COMP 598 Final Project - Data Science

1. Overview

The most salient topics discussed around the candidates of the 2020 US election appear to be concerns about Trump's legal undertakings and attempts to overturn the election, the actual results of the elections and vote count certifications, and the transition to the Biden administration. In general, liberals seem to be more concerned about Trump's moves threatening the United States' democracy. In our findings, we describe an apparent polarization of the election results based on the most relevant keywords extracted from two subreddits (/r/politics, and /r/Conservative) for each of the candidates (Trump and Biden).

2. Data

We used the reddit API to collect posts from the newest posts (/new section) of the /r/politics and /r/Conservative subreddits. For both subreddits, we collected 400 posts on the first day, 300 on the second, and 300 on the third, for a total of 2000 posts. We did not collect pinned posts. We filtered posts to those mentioning Trump or Biden in the title using the regular expressions "\D*[Tt][Rr][Uu][Mm][Pp].*" and "\D*[Bb][Ii][Dd][Ee][Nn].*". Then, we removed duplicate posts. We discuss and justify these design decisions in the Methods section. After collecting and filtering the data, there were 832 unique posts mentioning either candidate, of which 605 mentioned Trump, and 315 mentioned Biden. Of those, 88 (9%) mentioned both Trump and Biden. We had a total of 615 unique posts for /r/politics and 217 unique posts for /r/Conservative. Of the posts from /r/politics, 472 posts (77%) mentioned Trump, and 224 (36%) mentioned Biden. Of the posts from /r/Conservative, 133 posts (61%) mentioned Trump, and 91 (42%) mentioned Biden.

3. Methods

3.1 Choice of subreddit

After discussion, we decided to collect new posts (/new) rather than hottest posts (/hot). Although the hottest posts definitionally filter for the "trendiest" and most salient topics – the posts with the most interaction – and make extreme views more explicit, they suffer from multiple drawbacks.

Gathering from /hot would introduce significant overlap in the posts collected over the three days, as some posts would remain among the top posts. These duplicated posts would then have to be removed, otherwise their defined topic would be over-represented in the results, which would greatly reduce our sample size of posts to annotate. Furthermore, hottest posts could have been posted at an earlier date outside of the sampling period, including dates before the election and thus not representative of our period of interest, the “days following the US election.” The manual annotation that we performed was also more representative using newest posts, compared to /hot which filtered to more popular topics. Following annotation, the less relevant topics in /new would have lower counts, while the more relevant topics would stand out.

3.2 Pinned posts

We decided not to collect pinned posts, because these often deal with moderation and housekeeping topics of the subreddit. In addition, they would probably not have been posted on the sampling dates, and would be duplicated across days.

3.3 Definition of a mention

We defined a mention of Trump or Biden as containing either of these words, in a case-insensitive manner, and only retaining cases where the names are immediately preceded by non-digits. We did not impose any restrictions on the characters immediately following the names. These rules allowed us to capture hashtags like “#trump2020” and words like “Trumpism” as mentions of Trump. Using these rules, we are aware that we may also capture rare cases containing “Trump” or “Biden” as components, for example words such as “trumpet,” but decided to accept this drawback, and to compensate by discarding errantly captured content by annotating these posts into the “NA” category.

3.4 Filtering for mentions in the post title

We first intended to retain posts with mentions of Trump or Biden in the title, body, or both. However, we noticed that the great majority of posts on /r/politics and /r/Conservative only contain a

URL as the post body. Therefore, we decided to filter for posts with a mention of Trump or Biden in the title.

3.5 Collection of posts over three days

To collect a total of 1000 posts from each subreddit, we collected 400 posts on the first day, 300 posts on the second day, and 300 posts on the third day.

3.5 Removal of duplicate posts

We noticed that some posts were duplicated in our sample. This was due to an insufficient number of posts coming from each day. For example if there were 250 posts on November 21st, a day when we intended to collect 300 posts, we would have sampled the 250 posts from November 21st in addition to 50 posts from November 20th. To avoid diluting our sample (as discussed in section 3.1, choice of subreddit), we discarded duplicates before moving forward with annotation. As a result, we have less than 1000 posts in total for each subreddit.

3.6 Process used to select the topics

For open coding and topic definition, we sampled 100 posts at random from each subreddit (from the pool of posts mentioning the candidates), for a total of 200 posts. Each member of the team manually annotated these 200 posts. Then, we met to discuss our annotations and decide on a coding scheme to apply to the remainder of the posts.

Our goal was to choose comprehensive topics with clear inclusion and exclusion criteria, where any post could fit at least one of the topics. Additionally, we tried to choose objective topics. For example, we questioned whether “posts against Trump” could be a valid topic, and came to the conclusion that it wasn’t because it was subjective and was often very challenging to determine the opinion and political views of the author of a post solely from the title. We will elaborate further in the Discussion section.

In cases where team members had similarly-defined topics, we retained or merged the topics into a common and broader category. For example, we merged topics such as “COVID-19” and “health” with the broader “internal affairs”. Because some posts could belong to more than one topic,

we also defined an order of priority: if posts could be classified into foreign affairs, internal affairs, or transition and new administration, they should be assigned to foreign affairs. As we annotated the entire dataset using our topics, we re-evaluated our topics, refining some of the definitions and establishing the aforementioned order of priority to ensure that our typology worked. We made small adjustments, corrected our annotation accordingly, and annotated the whole dataset. We ensured that our typology was comprehensive by adding a “not applicable” (NA) category, identifying posts that did not fit into any of the categories, and confirmed that only a very small number of posts (72) had to be placed into the NA category. This was not the same as using an “other” category, where a larger number of posts would be expected to fit.

3.7 TF-IDF analysis

To understand the salient topics discussed according to each category we defined, we used the TF-IDF method to capture the most frequent words in the titles of the collected posts. We made three distinct evaluations for the input document of TF-IDF:

1. A standard analysis used the entire set of 2000 collected posts;
2. TF-IDF calculations considered only the posts were "Trump" was mentioned, and subsequently only the posts where "Biden" was mentioned;
3. TF-IDF calculations considered only the posts from /r/politics, and subsequently only the posts from /r/Conservative.

For the standard TF-IDF using the entire set of 2000 posts, we ranked the 10 most frequent words per topic (which we refer to as "keywords" in this report). Note that these results include the duplicated posts mentioned in the Methods. There were also posts that did not belong to any category because they did not mention "Trump" or "Biden". In the TF-IDF calculation, these posts received a phantom labelling of "OTHER". Apart from the standard TF-IDF calculation using the entire set of 2000 posts, we also calculated the TF-IDF targeting two different dimensions: one for posts mentioning each candidate (Trump/Biden), and another for posts from each subreddit (/r/politics and /r/Conservative).

