News Reporting on COVID-19: Final Report

Marzena Hurtado, Christian Kheav, Rina Kim, and Arthur Wu

INST 447 Spring 2021

Section 0101

Instructor: Prof. Aditya Bhat

May 15, 2021

Introduction

COVID-19 has been a huge topic for many news organizations around the world as the effects of this pandemic has been felt across the world for over a year. Throughout this pandemic, many different news articles have been written about death tolls, relief, vaccinations, reopenings, etc., and were exhibited through daily news cycles via the publishers' websites or via a social media platform. Social media currently plays a huge role in spreading news feed across the world due to the massive pool of users daily.

Our goal for this project has been to analyze different characteristics of top posts from news organizations related to COVID-19 on social media platforms. Our analysis consists of data from Facebook, which Pew Research Center states is regularly used as a source of news by a significant 36% of the American population, (Shearer & Mitchell, 2021). The characteristics of each post that we have chosen to include are the number posts, shares, audience interaction, and appearance of words signifying topic focus. The news organizations included in our data collection consist of CNN, The New York Times, The Washington Post, NBC News, Fox News, BBC News, Wall Street Journal, ABC News, The Guardian, and USA Today. Our main research questions are:

- **How do different major news organizations cover COVID-19 on Facebook over time?**
- **How do different news organizations differ in public engagement on Facebook?**

Data

The data we used for our project includes data on Facebook posts posted by the mentioned ten news organizations. We were able to obtain raw data on 16,822 posts, with various numerical and categorical characteristics of the post. After narrowing down the dataset to only posts related to the topic of COVID-19, our final dataset for analysis contained 4,557 entries and 13 columns. The variables that we used as columns include number of shares, number of comments, news organization, and post text. A complete list of our dataset columns can be found in the appendix.

Below includes the summary statistics for our numerical variables. Examining the initial statistics, we can see that posts by CNN have the highest average number of shares (2,431.80

shares) and posts by Fox News have the highest average number of comments (13,180.47). We can also see that the total number of posts differs between each news organization. For example, the dataset contains 690 CNN posts, but on the other hand, there are only 115 Fox news posts. We will account for this in our analysis by performing analysis based on the averages of the variables rather than analysis based on frequency. Here are summary statistics for our more notable numerical variables:

Shares:

News Org.	count	mean	std	min	25%	50%	75%	max
ABCNews	588.0	404.829932	1561.716320	2.0	46.00	119.0	323.50	32944.0
BBC	228.0	1546.197368	3590.458036	37.0	225.25	459.0	1315.50	42024.0
CNN	690.0	2431.800000	14009.135358	23.0	158.00	388.0	1017.25	295735.0
FOX	115.0	1824.826087	3152.908956	68.0	398.00	767.0	1931.50	26810.0
NBC	465.0	459.580645	1343.870335	1.0	43.00	121.0	365.00	16638.0
NYT	526.0	704.153992	2052.379358	3.0	58.00	155.0	472.00	21620.0
TheGuardian	603.0	283.456053	1075.046215	1.0	29.00	68.0	184.50	18481.0
USAToday	638.0	266.084639	1704.567940	6.0	46.00	79.5	171.25	40862.0
WASH	190.0	773.421053	2107.542462	6.0	95.25	269.5	648.75	25729.0
WSJ	919.0	167.095756	1104.625265	1.0	8.00	17.0	46.00	22492.0

Total Reactions:

News Org.	count	mean	std	min	25%	50%	75%	max
ABCNews	588.0	2061.226190	5408.932212	25.0	331.0	684.5	1797.00	94437.0
BBC	228.0	8275.956140	11994.669307	339.0	2092.0	3948.0	9073.75	85841.0
CNN	690.0	8654.410145	23061.333341	127.0	947.5	2437.0	6208.00	326292.0
FOX	115.0	13180.469565	19438.834266	438.0	3294.0	6192.0	12996.50	143914.0
NBC	465.0	1306.195699	2177.120771	5.0	251.0	589.0	1351.00	18835.0
NYT	526.0	3734.524715	8119.373915	8.0	434.5	1051.0	3038.50	71543.0
TheGuardian	603.0	1222.645108	2917.416183	22.0	199.5	432.0	1078.50	37923.0
USAToday	638.0	1280.769592	2270.428932	162.0	388.0	651.5	1221.25	26784.0
WASH	190.0	3057.689474	4258.546101	76.0	622.0	1473.5	3245.25	30940.0
WSJ	919.0	968.129489	4351.343866	11.0	72.5	150.0	419.00	65134.0

