we found “fraud”, “overturn”, “claims”, “coup” and “sue” as top ranked tokens across the ToB and T groups (Fig 4). These are all words that are related to corruption and accusation of such. For this topic, the conservative and liberal sentiment percentages, 45.5% and 51.2% respectively, are very similar (Fig 8c). Considering the high correlation between the “Trump” keyword and “S” topic posts (Fig 7 Left), the absence of “fraud”, “overturn”, “claims”, “coup” and “sue” for “S” topic posts mentioning Biden (Fig 4 Right), and almost even split of party sentiment across posts of these topics, we believe that “S” are the result of conservative and liberal redditors debating Trump’s claims of election fraud, as well as pointing out evidence that Trump himself may be guilty of election interference. For example, the post of “S” topic “Trump admits Michigan Congressmen met to discuss coup” suggests Trump may be guilty of election interference and the post “Battered by failures to reverse Biden win, Trump, allies persist with fraud claims” shows that Trump is pushing the narrative for the existence of election fraud being driven opposition. Results of our “N” topic posts analysis provide ample evidence that if either candidate were guilty of interfering with the election, it would be Trump. Firstly, we obtained 31 “N” topic posts mentioning Trump vs 3 mentioning Biden. Out of all “N” topic posts, 62.9% had a liberal sentiment and 37.1% were conservative (Fig 8b). We find the tokens “observers”, “recount”, “obstructing”, “officials” and “Wisconsin” in the top 10 TF-IDF token list for posts “N” topic posts

in the ToB and T group (Fig 6). These tokens come from multiple posts reporting that Trump election observers have been caught obstructing recounts in Wisconsin. Considering our evidence that Trump has actively been attempting to discredit the election results due to fraud and corruption, we find it very ironic that our results of “N” topic posts suggest that Trump is guilty of the election interference that he claims to exist.

Finally, we discuss the results of the “I” topic posts analysis. Posts of this topic are the most common, making up over half of all the posts in the final dataset. On November 7th 2020, Joe Biden was announced as President Elect after having reached 270 electoral votes. Since we began collecting posts on November 21st, reddit had a period of two weeks to discuss the legitimacy of the election. We believe that redditors slowly reached a consensus over the election legitimacy over this two-week period, leading to increasing number of posts not being related to the election results. Our 3-day time-series analysis of topic engagement supports this interpretation (Fig 6). As the number of “I” topic posts increased over time, the number of “S” posts decreased. We believe that as time went on, reddit users accepted the mainstream media’s consensus on Biden’s declaration as a winner and began engaging in “I” topic discussion instead. We also propose that some of the “S” topic users were finally convinced that the election results are legitimate. This is reflected by the fact that the number of “S” topic posts was overtaken by the number of “L” topic posts on the last day of post collection. Essentially, we propose that our time-series analysis of post topic popularity suggests that the majority of redditors have accepted Biden’s win and have moved on to discussing political topics unrelated to the election legitimacy.

6. Group Member Contributions

This project is a collaborative effort by all three team members. The work was evenly distributed and completed with utmost sincerity and honesty. Each member contributed equally throughout the project. Harmanpreet focused on data collection and cleaning. Then, data annotations were performed by combined effort of all members. Further, Yan focused on applying methods to the cleaned data. Analysis of the results were performed by all members. Lastly, Tristan contributed maximum to the report.

7. References

[1] http://userwww.sfsu.edu/efc/classes/biol710/timeseries/timeseries1.htm

[2] Zhong, Xiaoshi, Erik Cambria, and Jagath C. Rajapakse. "Named Entity Analysis and Extraction with Uncommon Words." arXiv preprint arXiv:1810.06818 (2018).

8. Appendix

Figure 9: Relative Topic Engagement per Partisanship