DIGI 405 Corpus Building Project: Building a Corpus on Articles Pertaining to the COVID-19 Vaccines Across the US Political News Spectrum.

Graham Greig

SN: 47022356

# Background

From the perspective of a Canadian, it is hard to avoid COVID-19 news from the United States. Much of this information is politically charged and fraught with misinformation. One area in particular is the discourse surrounding COVID-19 vaccines. Vaccines need a high uptake rate to be effective and produce herd immunity. Without this, populations allow for infections to continue to take place, viruses to mutate and vaccine efficacy can decrease (“SAGE”). Misinformation surrounding vaccines has increased immunization hesitancy in America, and allowed the continued death of its citizens. At the time of writing, roughly 73% of Canadians have been partially vaccinated with 64% being fully vaccinated. In the USA, these figures are just 59% and 50% respectively (Ritchie, Hannah, et al.). This is despite the US having better access to vaccine supply and starting their vaccination campaign earlier. In a quote from Dr.Anthony Fauci: “*Canada is doing better not because we are trying any less than they are trying. It's because in Canada, you don't have that divisiveness of people not wanting to get vaccinated, in many respects, on the basis of ideology and political persuasion*” (Fauci). Clearly, the problem of low vaccination uptake is coming from this indecisiveness. It is interesting to explore how “political persuasion” is represented in American media and how that media in turn is or is not encouraging vaccination reluctance.

For this corpus, an exploration will be performed across the political spectrum. This will allow for an investigation of differing perspectives on vaccines from the left and right. It’s also desired to compare across news outlets with varying reliability. This could allow one to clearly see misleading or hyperbolic claims when comparing less and more reliable media. Sources for building this corpus were found with the help of the Ad-fontes media bias chart. The chart displays news outlets in their relation to reliability (y-axis) and partisanship (x-axis) where the position of each outlet is the weighted average of it’s analyst reviewed articles. The scales go from -42 to 42 on political bias and 0 to 48 on reliability. Interestingly, there is a trend that as outlets become more partisan, they tend to become less reliable. The process of choosing sources also took into consideration the ease of scraping the sites by ensuring that they do not use “lazy-loading” and avoid paywalls. Firstly, the right leaning article was chosen: *The Federalist* which has a right skew of 20.3 and a reliability of 26.6. This puts it in the hyper partisan right with mid to low reliability. To match this on the left, the publication *Truthout* was chosen with scores of -20.1 and 29.3 for skew and reliability. Finally, *The Independent Journal Review* (IJR) was chosen to represent the middle with a bias of 0.6 and a reliability of 42.2.

# Corpus

The overall corpus has been split up into three sub corpora with each being from one source. The IJR corpus is made up of 319 articles with 136,913 tokens, The Federalist corpus consists of 131 articles with 92,205 tokens and the Truthout corpus is 54 articles and 71,375 tokens. Despite the difference in size of the corpora, they have a comparable number of distinct tokens with 8092, 8508 and 6936 respectively.

Some pre-analysis has been performed to find some interesting tokens across the corpora. An obvious token to use is “vaccine\*”. Looking at the cluster/ n-gram analysis shows that the 2-grams “vaccine doses” and “vaccine-passports” appear in the top ten two-grams of the IJR and Federalist corpora and could provide some unique insight. After applying a stoplist to all corpora, the word “people'' is in the top ten the most frequent tokens of all corpora. This could lend some interesting insight into how the authors address the citizens of america. Finally, “Trump” appears in the top twenty tokens of Truthout and “Biden” appears in the top twenty tokens of The Federalist. These tokens are both in the top ten most frequent words of the IJR corpus. Investigating these terms could show how the individual corpora are using politically charged messages to encourage or discourage vaccinations.

To analyze these tokens, a cross examination of their concordances in each corpora could be performed. This would give an idea of how they are used in the context of the articles and may give insight into misinformation and politically charged speech. More interesting word groupings around these root tokens could be found using collocate analysis to determine important nearby words. This could be followed up by a further concordance analysis to understand these collocates in context. Following this, a keyword analysis comparing each sup-corpora to one another could spot more interesting tokens which could be investigated using the methodology above. Finally, it is of interest to see if the discourse around vaccines has changed since Biden became president, replacing Trump. For this, each sub-corpora could be divided further into two sub corpora to compare against one another and see if positive and negative key adjectives flip usage.

# Method

The original web scraping method to be used for this project was Scrapy. This is because it includes a request to the website and parser functionality. However, this was abandoned early in the project due to some difficulties when working with the application. The main two reasons for leaving Scrapy behind was its somewhat hard to use framework and the need to reboot the kernel each time a spider is run. As well, because there was a breadth of knowledge given in lecture on Beautiful Soup and Requests, these seemed like the ideal scraping packages to use. To narrow the search window for articles on each news website, a root url was found using the respective website’s search function and the search term “covid vaccine”. From this url, the scraping methodology was as follows:

1.) Start a request on an initial root url using Requests.