Comments:

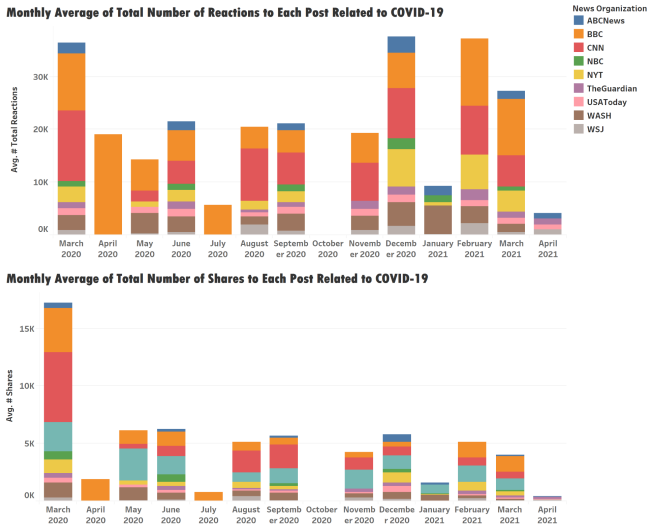
News Org.	count	mean	std	min	25%	50%	75%	max
ABCNews	588.0	130.707483	556.390255	0.0	4.0	12.0	45.25	8133.0
BBC	228.0	302.644737	1355.286848	0.0	8.0	27.0	98.00	16627.0
CNN	690.0	452.976812	2242.646247	0.0	6.0	23.5	114.00	33312.0
FOX	115.0	2712.426087	5785.519736	0.0	121.5	594.0	2507.00	46763.0
NBC	465.0	179.105376	474.496121	0.0	2.0	9.0	77.00	3389.0
NYT	526.0	386.129278	1820.877489	0.0	2.0	9.0	63.75	27911.0
TheGuardian	603.0	99.971808	334.726354	0.0	1.0	5.0	32.50	3035.0
USAToday	638.0	107.954545	349.077941	0.0	1.0	7.0	49.00	4432.0
WASH	190.0	631.836842	1631.985728	0.0	4.0	29.5	362.00	14117.0
WSJ	919.0	45.078346	334.133882	0.0	0.0	2.0	9.00	8520.0

Facebook Categorical Variable Summary Statistics for posts related to COVID-19 (before discarding posts with no text content):

News Organizations:

```
count     4962
unique      10
top        WSJ
freq       919
Name: News Org., dtype: object
```

At the beginning stage of our analysis, we ran preliminary visualizations for the monthly average number of reactions and shares on FB. These visualizations were created using Tableau.



From this initial graph, we can see that there are no posts for the month of October, and only BBC news posts for the months of April and July. This may be a feature of the Facebook API, where posts from some months are hidden, or we believe that this may be caused by the news organizations' undergoing mandated quarantine during those months, causing a lack of posts.

Data Fetching and Manipulation

We used Facepager to collect raw data on news organization posts related to COVID-19 on Facebook. Facepager is a tool that allows for easy data fetching from websites like Facebook, Twitter and YouTube through website APIs and web scraping. We chose to use Facepager because it allows us to fetch data using the Facebook API easily through an intuitive UI.

Facepager limits the fetching rate to 100 posts per run. Therefore, we decided to fetch 100 posts for every 7-day week starting from 03/01/2020 up to the week ending 04/10/2021. This allowed us to maximize the number of posts we fetch for each week for the entire year in which COVID-19 was a hot topic. Our team fetched data following this protocol from all news organization account pages, and saved each compiled dataset to .csv format by news organization account. After fetching data using Facepager, we had 16,822 rows of data.