We then broke down the frequency of these keywords to help us assess how candidates were being discussed, and relative engagement with the topics by liberals and conservatives. We broke posts down by subreddit to yield the keyword frequency overall, keyword frequency in /r/politics, and

keyword frequency in /r/Conservative. We broke posts down by candidate mention to yield the keyword frequency overall, keyword frequency in posts mentioning Biden, and keyword frequency in posts mentioning Trump.

4. Results

4.1 Topics Selected and Definitions

After our open coding and discussion, we settled on five topics: “transition and new government”, “election results and counts”, “election legal affairs”, “internal affairs”, and “foreign affairs”. These topics are broad enough to succinctly capture most of the posts gathered, while also providing information about salient topics around the election.

Table 1: Top 10 words by TF-IDF for each topic

Transition and new government	**Internal affairs**	**Foreign affairs**	**Election results and recounts**	**Election legal affairs**	**NA**
cabinet	prescription	Afghanistan	recount	Michigan	antibody
transition	costs	G	obstructing	lawyers	lara
picks	costs	drawdowns	certifies	lawsuit	fda
POTUS	lead	eat	Wisconsin	lawmakers	cocktail
Twitter	threatens	Eurasia	congratulates	lawyer	infected
inauguration	authority	golfing	requests	subvert	sister
Merrick	chances	hasty	results	Minnesota	ivanka
Garland	defend	immigrant	observers	coup	supporters
climate	learned	Iraq	certified	results	grants
asking	seals	Israel	Toomey	tactics	parody

4.2 Detailed Findings by Category

Transition and new government captured content related to the Biden administration during transition (e.g., content about Trump refusing to concede), its intentions once in office (e.g., policy proposals, Cabinet selections), and advice given to his administration. Positive examples of this category were “Merrick Garland on list to be Biden's attorney general: report” (related to Cabinet selection, intentions when in office), “Trump’s press secretary Kayleigh McEnany heckled in briefing:

‘When will you admit you lost?’” (related to Trump refusing to concede), and “How Joe Biden Can Use the Political Power of Obamacare to Expand It” (related to advice given to the Biden administration). Negative examples of this category would be anything not related to the Biden administration during transition or once in office. For example, “Joe Biden's 'Vote Joe' website defaced by Turkish Hackers” (about Turkish hackers, not the administration itself; category: NA). An example of an edge case was: “A destructive legacy: Trump bids for final hack at environmental protections”. Here, it was not clear whether the main topic is Trump’s current political decisions (hindering environmental protections), or the challenge they will pose to the government transition and to the Biden administration (“destructive legacy”). We reasoned that Trump’s main goal was to impede transition to a new government, and therefore included this post in “transition and new government”.

Election results and recounts referred to relatively objective factual information about vote counting and the election (e.g., demographic analyses of votes, news about officials certifying results, recounts). Positive examples of this category were “Georgia certifies that Joe Biden won the recount” (certifying results); “Georgia Gov. Kemp Calls for Audit, Says Trump Can Demand New Recount” (recount); and “Donald Trump Improved Standing With Latinos In 78 Of America’s 100 Majority-Hispanic Counties” (demographic analysis). Negative examples of this category would be posts that are not relatively objective, like “Republican voters say ‘no way in hell’ Trump lost” (unsubstantiated opinion, not a factual result; category: NA).

Election legal affairs captured content related to the ongoing attempts of the Trump administration to challenge the election results (e.g., news about the legal team, lawsuits filed by and against the Trump team, concerns about threats to democracy). Positive examples of this category were “Trump, Allies Take Frantic Steps to Overturn Biden's Victory” (attempts to overturn results); “Donald Trump Undercuts American Democracy as he clings to power” (threats to democracy); “Trump supporter election lawsuit affidavit seems to mix up Michigan and Minnesota” (legal team filing); and “Do Trump’s Lawyers Know What They Are Doing?” (legal team commentary). Negative examples of this category would be posts not related to challenging election results such as “Trump options narrow as Michigan backs Biden win” (refers to Trump’s options in trying to overturn the election, but is more focused on Michigan backing the election results; category: election results and counts). An example

of an edge case was: "Lin Wood: President Trump won 400 electoral votes" (factually incorrect claim, unclear who Lin Wood is, category: NA).

Internal affairs captured national policies enacted by the Trump administration (e.g., executive orders to reduce prescription drug prices). Positive examples of this category were "Peters criticizes Trump for not taking action after cyberattacks on hospitals, COVID-19 researchers"; "Trump, Pelosi barrel toward final border wall showdown"; and "Trump Makes Late-Term Bid to Lower Prescription Drug Costs." Negative examples of this category were things not related to policies that have been or are currently being enacted by Trump. For example, posts related to national policies Joe Biden intends to enact, such as "Biden appoints two former lobbyists to senior staff" were not internal affairs; they are captured by transition and new government. An example of an edge case was "A tragic farce: Trump's ignominious White House exit", where it was not clear whether the post should be included into internal affairs or NA. Here, Trump's exit from the White House was an action that Trump performed that was related to governance (exiting from it).

Foreign affairs captured content related to foreign actors (e.g., policy towards Iran, withdrawal from Afghanistan). Positive examples of this category were "Greece, Egypt seek Biden role in East Mediterranean dispute" (although it references Biden, this post was categorized as foreign affairs because it is related to the actions of foreign actors); and "Merkel's Germany Tells Trump Not to Bring Troops Home from Afghanistan." Negative examples were posts not related to foreign actors. For example: "Ilhan Omar to President-elect Biden: Seize 'Once-in-a-Generation' Chance to End Disastrous US Foreign Policy" (mentions foreign policy, but about a member of the House of Representatives asking for Biden to enact specific policies; category: new government and transition). An edge case for this category was "Putin refuses to recognize Joe Biden's victory." This is about both a foreign actor and the new administration. Given our order of priority (where posts that can be classified into foreign affairs, internal affairs, or transition and new administration should be assigned to foreign affairs), this post is a positive example of foreign affairs.

