Visualizing The Public Health Narratives Around Covid-19

Analysis of Media Coverage on Common COVID-19 Glossary

Ashish Patidar
Indiana University
apatida@iu.edu

Avinash Pawar
Indiana University
avipawar@iu.edu

Shardul Samdurkar
Indiana University
ssamdurk@iu.edu

Tanay Kulkarni
Indiana University
tankulk@iu.edu

Abstract

During COVID-19, practically everyone throughout the world and in the media had varied perspectives about almost everything related to COVID, whether it was vaccine information, mask usage, mask kinds, or variants of COVID. Varied news sources gave diverse information about COVID-19, thus it's interesting to learn about the different patterns of how different news outlets disseminate particular facts to the general public. The following study intends to uncover trends in the distribution of various news by the media over time, as well as the impact of this news on public opinion on various themes such as social distancing, masking, vaccination hesitancy, and so on.

Introduction

COVID-19 received widespread media coverage, and the media served as a conduit for many different narratives and points of view for the global populace. The study necessitated a collaboration of open source tabular data sources, extensive data preparation to meet stakeholder needs, and the application of multiple visual analyses to the data. The information for the news sources was acquired from both online and television sources. [1]

In terms of television stations, the entire archive of BBC News London, CNN, MSNBC, and Fox News from January 1, 2020, to the present, as well as a selection of days from CNN, MSNBC, and Fox News surrounding past disease outbreaks that received significant coverage on those stations since the beginning of the Television News Archive is being processed. The study began with exploration and understanding of the data we had, followed by a decision on how to preprocess the data and choose which analyses and visualizations would best answer the research questions. Following our analysis of the data, we discovered that it is a time-series dataset with four separate datasets.

The information included what was portrayed as a news source on the internet or on television. We had to assess whether we needed to do a comparative study with earlier pandemics, or if the previous records would have any impact on the current research question because the data source also provided data prior to 2020.

We choose to restrict the data only to the COVID time period, which is roughly after 2020. Temporal analysis, geospatial analysis, topical analysis, and word clouds were the types of analysis we chose for the project. In terms of diverse patterns observed in different timeframes and specific circumstances that took precedence over others, the study provided some amazing insights. In addition, the study revealed some highly fascinating changepoint timeslots, which were also compatible with the real world. [2]

1 Insight Needs

For insight needs we went through a series of discussions amongst the team and we also had a chance to understand what the stakeholders needed in terms of an outcome. The initial traction of thoughts included questions like, how might the rapidly evolving narratives surrounding a topic like Covid-19 be visualized? Imagine the possibilities for public health if you could automatically extract a list of common vaccine hesitancies from news content and visualize them in a narrative map, such as a graph structure, that showed how those concerns were related in terms of cooccurrences in news coverage or semantic connections? Understanding how COVID-related keywords evolved during these three phases of COVID, as well as which media outlets are the most popular (with the most user views) and what narratives they have spread through them, enhanced the 'why' component of our research. To meet these requirements, we decided to focus our visualizations on Stream Graphs, Bar Plots, Line Charts, and Word Clouds.[3][4]

2 Data Acquisition & Data Pre-processing

For all news channels and articles, data is available as JSON objects. Each JSON object contains data of snippet for 30 minutes, so for a day, we will have 48 JSON objects for one news channel. This is evolving data that is updated on daily basis.[5]
Google BigQuery is used to deal with JSON objects as it is easy to generate tables from JSON objects on BigQuery. [6]

2.1 Description of Data

For covid-19 there are three tables available publicly which are extracted from the JSON objects, these tables are mainly used for visualization purposes.[6]

The important tables used for analysis and visualization are

1. Onlinenews dataset: This dataset contains 80219314 entities with 5 attributes, which are:
   - Topic: topic to which snippet is related to
   - Url: url to the article
   - Datetime: date and time of the article
   - Title: title of the article
   - Context: actual snippet or content of the article

   Apart from URL, all attributes are important for visualization from this dataset.

