Alex Bishka, Eve Kazarian
Math 189Z

Math 189Z Final Report

Topic and Sentiment Distribution Analysis across Various COVID-19 Media

Motivation

The motivation behind this project comes from one of the reasons why climate change has taken so long to become recognized, which is miscommunication between the scientific community and the general public. Therefore, we wanted to inspect this situation with COVID-19. More specifically, we were curious to see if the public, media, and scientific community all have the same understanding of the situation, and to what degree miscommunication is present. In order to explore this question, we performed topic and sentiment distribution analysis across those three levels of society on various texts: tweets for the public, articles for the media, and scientific articles for the scientific community.

Methods

The data we worked with were a collection of scientific articles, TV broadcasts, and tweets from recent months (January - April 2020). We divided each media file into smaller data frames by month. These data frames ultimately contained a collection of text snippets from their corresponding media: abstracts from scientific articles, fragments of conversation from broadcasts, and tweets from Twitter. Arranging the data in this way allowed us to perform topic and sentiment analysis on the media content. For this project, we had a lot of TV broadcast sources, so we chose to perform analyses on 3 different sources instead of all of them: BBC, Al Jazeera, and Reuters.

For our topic modelling, we chose to use the popular modeling technique known as Latent Direlect Analysis (LDA). This technique breaks down documents into a hidden layer of topics, which then provides a set of words that is representative of the topic. Thus, both the words and documents are modeled by a set of topics. Words removed from any of the sets were represented in our list of stop words, which was consistent throughout each LDA. In our LDA, we used three documents to generate five sets of words and then based off of those sets qualitatively determined the topics from that pool of documents. Typically, our three documents represented all of the posts (tweets, news articles, or scientific articles) from the three previous months before the shutdown - January, February, March - with each document representing one month. However, certain sources of data did not always represent those three specific months and so what time period of posts each document represented shifted around, but we attempted to keep the timeline of posts as close to possible to the COVID-19 shutdown. There was one exception to this rule, where for the scientific community we had data that ranged several years

into the past, so we decided to have one document represent the set of scientific articles in the past ten years, another for the past five years, and the last for the past year. For the sake of brevity, we shall only discuss the LDAs for the scientific articles for the time around the COVID-19 shutdown, tweets, and a few example news outlets. Additional details about the news sources and LDA can be found in our jupyter notebook.

To analyze the sentiment across the scientific articles, TV broadcasts, and tweets, we used VADER, the Valence Aware Dictionary and Sentiment Reasoner for Python. This module contains a tokenizer which allows us to split text into words. We tokenized the text snippets for each media and computed how positive, negative, and neutral a snippet was. We categorized a text snippet as positive if its positive score outweighed its negative and neutral scores (we applied the same logic to negative and neutral text snippet categorization). Each time a snippet was categorized, we added to a corresponding sentiment count (e.g. the count for positive text would increase by 1 if we encountered a positive text snippet). After searching through all text snippets for a media, we computed its percentage of positive, negative, and neutral snippets. We also computed a negative-to-positive sentiment ratio (if applicable) to see which sentiment was predominant in that media. To understand how VADER was classifying our text, we also printed out the positive and negative text snippets to view them.

