Covid And Misinformation

Bilal Tariq

2024-05-07

Introduction

We all remember COVID-19 as the days we remained closeted up alone in our rooms with nothing to do but doomscroll through countless Youtube and Tiktok videos. Under such circumstances, we were consuming content at immaculate rates, but how many times was it that you stumbled across a piece of media that seemed dubious? Although it may have seemed dubious to us, it turns out that "over 78% of the public believes or is unsure about" at least one piece of COVID-19 misinformation (KFF, 2021). Moreover, the situation was only worse in developing countries with poorer educational structures, which made misinformation about COVID only that much more potent. Having come from a developing country myself (Pakistan), I was well aware of the crises at hand there: people using non-scientific ailments, conspiracies against governments of other countries and not following the protocols of quarantine.

For my Final Project, I intended on looking at COVID-19 spread and misinformation (in 2020). In particular, the highlight of my project was performing a topic analysis on the titles of mininformation articles that had been compiled into a dataset, i.e looking at what general 'topics' could be extracted from them. Not only did I perform a topic analysis of misinformation across the globe, but the more important idea was looking at how what misinformation topics were spread across certain countries too. I wanted to see if there was any association between the cultures, geographies, number of COVID cases and development of the country with the topics of misinformation that were being spread across those regions. To aid my analysis with respect to the spread of COVID cases, I also developed a spatial visualization of the cases on a world map.

Data

This project entailed having two primary datasets. The specific wrangling code for both can be found in "Data Wrangling - Final Project .Rmd". The links to both the primary datasets may be found in the references.

The first, by the European Center for Disease Prevention and Control, included the number of daily cases per country in 2020. This was collected during the COVID-19 epidemic in order to help note the testing efficacy of countries around the world and improve it. For my specific case, I needed this dataset in order to display a visualisation of the COVID spread in 2020 using a map. So we had to condense the daily cases per country into a number for the yearly cases per country. The top 5 rows of the intermediate dataset looked like this:

```r
# Importing dataset of covid cases
casesCountryData <- read_csv("COVID Datasets/dailyCases.csv")
casesCountryDataSelected <- casesCountryData %>%
    select(cases, deaths, countriesAndTerritories)

# Since the cases dataset is of daily cases per country in the year 2020 we are
# going to have to summarise the data and group by country

casesCountryDataSummarized <- casesCountryDataSelected %>%
    group_by(countriesAndTerritories) %>%
    summarise(Cases = sum(cases))

# Only displayed the first few rows
top_5 <- head(casesCountryDataSummarized, n = 5)

kable(top_5)
```

countriesAndTerritories	Cases
Afghanistan	49273
Albania	48530
Algeria	92102
Andorra	7338
Angola	16188

Furthermore, I had to manually change some names of countries (which were inconsistently named and had underscores) so that I could merge it with a dataset of latitudes and longitudes. Merging it gave it spatial coordinates, which allowed it to be used to plot in leaflet. The first wrangled dataset is shown below.

Countries	Cases	Latitude	Longitude
Afghanistan	49273	33.0	65.0
Albania	48530	41.0	20.0
Algeria	92102	28.0	3.0
Andorra	7338	42.5	1.5
Angola	16188	-12.5	18.5

With the cases dataset ready, it was time to move onto the misinformation dataset. This data was provided by UNESCO and compiled by the Empirical Studies of Conflict (Princeton University), so it can be considered a trustworthy source. It was collected from the early days of he pandemic through the end of 2020 as well with over 5,600 misinformation stories. Although trustworthy, we must note that what is deemed as "misinformation", ultimately fell into the hands of the organization at Princeton that compiled the data, so it should be acknowledged there is an aspect of subjectivity as to what inherently is misinformation, with some topics being slightly more vague than others.

There was some wrangling required for this part to simply filter out any words that were non-english (as the topic analysis we were performing may only be performed on English words). This does create a marginal issue with our analysis where we may not be able to fully get a grasp over the misinformation on a country as we are just limiting our analysis to English topics. However, for our purposes, it is a very minimal effect. Below is a sample of what the Title of an Article looks like and which country it is from.

Title	Primary_Country
India Is In The Middle Of A Coronavirus YouTube Frenzy, And It's Going To Get People Killed	India

This now meant that we were ready to begin our journey on visualizing and investigating the spread of COVID and its misinformation throughout the world.

Methods

Shiny Map Visualization

In a similar vein to my data structuring and wrangling, my methods were divided into two parts. The first was the visualization of COVID spread using a Shiny Application, and the second was misinformation topic analysis using Latent Dirichlet Allocation.

The first method was the Shiny Application. The goal was to create a map using Leaflet that would show the general outbreak on a larger scale. The initial idea was to make a a map that was colorized on the basis of the outbreak; however, implementing that in Shiny proved to be too difficult an implementation. Rather, markers were used to aid the visualization.

The map has circular markers on it that represent the number of COVID cases in a given country. Those markers were meant to be colorized according to the number of cases as well, with blue representing a fewer number of cases and red representing a large number of cases. Furthermore, the map was built to make the markers clickable; they display the country name and the number of cases in that country. There is also a drop down on the sidebar of the shiny app, where the user may select a country they want to see the cases for and a view box that displays the number of COVID cases in that country.