2. Onlinenewsgeo dataset: This dataset contains 175630971 entities and 11 attributes, attributes are as follows:
   - Datetime: date and time of the article
   - URL: url to the article
   - Title: title of the article
   - SharingImage: images associated with the article
   - LangCode: code of the language in which the article is published
   - DomainCountryCode: country code of the domain of the article
   - Location: location from where the article is published
   - Lat: latitudinal coordinates of the location
   - Lon: longitudinal coordinates of the location
   - CountryCode: country code from where the article is published
   - ContextualText: the actual text of the article

Out of these 11 attributes DateTime, title, DomainCountryCode, Location, Contextualtext, lat, and lon are important for visualization and analysis.

3. Tvnews dataset: This dataset contains 4252351 entities and 7 attributes which are as follows:
   - URL: url to video of news
   - MatchDate: date of the broadcast of the video
   - Station: name of the station on which the video was broadcasted
   - Show: name of the show under which video was broadcasted
   - IAShowID: ID of the show
   - Snippet: OCR converted text/caption from the video
   - MatchDateTimeStamp: tine stamp of the video

Of these 7 attributes, all are important except IAShowID.

We have used the above mentioned dataset and performed data pre-processing on the same. We first excluded the redundant columns from the dataset. We also filtered the missing data. We performed aggregations on the dataset using SQL queries and window functions and acquired the required attributes.

3 Analysis Methods

Different analysis methods and algorithms were used for different types of analysis based on our data.

- For geospatial analysis of articles, onlinenewsgeo table is used to create a choropleth map of the number of articles published in a country related to covid.
- For statistical analysis, A bar graph is created from tvnews dataset to find out the top 5 stations that published news regarding covid.
- There were multiple temporal analyses performed using all three datasets:
- Streamline graphs were created from onlinenews and tvnews datasets to see the number of occurrences of the common covid words during different waves of covid. Line graphs are created along with a choropleth map to see the number of occurrences of a particular word used in articles or news stations of a particular country or region.
- For topical visualization and comparison word cloud is plotted for the common words used during covid times. Tvnews dataset is used for this visualization, the number of occurrences of common words is counted using SQL queries from the data set to create a word cloud based on that.

Some of the common covid words are covid19, corona, quarantine, vaccine, covid cases, etc.
Due to time and computational restrictions, snippets and articles of only english language are considered.

4 Visualizations

Depending upon the data we have implemented visualizations such as Line graphs, choropleth map, streamline graphs, word cloud, and bar graphs.

We will include some of the crucial visualizations here. But keep in mind that we have only analyzed the top 5 News agencies from our data which are BBC News, CNN, FOX News, MNCBC, and ALJAZ.

Tableau and Google Data Studio were primarily used for the visualizations.

4.1.1 Line Graphs

For the Temporal data visualization as the data is categorized on daily basis. We have decided to implement the Line graphs for the word Omicron and Delta along with the commutative graph of all 12 words. The data contains the commutative frequency of Covid-19 related word use on a daily basis.

By analyzing the graphs we can clearly see the trends when Omicron and Delta variants were identified, hence the sudden increase in the frequency of that word. Also by looking at the whole cumulative frequency of word usage we can identify the sudden increase in the coverage of Covid-19 related words as more cases are emerging worldwide.

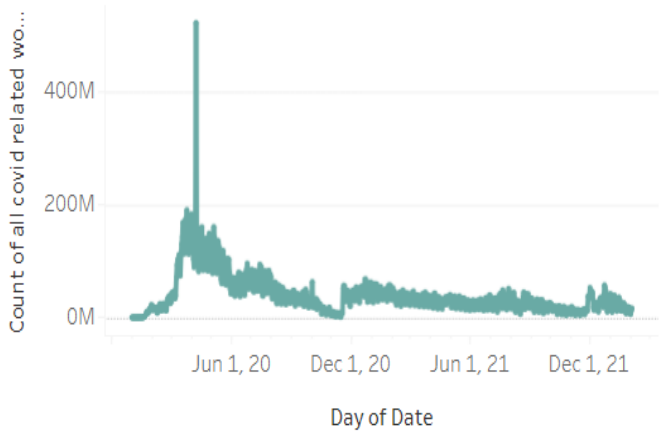

Fig. 1. a. Line graph of the commutative frequency of Covid-19 related words.