In addition to these five topics, we had a "not applicable" category. ***Not applicable*** captured irrelevant posts, including those that did not relate to our other five topics, but mostly posts captured by mistake. Examples of irrelevant posts were "Berkeley professor sparks outrage by claiming Trump supporters who have an 'indifference' to COVID are like Germans who 'ignored and profited from the

Holocaust'" (not about Trump or Biden specifically, vaguely about “Trump supporters” and a university professor) and “Bigot: Joe Biden Lackey Wanted to Destroy Amy Coney Barrett Because She's Catholic” (conjecture about someone tangential to Biden). An example of a post captured by mistake would be something unrelated to Trump or Biden that contains the word “trumpet,” which would be captured by our regular expression filter designed to capture hashtags like #trump2020.

As discussed during the lectures of this course, there is no true objectivity in coding. As coders we were biased towards finding certain information believable and objective. We attempted to minimize our bias by leaning on factual reporting where possible (e.g., whether a state’s race has been certified is a matter of fact). Additionally, although /r/politics is overrepresented in our annotated data, we did use all 2000 posts (split evenly between subreddits) collected as the standard TF-IDF calculation to dilute the influence of posts from /r/politics on our keyword lists.

4.3 Topic Characterization and Engagement

4.3.1 Election Legal Affairs

The keywords for election legal affairs were: “Michigan”, “lawyers”, “lawsuit”, “lawmakers”, “lawyer”, “subvert”, “Minnesota”, “coup”, “results”, and “tactics”. These terms speak about attempted election theft (coup, results, subvert), the mechanism (lawyers, lawsuit, lawmakers, lawyer, tactics), and the targets (Michigan, Minnesota). The terms appear much more often in posts mentioning Trump. There is a split between subreddits: /r/politics mentions terms related to election theft and the target more often; /r/Conservative mentions the mechanism more.

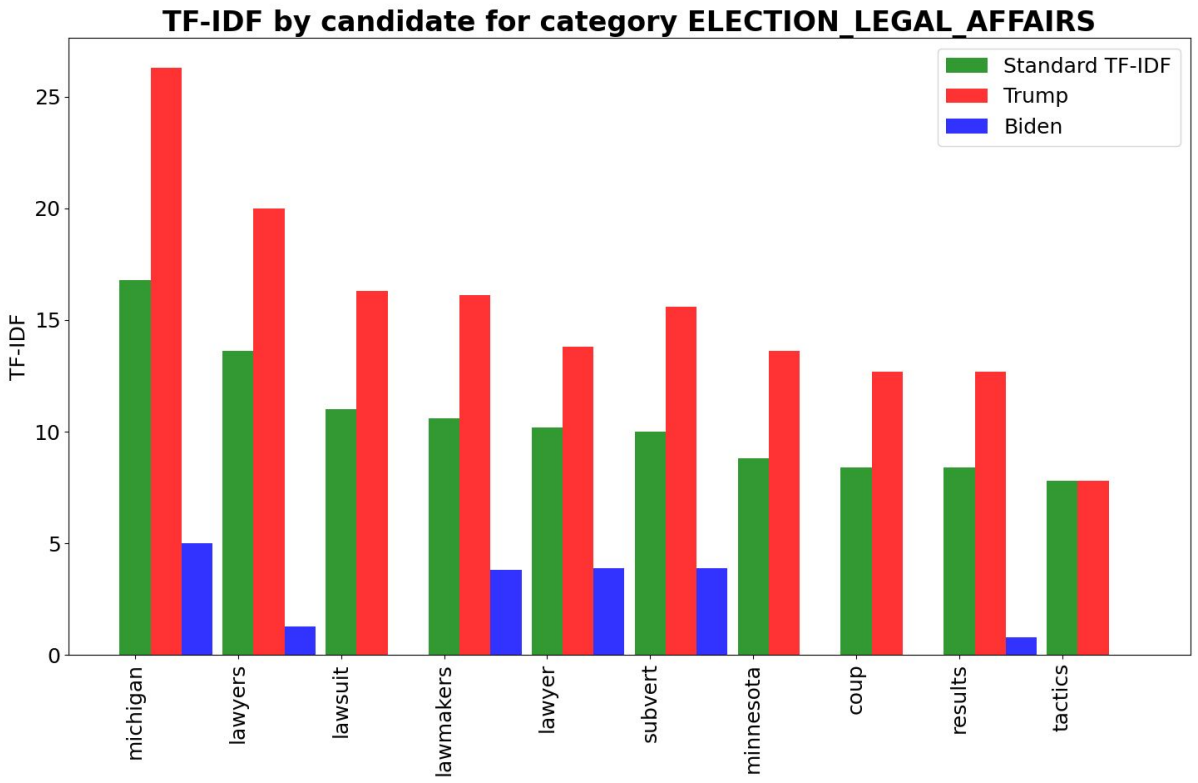

Figure 1: TF-IDF by candidate for Election Legal Affairs.