To clean our Facebook data for analysis, we first combined the datasets of all news organizations into one large dataset. We then confirmed that all columns matched the correct data type; we converted all numerical columns to the float64 data type to maintain consistency across datasets, and converted the 'time posted' column to datetime format. We also removed all invalid rows by filtering out rows with empty cells. Lastly, we filtered our dataset to contain only posts that contained a keyword (covid, 'coronavirus', 'pandemic', and 'vaccine') in the 'Post's title' or

'Post's text' columns. After cleaning and manipulation, our Facebook dataset contained 4,557 entries and 13 columns.

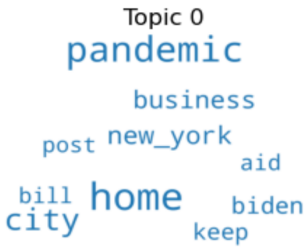
Topic Analysis

We conducted topic analysis using skills and knowledge learned in the course of this semester's INST447 class lectures as well as instructions in the online data modeling article by Machine Learning Scientists, Selva Prabhakaran (Prabhakaran, 2019). To prepare the text in each post and it's title for analysis, we used the following libraries available in python: numpy, pandas, datetime, regex, nltk, spacy, gensim, pyLDAvis, matplotlib and wordcloud. The pre-processing of text data included removal of URLs, whitespace and non-alphanumeric characters, as well as transformation of text to lowercase, tokenization of the content, and finally running a bigrams detection model using gensim's Phraser() and applying them to the textual data. To avoid unnecessary noise, we also conducted the removal of stop-words using nltk's standard English stop-words library with added few custom words which were determined as frequently appearing yet irrelevant in terms of topical analysis (e.g., "HAPPENING NOW", "Breaking News").

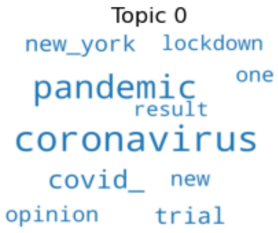
The topics characterised by a set of keywords returned from the gensim's LDA models were ultimately visualized using pyLDAvis and the word-clouds shown in the appendix of this document as well as the project's Jupyter notebook. Based on evaluation of these visual presentations, it is evident that the three main topics addressed by each organization are quite distinct, even though they share some of the keywords. We can make this assumption based on the observation of distances between circles corresponding to each topic in pyLDAvis models for each of the ten organizations (i.e., no overlapping circles). The differences between organizations are more subtle, but some trends are noticeable. For example, the term "vaccine" is the most used and among top keywords for aggregated posts by most analyzed organizations except FOX and Washington Post for which the term "vaccine" doesn't even appear among the top eight used keywords.

Another interesting example of insight is comparison of main topics addressed by two organizations, Wall Street Journal and the New York Times (both based in New York City). While each organizations' main topic related to COVID-19 circles around local aspects, the tone sensed from the respective word-clouds is strikingly different.

Wall Street Journal

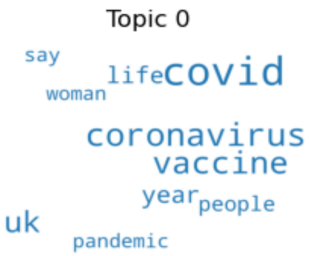


The New York Times



Similarly, looking at the main topics addressed by The Guardian and BBC, both British organizations, the trend in choice of words has a different feel:

The Guardian

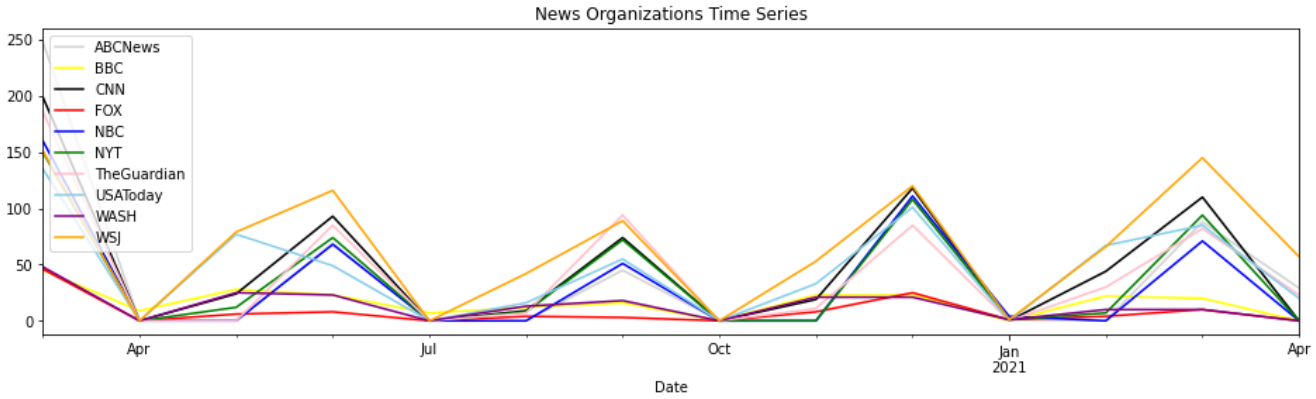


BBC

Another interesting observation is that while the time of posting that we are analyzing encompasses mostly the period during the Trump presidency, the keyword “biden” appears more often (6 occurrences) than the keyword “trump” (4 occurrences) among the keywords of top three topics for all ten organizations.

Lastly, we discovered that the “death” keyword holds a prominent place in posts by the Washington Post (7th most used term) and the New York Times (9th most used term).

Time Series Analysis

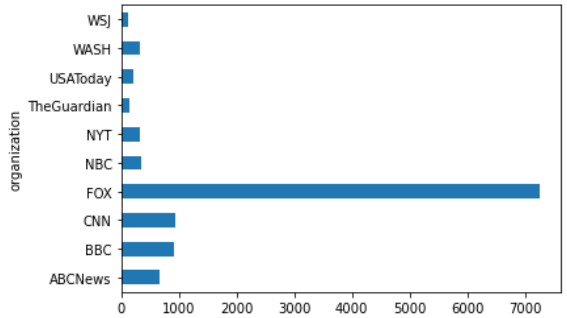


Covid-19 has been a topic of concern for over the past year as it has driven the world into a global pandemic which has caused billions to be affected and millions of deaths across the world. Death is just one of the topics that news organizations like ABC News and CNN have covered and posted about throughout Facebook over the past year. We used monthly time series analysis through the use of a multiple line graph to analyze the number of posts that each news organization has posted related to any topic of Covid-19 over a year worth of data collected. From the time series visualization we can see that after the month of April 2020 which was the beginning of the national lockdown, Wall Street Journal is the news organization that has posted the most about the topic of Covid-19 out of the other 9 most popular news organizations we had used for the study. The organization with the second highest number of Covid-19 related posts was CNN after the month of April 2020. While Covid-19 has been a topic of news for the past year it is surprising to see some news organizations not covering the topic like other news organizations.