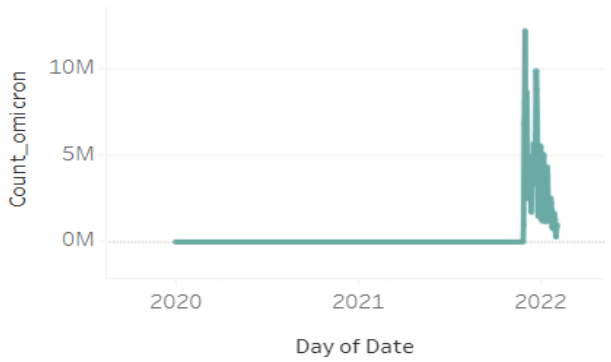

Fig. 1. b. Line graph of the commutative frequency of the Omicron word.

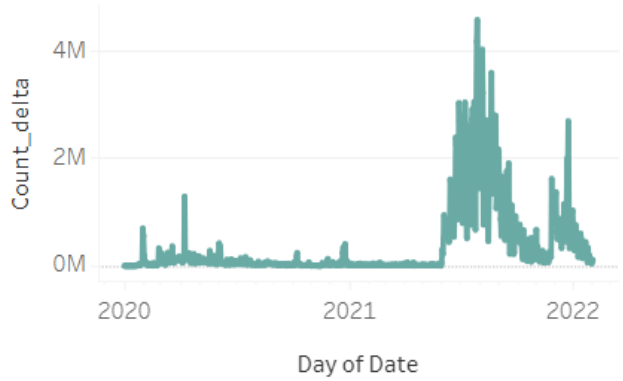

Fig. 1. c. Line graph of the commutative frequency of Delta word.

4.1.2 Choropleth Map

In the case of geospatial data, we have analyzed which county has used Covid-19 Related words and how frequently. We have constructed the choropleth map. So, the results can be easily seen. As we have chosen the English language as our language of study we can see non-English language countries are not on the map. Also, the sheer amount of content in English from the USA dominates the map.

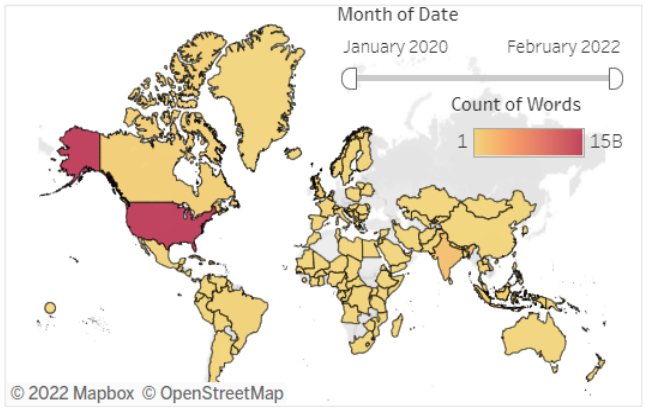

Fig. 2. Choropleth Map with interactive slider.

4.1.3 Streamline Graph

We have decided to implement the Streamline graphs for the 12 Covid-19 Related words. The data contains the commutative frequency of Covid-19 related word use on a daily basis. The chosen words for study are Covid, Corona, Death, Recovery, Relaxation, Social Distance, Vaccine, Mask, Lockdown, Omnicorn, and Quarantine

By analyzing the graphs and by looking at the whole commutative frequency of word usage we can identify the sudden increase in the coverage of Covid-19 related words as more cases being emerging worldwide. And other such Trends can be found. Also among the stations, we have chosen we can observe that ALJAZ has most of the coverage in all Covid-19 Related words.