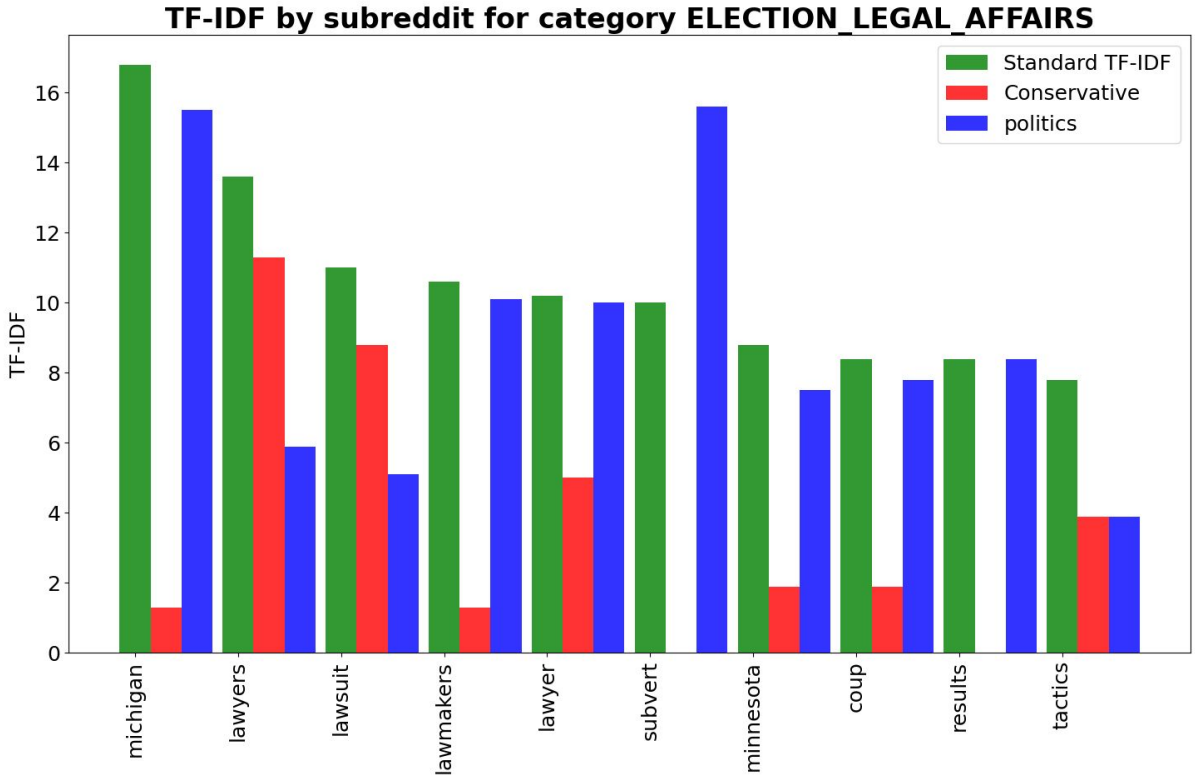

Figure 2: TF-IDF by subreddit for Election Legal Affairs.

4.3.2 Election Results and Recounts

The keywords for election results and recounts were: “recount”, “obstructing”, “certifies”, “Wisconsin”, “congratulates”, “requests”, “results”, “observers”, “certified”, and “Toomey”. These terms reflect a relatively dry picture of the count (recount, certifies, certified, congratulates, results), with an allusion to election theft (obstructing, observers). The terms appeared more often in posts mentioning Trump, and in posts in/ r/politics.

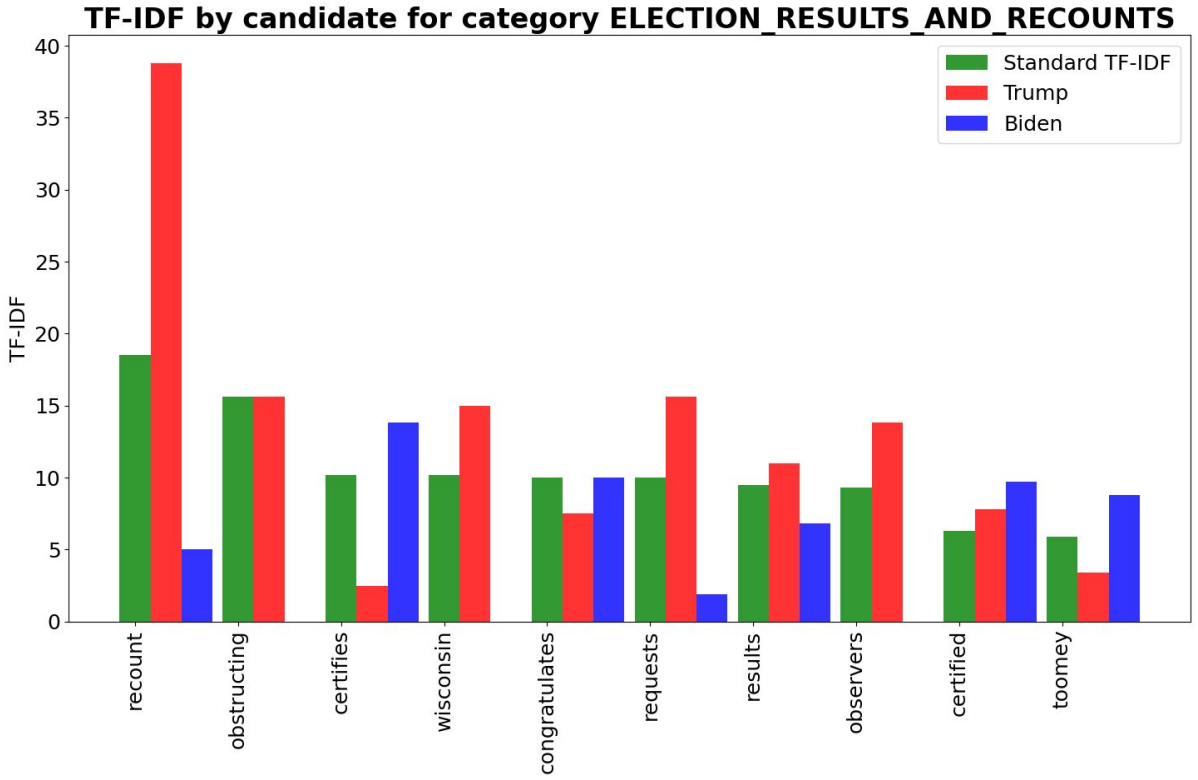

Figure 3: TF-IDF by candidate for Election Results and Recounts.