Results

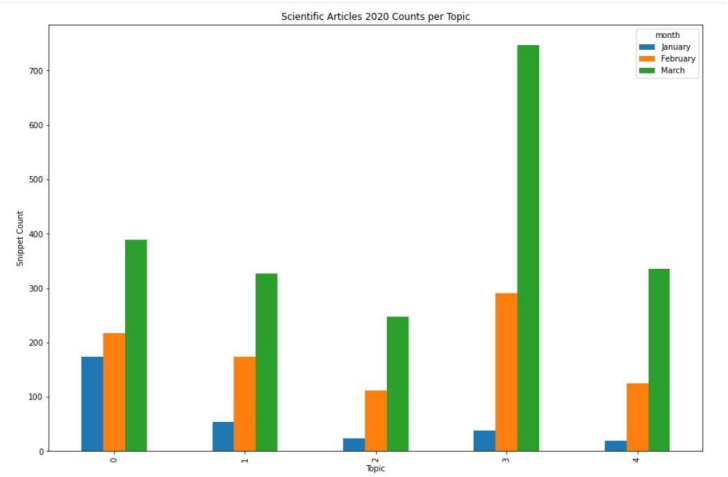

Figure 1. Scientific Article LDA

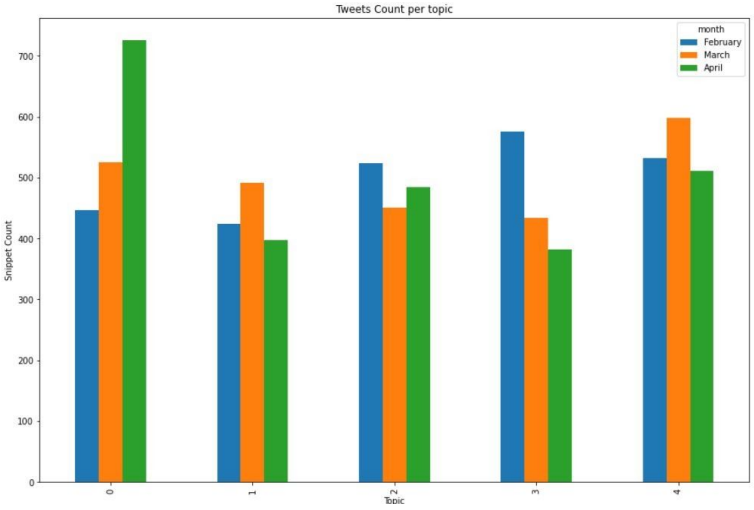

Figure 2. Twitter LDA

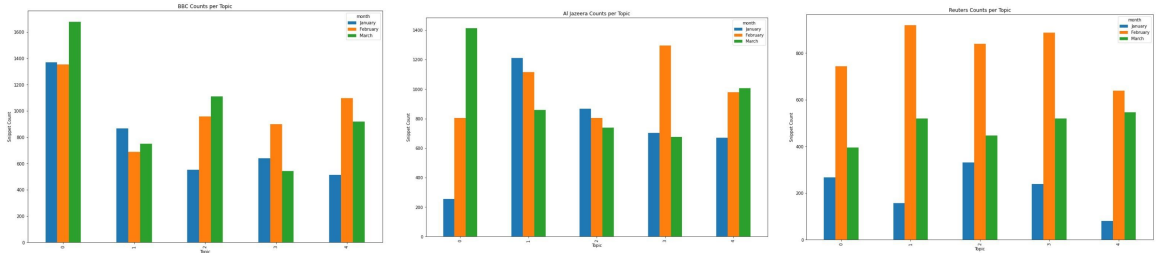

Figure 3. From left to right, LDA, for BBC, Al Jazeera, and Reuters

Our LDA plots indicate that there was a lot of coverage of COVID-19 in the media prior to the month of shutdown, March (Figure 3). Additionally, social media seemed to reflect this trend (Figure 2), while the scientific community seemed to disseminate most of its information on COVID-19 during March (Figure 1).

What we qualitatively determined each topic to be is shown in Figure 4.

Source	**Topics**
Scientific Articles	**Topic 1:** how infection works, **Topic 2:** potential vaccination, **Topic 3:** lifecycle of the virus, **Topic 4:** global consequences, **Topic 5:** effects of the virus and how to detect infection
Twitter	**Topic 1**: initial response in Wuhan, **Topic 2:** government responses to crisis, **Topic 3:** UK's reaction and world wide quarantine measures, **Topic 4:** new discoveries on Wuhan situation, **Topic 5:** Trump's response to the situation
BBC	**Topic 1:** initial reaction to COVID-19 situation, **Topic 2:** outbreak in