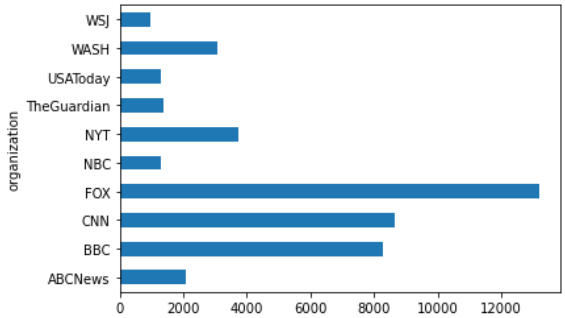
Public Engagement Analysis

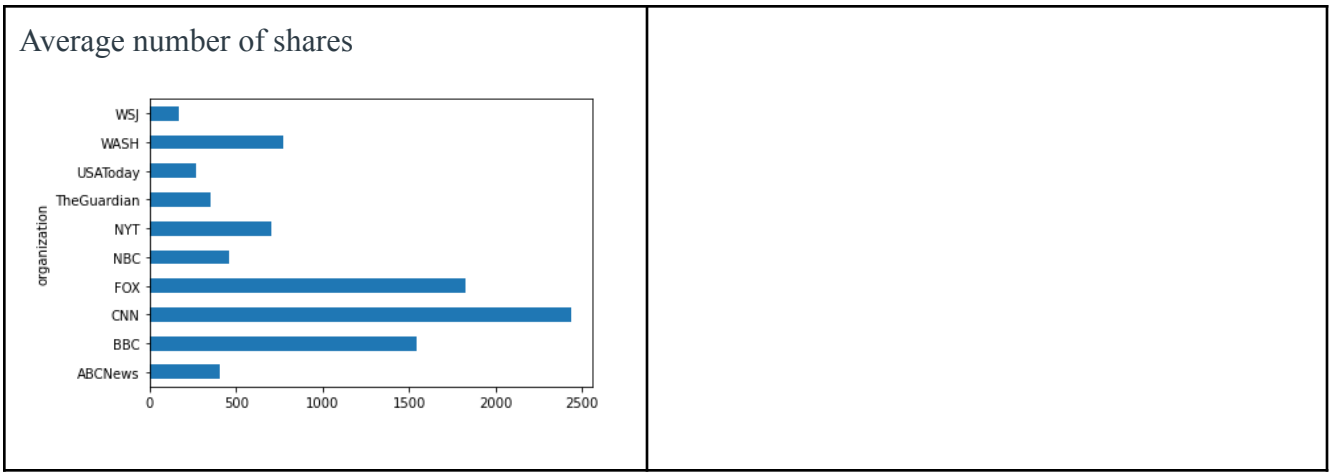
Another area of our data that we wanted to study was the public engagement with posts posted by different news organizations on Facebook. To do this, we analyzed the numerical variables of our dataset including columns such as number of shares, number of total reactions and number of comments. We used bar graphs grouped by organization and line graphs spanning the date time variable to show how the average number of comments, reactions and shares differed among news organizations.

Average number of comments



Average total reactions





From the bar graphs, we can see that FOX news leads all other news organizations in average number of total reactions and average number of comments. This suggests that the Facebook community of Fox News followers is larger and more active than other news organizations like TheGuardian and New York Times. We can also see that CNN posts have the highest average number of shares. Even though CNN may not have as many reactions or comments, the articles and posts that CNN posts on Facebook may be more shareable than those posted by other news organizations. Organizations such as WSJ and The Guardian have a very low average number of reactions, shares and comments, which is understandable since they are less popular than large news organizations like Fox and CNN. They may also not cover COVID-19 as diligently as other news organizations.

Conclusion

Through our research and analysis, we wanted to determine how different major news organizations cover COVID-19 on Facebook over time and how news organizations differ in public engagement on Facebook. We believe that through our topic, time series and engagement analysis, we were able to discover interesting results related to our research questions. Despite all covering COVID-19 as a general topic, we can see that major news organizations focus on different aspects within COVID-19 in their posts on Facebook. Some news organizations such as The Guardian cover COVID-19's impact on daily life, while other news organizations such as Fox news cover the political aspect of the pandemic. Through our time series analysis, we found that surprisingly, Wall Street Journal led other organizations in COVID-19 posts after April 2020. Lastly, we discovered trends in public engagement where Fox news posts receive the most

engagement based on average number of reactions and comments, but CNN posts surprisingly had the most average number of shares. This analysis can be further improved through additional Facebook post data from each news organization. Our data only contains approximately 5,000 entries, and the distribution of posts between organizations are not similar. In the future, efforts to obtain more data could be made in order to refine our analysis and results.