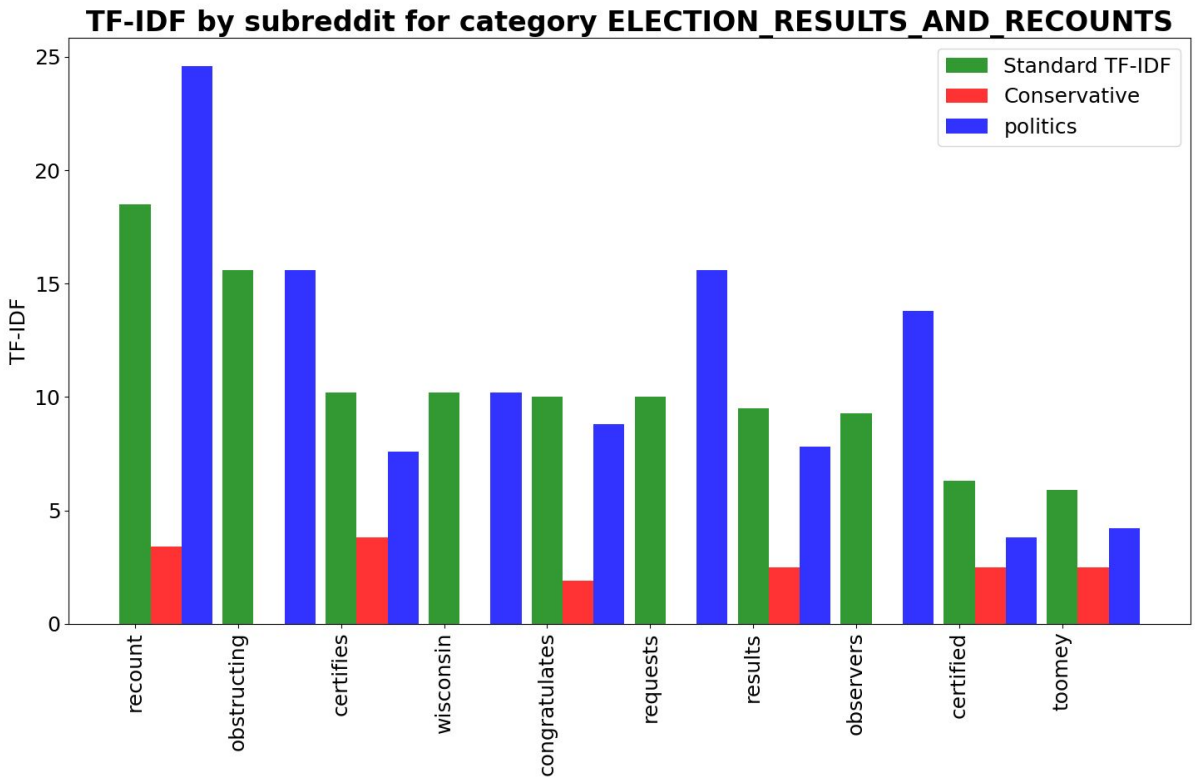

Figure 4: TF-IDF by subreddit for Election Results and Recounts.

4.3.3 Transition and New Government

The keywords for transition and new government were: “cabinet”, “transition”, “picks”, “POTUS”, “Twitter”, “inauguration”, “Merrick”, “Garland”, “climate”, and “asking”. These terms reflect a focus on the new team (cabinet picks, Merrick Garland, POTUS), the transition process (transition, inauguration), and policy intentions (climate). The terms appeared more often in posts mentioning Biden, exclusively so in the case of Merrick Garland and climate. Discussion of these keywords was more frequent in /r/politics.

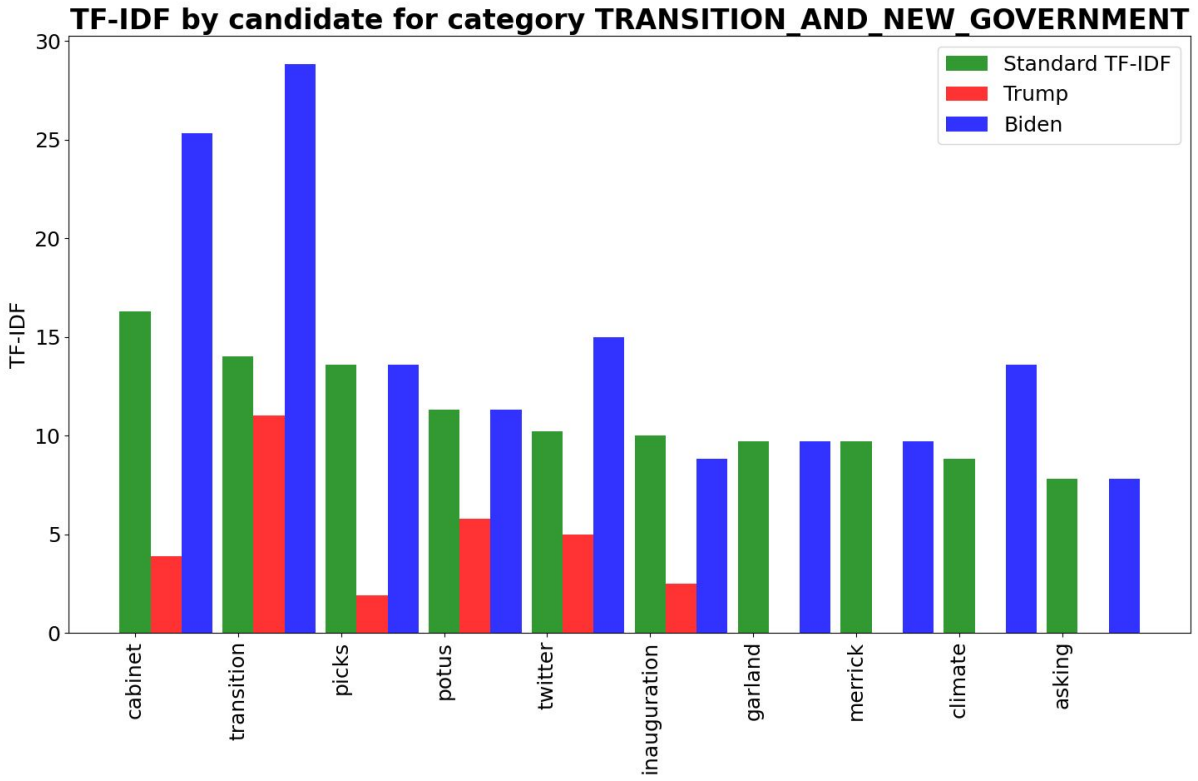

Figure 5: TF-IDF by candidate for Transition and New Government.