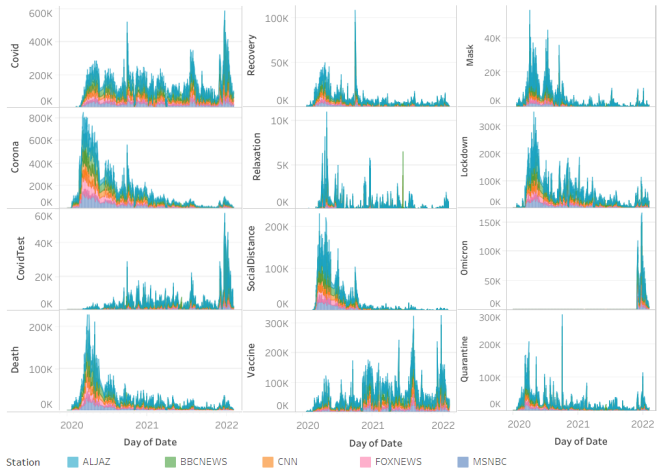

Fig. 3. Streamline Graph of 12 Covid-19 Related words.

4.1.4 Word Cloud

Word cloud is created on the basis of the frequency of the word that occurred on a news channel or snippet. The virus is the most used word during the times of covid followed by covid and corona. Words like Quarantine, Lockdown, and Vaccine were also used frequently. It is an interactive word cloud where a change in date may result in a change in the frequency of the word.

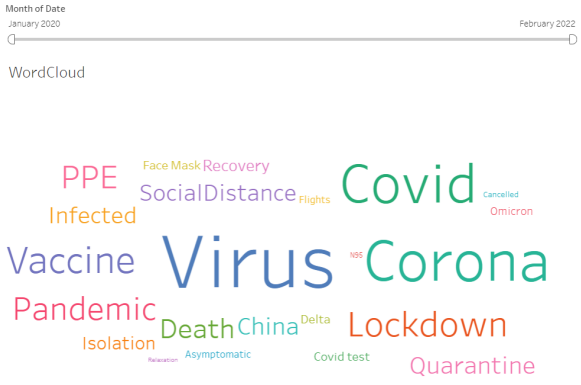

Fig. 4. Word Cloud with interactive slider.

4.1.5 Bar Graph

The bar graph below shows the number of snippets shown by news channels from December 2019 till now regarding covid, main purpose of this bar graph is to select the top 5 news stations on basis of the snippets shown by them related to covid during covid times.

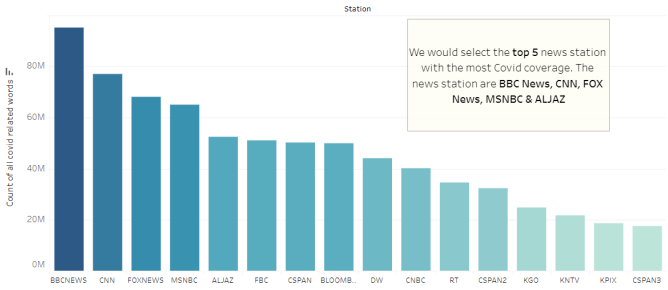

Fig. 5. Bar plot for TV station frequency analysis

5 Interpretation of Results

BBC News, CNN, FOX News, MSNBC, and AlJazeera were the top five news stations according to Covid coverage. The frequency of terms within a certain period can be determined via a word cloud analysis of Covid-related words. During the previous two years, the terms "Corona," "Covid," and "Virus" were more frequently used. After October 2020, the use of the word "vaccine" grew. The word frequency streamgraphs provide insight into the events that occurred throughout the Covid era, as well as their intensity. The impact of the second Covid wave of the Delta variant and the third Covid wave of the Omicron variant is clearly matched by the streamgraphs of the "Delta" and "Omicron" words. The streamgraph of the word "Vaccine" depicts vaccine discussions during the Covid period, as well as how they increased once vaccinations became available. According to the World Map research, Covid had the largest coverage in the United States and India, which correlates to the number of cases in both countries. By merging the geographical map with the frequency plot, the analysis also indicates the influence of Covid in countries such as the United States, India, and Australia, as well as the timing of the second and third waves in the same.