Appendix

Our team initially wanted to compare Facebook post data with Twitter post data. However, our team had difficulties in fetching data from Twitter. The Twitter API limited the amount of posts we could scrape and prevented us from obtaining an adequate amount of data to perform analysis on. At the end, our team decided to only focus on Facebook data. We attached our twitter data csv and codebook for reference to the attempt.

Variables utilized in our dataset:

Variable	**Description**
organization	Categorical. New organization that made the post/tweet.
date	Date. Date that the post or tweet was posted.
title	String. Title of the post.
post	String. Text content in post
shares	Numerical. Number of times the post/tweet has been shared or retweeted.
comments	Numerical. Number of comments people have left on the post.
total_reactions	Numerical. Number of total reactions.
like	Numerical. Number of times the post has been liked.

care	Numerical .Number of times the post received a care reaction.
haha	Numerical. Number of times the post received a haha reaction.
wow	Numerical. Number of times the post received a wow reaction.
sad	Numerical. Number of times the post received a sad reaction.
angry	Numerical. Number of times the post received an angry reaction.
love	Numerical. Number of times the post received a love reaction.

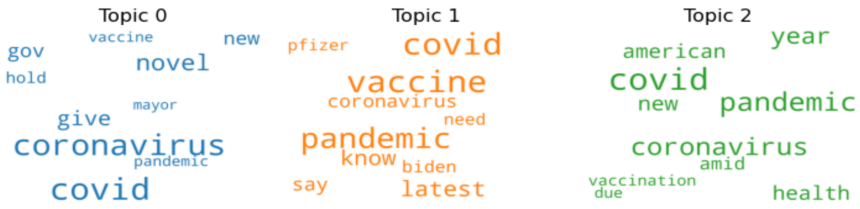
Number of Covid-19 related FB posts per month by news organizations

yearMonth organization	2020-03	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09	2020-11	2020-12	2021-01	2021-02	2021-03	2021-04
ABCNews	248	0	0	69	0	0	45	0	108	1	0	88	29
BBC	45	9	28	23	7	12	16	23	23	0	22	20	0
CNN	199	0	24	93	0	9	74	19	118	0	44	110	0
FOX	46	0	6	8	0	4	3	8	25	1	4	10	0
NBC	159	0	0	68	0	0	51	0	111	4	0	71	0
NYT	150	0	12	74	0	8	72	0	108	1	7	94	0
TheGuardian	81	0	0	18	0	1	18	2	30	0	11	31	7
USAToday	135	0	77	49	0	16	55	33	101	0	67	85	20
WASH	48	0	25	23	0	13	18	21	21	1	10	10	0
WSJ	151	0	79	116	0	42	89	53	120	1	66	145	57

The topic analysis and word models can be located in the topic analysis section of the codebook and below is an example of the top words used by ABC News and the word model of each topic. Main topics and keywords (Example 1):

ABCNews

Overall 8 Top Terms: 1. novel 2. give 3. gov 4. coronavirus 5. vaccine 6. hold 7. year 8. know



References

Mitchell, A., Gottfried, J., Kiley, J., & Matsa, K. E. (2014, October 21). *Media Sources: Distinct Favorites Emerge on the Left and Right*. Pew Research Center. https://www.journalism.org/2014/10/21/section-1-media-sources-distinct-favorites-emerge-on-the-left-and-right/

Prabhakaran, S. (2019). *Topic modeling visualization – How to present the results of LDA models?* Retrieved from ML+ Let's DataScience: https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/

Shearer, E., & Mitchell, E. (2021, January 12). *News Use Across Social Media Platforms in 2020*. Pew Research Center: Journalism & Media. Retrieved March 12, 2021, from https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/

Turvill, W. (2020, October 20). *The top ten most-followed news accounts on Twitter*. Press Gazette. https://www.pressgazette.co.uk/the-top-ten-most-followed-news-accounts-on-twitter/

Watson, A. (2020, September 22). *Most popular news websites in the United States as of August 2020, by unique monthly visitors*(in millions)*. Statista. https://www.statista.com/statistics/381569/leading-news-and-media-sites-usa-by-share-of-visits/