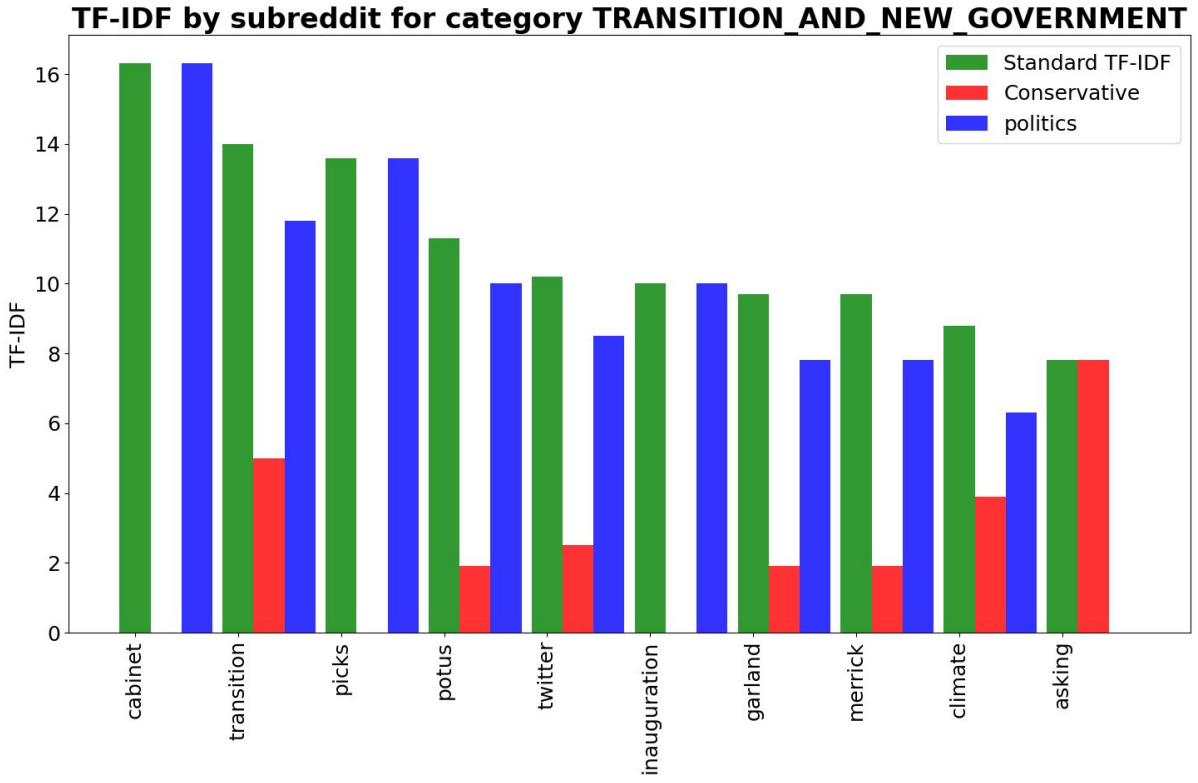

Figure 6: TF-IDF by subreddit for Transition and New Government.

4.3.4 Internal Affairs

The keywords for internal affairs were: “prescription”, “saved”, “costs”, “lead”, “threatens”, “authority”, “chances”, “defend”, “learned”, and “seals” (referencing Navy Seals). These terms reflect internal actions relevant to the Trump administration, notably Trump’s executive action to reduce drug prices (prescription, saved, costs). The terms appeared more often in posts mentioning Trump, as expected given the category definition. Terms were mentioned relatively evenly across subreddits, with slightly more focus on drug-related terms in /r/Conservative. We determined that the results from this category were not informative with respect to election legitimacy, so placed the graphs of TF-IDF results in the Appendix.

4.3.5 Foreign Affairs

The keywords for foreign affairs were “Afghanistan”, “G” (although seemingly an outlier, this stands for G20), “drawdowns”, “eat” (outlier resulting from small size of the foreign affairs category), “Eurasia”, “golfing”, “hasty”, “immigrant”, “Iraq”, and “Israel”. These terms reflect key foreign agents and responsibilities. The terms appeared evenly in posts mentioning Trump and those mentioning Biden. Discussion of these keywords was relatively equal across subreddits, excepting the G20 and golfing, which were discussed more in /r/politics. Like for internal affairs, this category was not informative with respect to election legitimacy, so the graphs of TF-IDF results are in the Appendix.

5. Discussion

Before looking into the topics that were discussed around the candidates, we expected differences in how each candidate would be approached in each subreddit. For example, it was apparent that both liberals and conservatives mentioned Trump more than Biden, even though Biden is the incumbent President. Indeed, in both /r/politics and /r/Conservative, there were more posts mentioning Trump (77% in /r/politics, 61% in /r/Conservative). After analyzing the collected data, other differences in the points of view of each political group became apparent in each category.

The “Election Legal Affairs” category speaks directly to perceptions of election legitimacy and divergent engagement between liberals (/r/politics) and conservatives (/r/Conservative). The keywords **subvert**, **coup**, and **results** refer to attempted election theft. The **tactics** of the campaign were to

strongarm **lawmakers** and have **lawyers** file various **lawsuits**. The Trump legal team erroneously confused **Michigan** and **Minnesota** in a court filing alleging voter fraud.

Every term was more commonly used in posts mentioning Trump, reflecting that Trump and his legal team were the ones pushing legal filings through courts. Liberals used terms in the category more overall, and specifically used terms like **subvert** and **coup** much more than conservatives. Conservatives used terms in this category less than liberals did, with the most relative focus on the legal process, with terms like **lawyers** and **lawsuit**.

These results paint a polarized picture. Liberals were more engaged with the topic of election legal affairs overall, with particular focus on the legal affairs as a method of stealing an election, bungled by unforced errors in court filings. Conservatives focused less on the topic. To the extent they focused on it, they discussed the process, but left out legal filing gaffes and the conceptualization of election theft.

For the topic of "Election Results and Recounts", the standard TF-IDF calculation listed **obstructing** as the second most frequent word. The TF-IDF analysis by candidate showed that **obstructing** was equally relevant for the set of posts mentioning Trump, whereas it had no representative value for the set of posts mentioning Biden. We found that liberals were the only source of mentions to **obstructing**. Although we did not calculate a correlation between Trump and **obstructing**, it seemed that liberals related **obstructing** to Trump. We performed a similar analysis with the keyword **requests**, possibly indicating that according to liberals the majority of **requests** for **recounts** were coming from posts mentioning Trump. Overall, liberals were much more engaged with the most frequent keywords obtained by the standard TF-IDF results, whereas conservatives seemed to have a variety of relevant keywords.

Considering the topic of “Transition and New Government,” liberals appeared to be more attentive overall compared to conservatives, as expected. Posts annotated into this category referenced Biden much more often. As expected, Trump was barely mentioned in posts from this topic, given that we defined this topic as content related to the Biden administration, specifically. The keywords appear to represent some of the most salient topics discussed around Biden and the topics that liberals are most concerned about regarding the **transition**, such as Biden’s **picks** for his **cabinet**. This includes Judge **Merrick Garland**, who is being considered for the position of attorney

general in Biden's administration. **Climate** issues and the **@POTUS Twitter** account (which Twitter confirmed will be handed to Biden regardless of the outcome of Trump's attempts to steal the election), are other topics that liberals appear to be particularly interested in, related to Biden.

Overall, both liberals and conservatives focused a lot of attention on the outgoing President, Donald Trump. In particular, he was mentioned very frequently in posts related to his attempts to challenge the election and obstruct the democratic process. Fewer posts related to President-Elect Joe Biden. Biden was mentioned particularly frequently in posts related to the election transition and new government. Polarization around these topics reflects the larger polarization in United States politics.