2.) Make a soup object from the page using Beautiful Soup.

3.) Extract all of the links from the page using the and add them to a list of links if the links are not already in the url list and they contain the word “vaccine” in the title.

4.) If the next page of links is available, go to it and continue to gather links to articles.

Else...

5.) Start iterating through the list of links.

6.) Extract the text, title, date and author from the page.

7.) Save each visited page as a text file with the format: (Date\_ArticleName\_AuthorName.txt)

The proper CSS or class selectors were found using the built-in Firefox inspector and a trial and error process of choosing logical selectors from reading the html data. In some instances (ex. author tags) this was cumbersome and an auto generated CSS was used. This was a pretty straightforward process but it did run into a few issues. For The Federalist and Truthout websites, the requests module was running into an <Response 403> error. This was fixed by explicitly setting the user-agent for the response header. Another issue that arose was some files were not being saved, as the article titles contained illegal characters. This led to stripping of all whitespace, semicolon and quotation characters in the file names. In the raw text data, some text had twitter links and these were removed by identifying lines with “@” or ".twitter” in them. Finally, many articles have common text at the bottom of the article, often about the article author or the outlet's policies. This text was stripped by finding common identifiers and stopping the text saving process early when they were identified.

# Limitations

One of the largest limitations is the reliance on the Ad Fontes media bias chart to determine the overall reliability and skew of the news articles. The chart average is determined from a set of ~10 -15 articles read by three different analysts so, the actual average of the articles could be in-exact. For example, some IJR articles appear to be somewhat right leaning and sensational, which could place its true skew right of center. The size of the articles is also very different. The IJR articles are clearly much shorter than both the Federalist and Truthout articles based on the number of articles and and the corpus token volumes. This could mean the IJR corpus articles are more to the point vs. the other two corpora. A keyword comparison could pick up words that mainly come from the difference in writing styles as compared to the difference in content which is desired. However, this could provide interesting concordance comparisons since the hyper-partisan articles are likely more charged. Another limitation is the size of the Truthout corpus when compared to the other corpora. It is about half the size of the IJR corpus in terms of tokens and close to 30% smaller than that of the Federalist. When comparing corpora for keyword analysis there is some debate on the need for a large reference but having such a small corpora could reduce the number of keywords that are found (Baker, 134). This also translates into collocation analysis. Because all of the corpora are quite small, the log-likelihood estimator and mutual information scores could miss low frequency but interesting collocates.

## References

Baker, Paul. *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press, 2010.

*REPORT OF THE SAGE WORKING GROUP ON VACCINE ..* www.who.int/immunization/sage/meetings/2014/october/1\_Report\_WORKING\_GROUP\_vaccine\_hesitancy\_final.pdf. Accessed 18 Aug. 2021. Accessed 18 Aug. 2021.

Ritchie, Hannah, et al. “Coronavirus (Covid-19) VACCINATIONS - Statistics and Research.” *Our World in Data*, 5 Mar. 2020, ourworldindata.org/covid-vaccinations?country=USA.

“Fauci: Politics Is the Reason Canada Has Surpassed U.S. on COVID Vaccines | CBC News.” *CBCnews*, CBC/Radio Canada, 21 July 2021, [www.cbc.ca/news/world/fauci-us-canada-vaccines-1.6109362](http://www.cbc.ca/news/world/fauci-us-canada-vaccines-1.6109362).

“Interactive Media Bias Chart.” *Ad Fontes Media*, 16 Aug. 2021, [www.adfontesmedia.com/interactive-media-bias-chart/](http://www.adfontesmedia.com/interactive-media-bias-chart/).

“Python Requests. 403 Forbidden.” *NewbeDEV*, newbedev.com/python-requests-403-forbidden.

Ignatow, Gabe, and Dragomir Radev. *Text Mining: A Guidebook for the Social Sciences*. SAGE, 2017.

“CSS Selector Reference.” *CSS Selectors Reference*, [www.w3schools.com/cssref/css\_selectors.asp](http://www.w3schools.com/cssref/css_selectors.asp).

“Scrapy 2.5 Documentation.” *Scrapy 2.5 Documentation - Scrapy 2.5.0 Documentation*, 7 Apr. 2021, docs.scrapy.org/en/latest/.

“Beautiful Soup Documentation.” *Beautiful Soup Documentation - Beautiful Soup 4.9.0 Documentation*, www.crummy.com/software/BeautifulSoup/bs4/doc/.