	Wuhan, **Topic 3:** world reaction to COVID-19, **Topic 4:** decision to quarantine, **Topic 5:** response in UK and Italy:
Al Jazeera	**Topic 1:** world's initial reaction to COVID-19, **Topic 2:** government uncertainty over the situation, **Topic 3:** ground zero, **Topic 4:** situation in China around time of outbreak, **Topic 5:** global emergency
Reuters	**Topic 1:** WHO's response to the crisis, **Topic 2:** situation of world economy, **Topic 3:** quarantine situation and its impact on Europe, **Topic 4:** health of markets as a result of growing fears, **Topic 5:** impact of Russian oil problems on the economy

Figure 4. Topics for each LDA

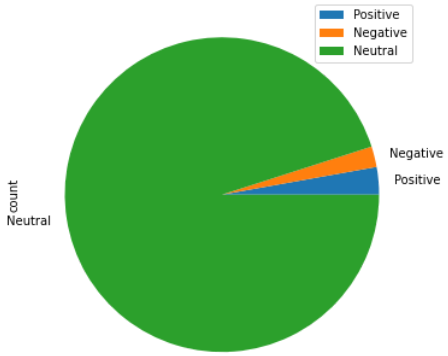

Figure 5. Tweet sentiment distribution.

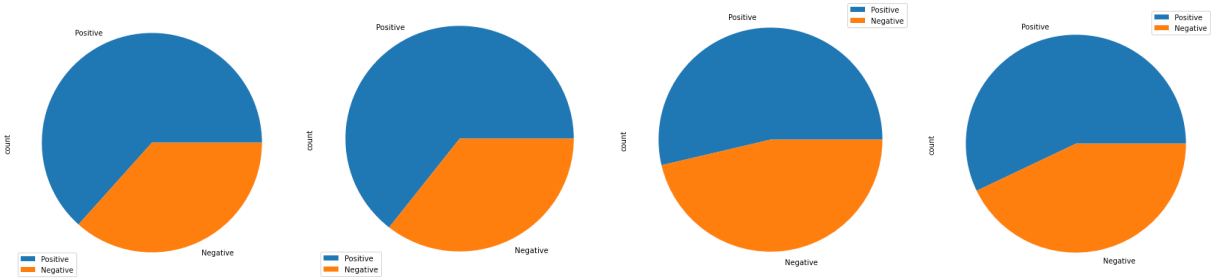

Figure 6. From left to right, sentiment distributions for BBC, Al Jazeera, Reuters, and Tweets.

Source	**% Positive**	**% Negative**	**% Neutral**
BBC	1.12	0.65	98.23
Al Jazeera	0.49	0.27	99.24
Reuters	0.97	0.84	98.19

Twitter	2.83	2.12	95.05
Scientific articles	0	0	100

Figure 7. Sentiment distribution as percentages.

Our sentiment analysis indicated that all of the media-- the scientific articles, TV broadcasts, and tweets-- had predominantly neutral content. In Figure 5, we see a representative sentiment distribution. However, the TV broadcasts and scientific articles notably contain more neutral content than Twitter. The scientific articles are 100% neutral, the new sources range from 98-99% neutrality, and the tweets are 95% neutral (Figure 7).

Furthermore, the ratio of negative-to-positive sentiment text differs among the media. BBC and Al Jazeera had notably more positive content than Reuters and Twitter (Figure 6). Some examples of the text snippets that qualified as negative and positive for each media are shown in Figure 8.

Media	**Positive Sentiment Content**	**Negative Sentiment Content**
BBC	can ai help to treat it?	but the worst affected place outside china is italy where148 people have now died.
	a lucky survivor being pulled alive from the rubble.	angry protests on the streets of jericho.
Al Jazeera	throughout the day safe and saved.	quite problematic logistically right.
	or miracle that is.	brazil's worst industrial disaster almost a year ago.
Reuters	a pleasure thank you.	brazil's worst industrial disaster almost a year ago.
	welcome.	net to the collapse.
Twitter	Please join me in lifting our voices to God.	Now, Spain is desperately fighting COVID-19.
	corona can’t stop our love!	Biden got it wrong.

Figure 8. Selected text snippets from different media grouped by sentiment categorization.