6. Group member contributions

Hanneli: Contributed to design decisions, wrote the data collection and filtering script, participated in definition of the topics, wrote the TF-IDF script, participated in writing the discussion and participated in report revisions.

Mairead: Contributed to design decisions, participated in definition of the topics, wrote the Results section, sent out regular meeting follow-ups outlining tasks for each member, participated in writing the Discussion revising the report.

Gabrielle: Contributed to design decisions, participated in definition of the topics, wrote the Data section, wrote the Methods section, participated in writing the Discussion and participated in report revisions.

Appendix A - Charts

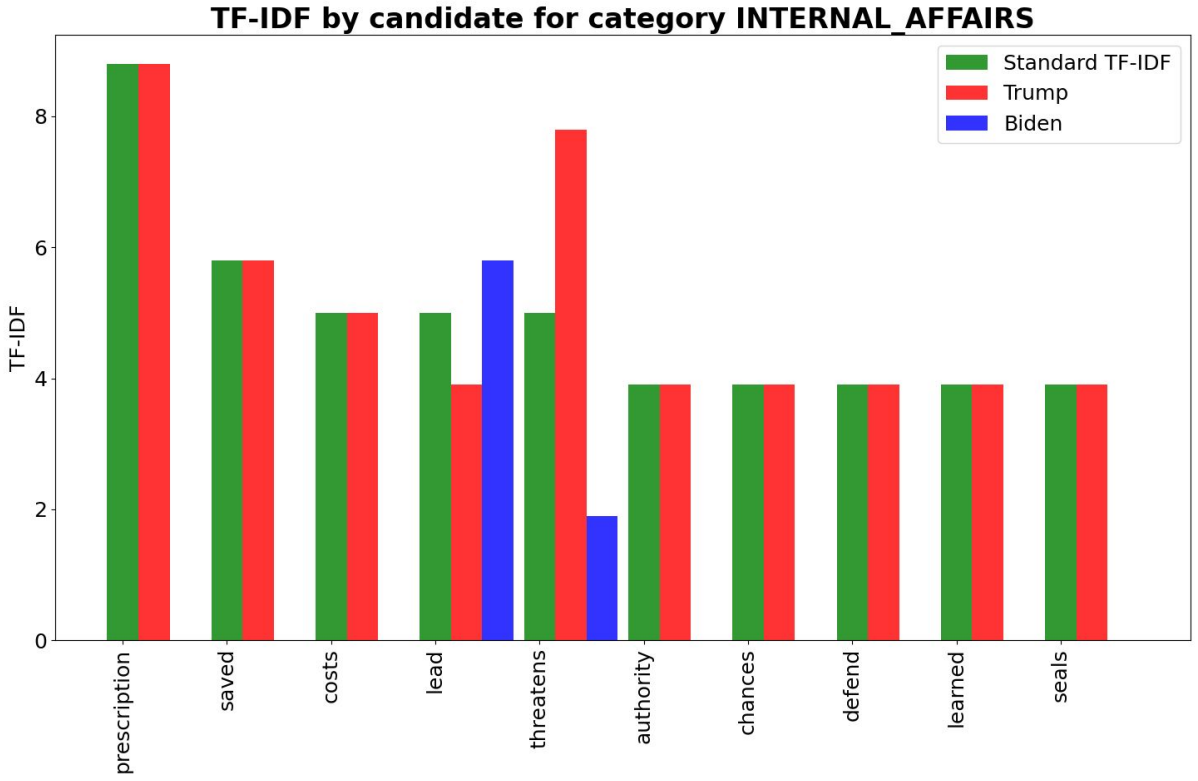

Figure 7: TF-IDF by candidate for Internal Affairs.

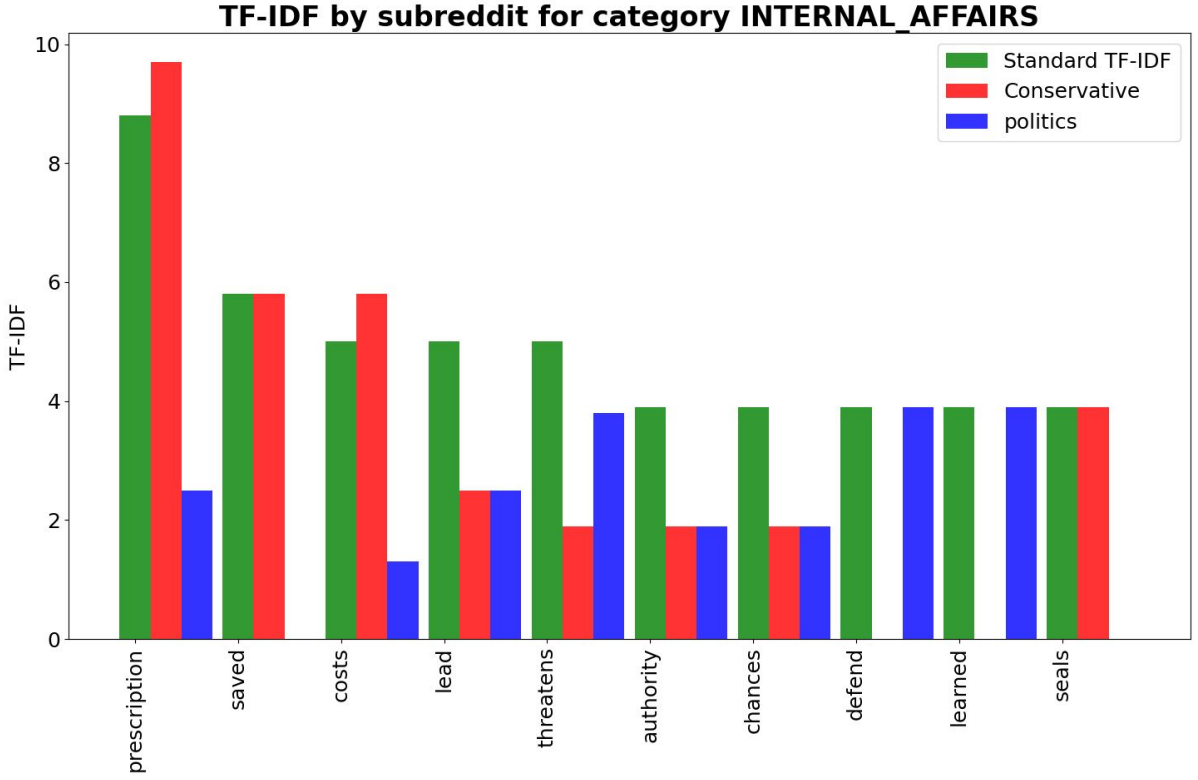

Figure 8: TF-IDF by subreddit for Internal Affairs.

