Team Members: Alfredo Gomez, Bryan Uribe

News Broadcast's Topics in Times of Pandemic

Introduction

In mid April 2020, the Becker Friedman Institute published a study on the effects of misinformation during pandemic (Bursztyn et al). This study looked at two popular news shows, Hannity and Tucker Carlson Tonight. While both of these are from Fox news, in the early stages of the pandemic Carlson warned of the dangers of the pandemic while Hannity dismissed any concern for it. The study found that greater viewership of Hannity relative to Tucker Carlson Tonight is strongly associated with a greater number of COVID-19 cases and deaths in the early stages of the pandemic.

Seeing the drastic effects of misinformation when it comes to contracting COVID-19, or, in some cases, dying from it, one comes to ask what these news shows are even saying. Our work here attempts to answer this question through a statistical method from the world of NLP, topic models.

Data

Our data was acquired through web scraping CNN and Fox news transcripts. In total, we acquired 354 different news articles or TV broadcasts from early January to May 13th having no bias on the articles chosen. The data contains most of all articles, leaving out some parts of texts which were unreachable through a web scraper. This text was not inherently the body of the article, only an introduction. Moreover, since we used a web scraper, parts of our data has unimportant text such as audio cues, and names signifying the person who spoke the text. We emphasized to eliminate some of this noise through effective tokenizing which will cut out meaningless text for LDA

Methods

As mentioned earlier, our main method for analysing our data is through the use of topic models, specifically we will use Latent Dirichlet Allocation (LDA) to probabilistically generate topics trained on our entire dataset. Once we have our list of topics, we select the most probable topic for a given transcript. This allows us to see how the topics are distributed across the two different news stations. Finally we qualitatively analyze the aforementioned distributions and discuss the trends we find.

Results

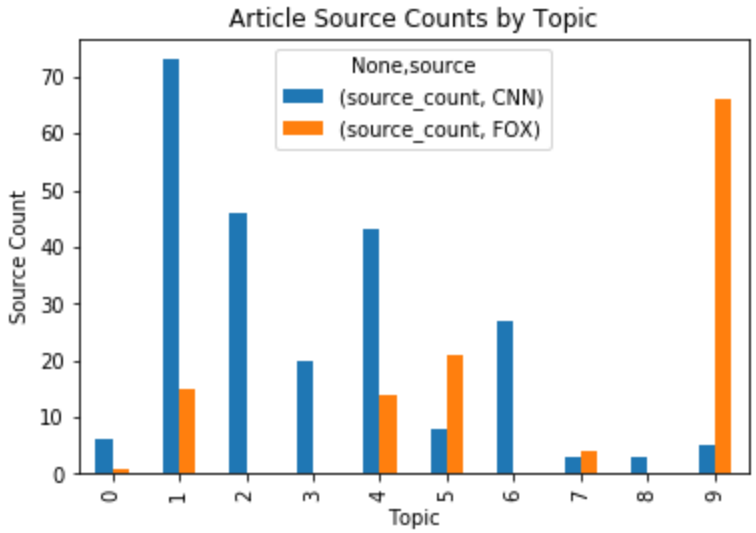
Upon using initially using LDA to find our initial list of topics we found the following:

Topic #												
0	like	think	know	greg	right	going	dana	yes	gutfeld	williams	perino	watters
1	new	10	cnn	thank	using	think	know	said	people	cuomo	vaccine	sued
2	virus	click	news	said	people	greg	china	2020	president	trump	media	opinion
3	new	said	january	people	china	president	trump	000	say	obama	stone	facts
4	house	said	president	trump	trial	senate	evidence	ukraine	biden	schiff	impeachment	witnesses
5	like	virus	cnn	medical	dr	gupta	know	people	health	19	really	care
6	like	think	know	said	want	people	president	trump	say	right	going	unidentified
7	like	think	know	people	president	say	right	going	biden	gutfeld	williams	watters
8	like	think	know	lot	want	people	iran	work	right	going	ve	really

We immediately noticed we would need more stop words to account for the fact that the news transcript would often feature the names of the people speaking as well as the name of the news station. Making that adjustment gives us the following set of new topics:

Topic #												
0	dr	gupta	think	know	people	time	right	ve	jones	pete	marie	lawrence
1	new	10	click	news	thank	using	good	said	people	2020	trump	sued
2	according	transcript	house	said	january	told	president	trump	administration	obama	ukraine	facts
3	house	said	case	president	trump	trial	senate	republicans	impeachment	democrats	witnesses	bolton
4	virus	dr	think	know	want	people	president	say	right	going	ve	really
5	click	news	said	people	2020	president	trump	media	democratic	newsletter	bernie	sanders
6	world	international	10	version	transcript	news	events	president	iran	prince	soleimani	iranian
7	good	think	got	know	right	going	oh	yeah	yes	jones	laughter	hegseth
8	says	said	white	vaccine	trump	yeah	impeachment	laughs	book	waters	wine	cave
9	think	know	video	people	president	trump	say	right	going	clip	yes	crosstalk

Nowe we count the number of transcripts per news station for each topic to obtain the following plot:



Discussion

Our goal was to use LDA to find topics that news outlets were publishing to the public in an attempt to identify misinformation. Through the use of web scraping we were able to acquire sufficient data to highlight the prime topics that were prevalent in news outlets during this COVID 19 period. While web scraping was efficient at gathering data, we realized that data cleaning was a key role to predicting the key topics in our articles that we studied.

All in all, the topics that we found to be most common dealt with flashy news headlines and articles that aim to grab the public's attention. While this may not be different than pre-pandemic and while there may not be evidence for misinformation, we found to our disappointment that these large news outlets did not publish more information on public safety. The article by Becker Friedman Institute highlights that misinformation is a cause for early stage pandemic deaths, and furthermore we argue that subsequent deaths may be caused from a lack of information.

References

Bursztyn, Leonardo and Rao, Aakaash and Roth, Christopher and Yanagizawa-Drott, David, Misinformation During a Pandemic (April 19, 2020). University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2020-44. Available at SSRN: https://ssrn.com/abstract=3580487 or http://dx.doi.org/10.2139/ssrn.3580487

Lettier. “Your Guide to Latent Dirichlet Allocation.” Medium, Medium, 31 May 2019, medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d.