Discussion

With respect to the LDA analysis, we chose to focus on three seemingly different news outlets: Al Jazeera (a Qatari TV channel), BBC (a British news source), and Reuters (a news source largely concerned with financial market data). The characteristics of these outlets are reflected in the topics uncovered using LDA. The stop words we used were common to all the news sources so that we could uncover any true topic differences between them. Interestingly enough, they have all seemed to focus on COVID-19 at separate times, with BBC spending the most time on it in January, Reuters in February, and Al Jazeera in March. However, all three of these news sources do have coverage throughout these three months. Furthermore, they focus on issues more relevant to their purpose, with BBC focusing on the UK, Al Jazeera on the Middle East, and Reuters on the economy.

Twitter seems to be most interested in COVID-19 in April, but it still seemed to be a popular topic in February and March. The twitter space focuses a lot on the initial outbreak and government response to the outbreak, with a bias towards how the US responds.

The scientific community has published most of its papers in March, likely because of the time it takes to conduct a study on the current COVID-19 strain and publish a scientific article. Additionally, the scientific community focuses more on topics that deal with infection, understanding, and treatment of the disease.

Overall, the topics covered between the media and the public are pretty similar. Both tend to focus on the initial outbreak and government response to the situation, basically the impact of the virus on society. The scientific community focuses more on how the virus impacts an individual and treatments to the virus. Therefore, there does not seem to be too much miscommunication between the three levels of society. It probably would be better - from a communication and understanding standpoint - if the media and the public topic space reflected more on how the virus works, rather than how to respond to it and how it impacts the world. However, it does seem the scientific community has gotten the point across of the urgency and danger of the situation, and so it appears that the information on COVID-19 has been properly disseminated between these levels of society.

With respect to the sentiment analysis, the fact that all media have majority neutral content is a good sign that the analysis is working properly. Scientific articles are meant to be written in an unbiased tone. The TV broadcasts are reporting on the status of the virus, rather than giving an opinion on it. Twitter should and does have the least amount of neutral content out of all of the platforms because it not only propagates news but also allows people to share their worries, support, exclamation, and prayers.

As we see with the selected positive and negative tweets, they are usually very emphatic: “corona can’t stop our love!” and “Biden got it wrong.” We note that the amount of positive and negative tweets is roughly the same because for every worry there appears to be a tweet about hope and support.

Additionally, the nearly one-to-one ratio of negative-to-positive content for Reuters likely occurs because anything that is classified as positive is a standard greeting or polite remark (which is expected from a conversation), but anything negative is related to the actual status of the virus. Reuters, out of all of the platforms, is doing the most reporting on the state of the financial market; this negative financial content is indicated by words like "industrial," "collapse," and "net" in the selected negative tweets. Thus, it should have a higher negative-to-positive content ratio.

The other TV broadcasts not only have the positive cordialities but also positive and hopeful content. For example, Al Jazeera and BBC have content about potential solutions and saved lives: "can ai help to treat it?" and "or miracle that is." Thus, their ratios of negative-to-positive content are lower.

Conclusion

Our motivation was to understand whether these various media outlets were on the same page when covering coronavirus, and it appears that each platform is responding to the virus accordingly. Twitter and news outlets were quick to disseminate information while scientific articles are coming out with hopeful research breakthroughs. There are interesting trends we would expect to see, such as different sentiment and topic distributions across the platforms. The different sentiments and topics are expected and indicate that each platform is spreading COVID-19 information in a unique yet comprehensive way.

References

"Twitter Sentiment Analysis Using Python." *GeeksforGeeks*, 7 Feb. 2018, www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/.

"Text Analysis Part Two: Sentiment Analysis With the Natural Language Toolkit." *Sentiment Analysis With the Natural Language Toolkit | Archives Unleashed*, cloud.archivesunleashed.org/derivatives/text-sentiment.

Malik, Usman. "Python for NLP: Sentiment Analysis with Scikit-Learn." *Stack Abuse*, Stack Abuse, stackabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/.

Ganegedara, Thushan. "Intuitive Guide to Latent Dirichlet Allocation." *Medium*, Towards Data Science, 27 Mar. 2019, towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158.