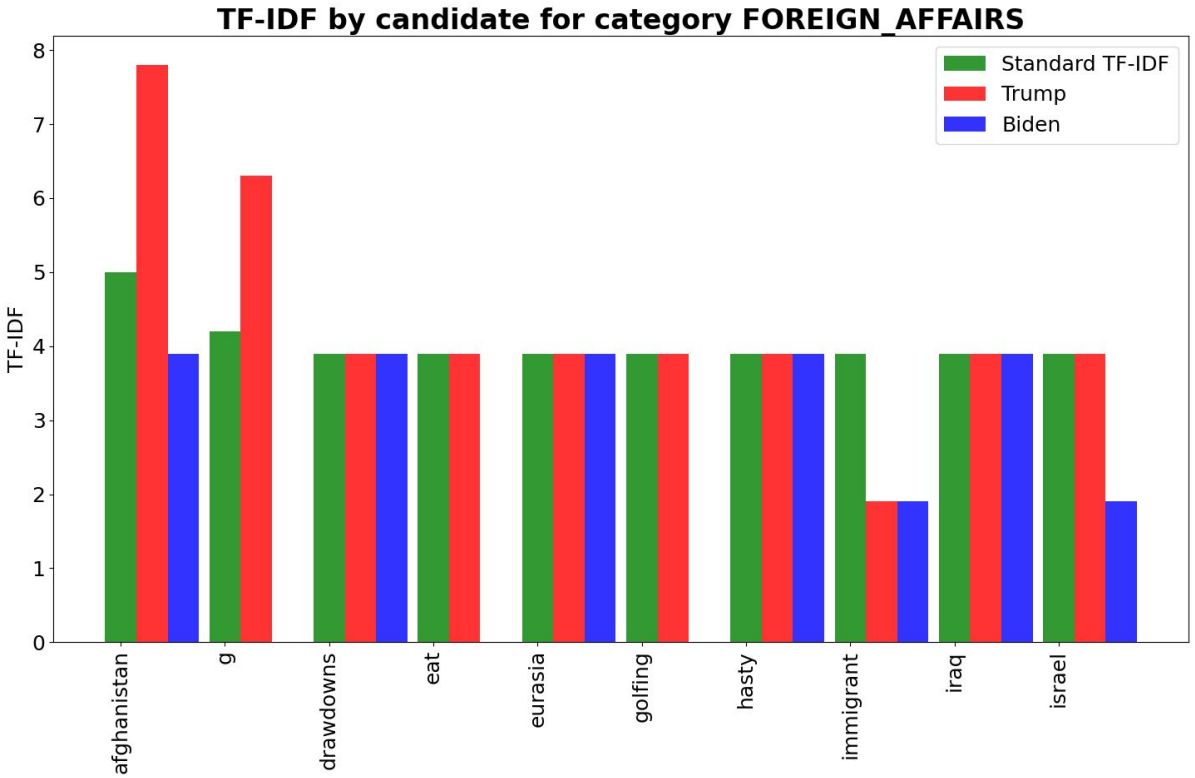

Figure 9: TF-IDF by candidate for Foreign Affairs.

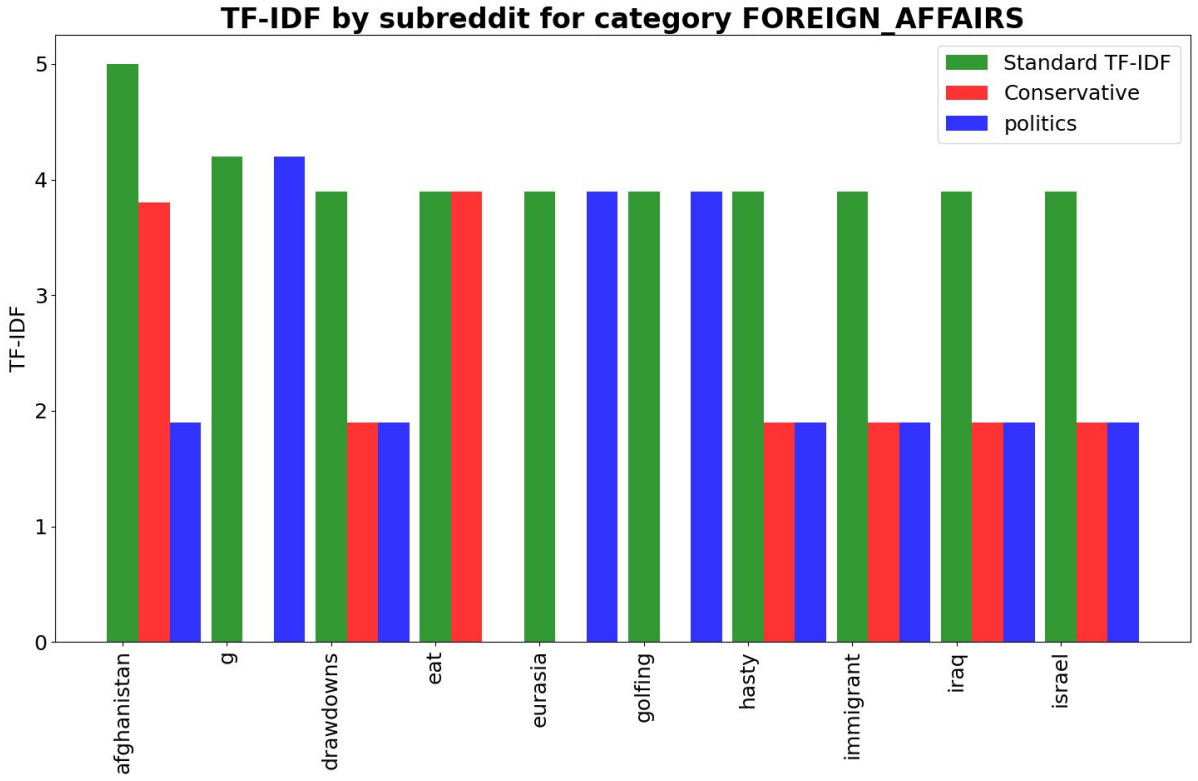

Figure 10: TF-IDF by subreddit for Foreign Affairs.

Appendix B - Source Code

The source code for the project can be found on https://github.com/hannelita/comp598. The project will be made public after the submission deadline has passed.