6 Challenges and Opportunities

There are several challenges we are facing during this project, but the most challenging part is the size of the data. Since data is too large it is difficult to use the data on the local machine, therefore, we have to use BigQuery or other cloud platforms for data analysis and visualization. Another, major challenge we are facing is data is available for all the languages and we are only doing visualization for English, therefore most of the data is of no use increasing time complexity. One more challenge is that this data is evolving data and is updated each day so it is very difficult to cope with it because of its size.

The major opportunity we got from this project is that we have open data for everything, so we can extract it in whatever way we want for analysis. Although the size of data is an issue, it comes with an opportunity.

7 Complexity and Scaling issues

There are thousands of words associated with covid but it is very difficult to create visualizations for all of them. That's why we had to reduce the scale to select the top 12 words used during covid. There are many synonyms for a single word, for example, covid, covid-19, SARS, and corona can be used to represent covid-19 so it increases the complexity while visualizing. Each row of data contains a text column with 50-100 words which increases the time complexity of the query to execute. This data is best accessible on Google BigQuery because of which we are facing some connection complexities on the tableau and other visualization software, restricting us to majorly dependent on Google Data Studio for visualization purposes.

8 Promising Avenues for Future Work

Each of the words selected as covid glossary can have synonyms and it is possible for words to have different meanings in different respects which is not considered in this project. In the future, some NLP techniques could be used to handle this problem in an effective manner.

Acknowledgments

The authors wish to thank Prof. Andreas Bueckle and Prof. Michael Ginda for their guidance and support throughout the project. And we want to thank our client Dr. Kalev Leetaru for his continuous support and feedback during the development of this project.

References

[1] "Coronavirus in the U.S.: Latest map and case count" New York Times (February 18, 2022). https://www.nytimes.com/interactive/2021/us/covid-cases.html

[2] Katy Börner: Atlas of science: visualizing what we know: The MIT Press, Cambridge, MA/London, UK, 2010, Scientometrics. 88. 675-677. 10.1007/s11192-011-0409-7.

[3] https://blog.gdeltproject.org/quantifying-the-covid-19-public-health-media-narrative-through-tv-radio-news-analysis/

[4] Ng R, Chow TYJ, Yang W (2021) News media narratives of Covid-19 across 20 countries: Early global convergence and later regional divergence. PLOS ONE 16(9): e0256358. https://doi.org/10.1371/journal.pone.0256358

[5] What Google's Cloud Video AI Sees Watching Decade Of Television News : https://blog.gdeltproject.org/what-googles-cloud-video-ai-sees-watching-decade-of-television-news-the-visual-global-entity-graph-2-0/

[6] Google Big Query : Database https://console.cloud.google.com/bigquery?project=gdelt-bq&p=gdelt-bq&d=gdeltv2&t=vgegv2_iatv&page=table

[7] Abdel-Latif, M. M. M. (2020). The enigma of Health Literacy and COVID-19 Pandemic. Public Health 185, 95–96. doi:10.1016/j.puhe.2020.06.030

[8] Aisami, R. S. (2015). Learning Styles and Visual Literacy for Learning and Performance. Procedia Soc. Behav. Sci. 176, 538–545.

[9] Dayan, Z. (2018). Visual Content: The Future of Storytelling. Forbes. doi:10.33107/ubt-ic.2018.360

[10] McFadden SM, Malik AA, Aguolu OG, Willebrand KS, Omer SB. Perceptions of the adult US population regarding the novel coronavirus outbreak. PLOS ONE. 2020;15: e0231808. pmid:32302370

Link to Visualization

https://public.tableau.com/app/profile/shardul.samdurkar/viz/CovidAnalysis_16506336277170/Story?publish=yes