Below are images of what the general world map looks like *(Figure 1)* and the interactability of the shiny map *(Figure 2)*. The implementation of Leaflet and its interactibility was also a part of my "Go Beyond" aspect, where I hadn't implemented spatial visualizations using leaflet in class before.

LDA Topic Analysis

The core part of the project lies here, in the LDA Topic Analysis. The way LDA topic analysis works is by tokenizing words and using a mathemtical network analysis (Grün and Hornik, 2011) extract topics of terms which go together within it.

To execute this, we first had to tokenize our wrangled dataset, by unnesting tokens from the tidytext package. This allows for us to break up the titles into individual words for each title. Next, we had to edit this set of unnested words by removing the stop words and also removing words like "covid", "19" and "coronavirus". This is because we don't want words that don't convey any information relevant to misinformation. The stop words have a similar situation, where they are common words and don't add much value to the topic analysis.

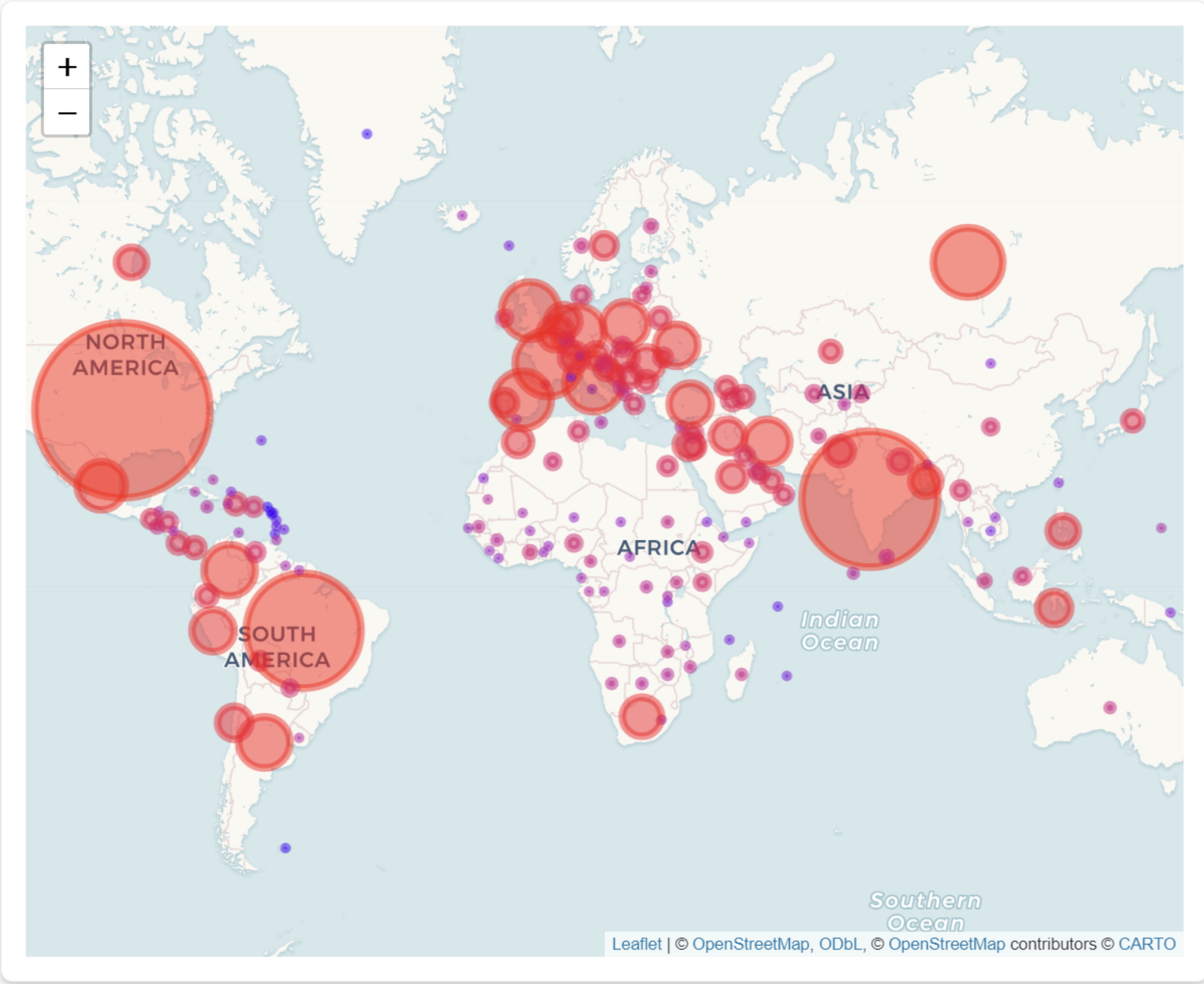


Figure 1: Shiny World Map

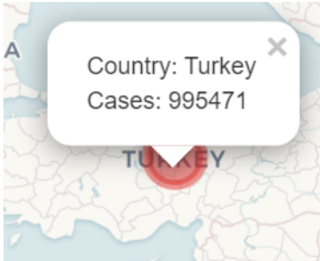


Figure 2: Map Marker

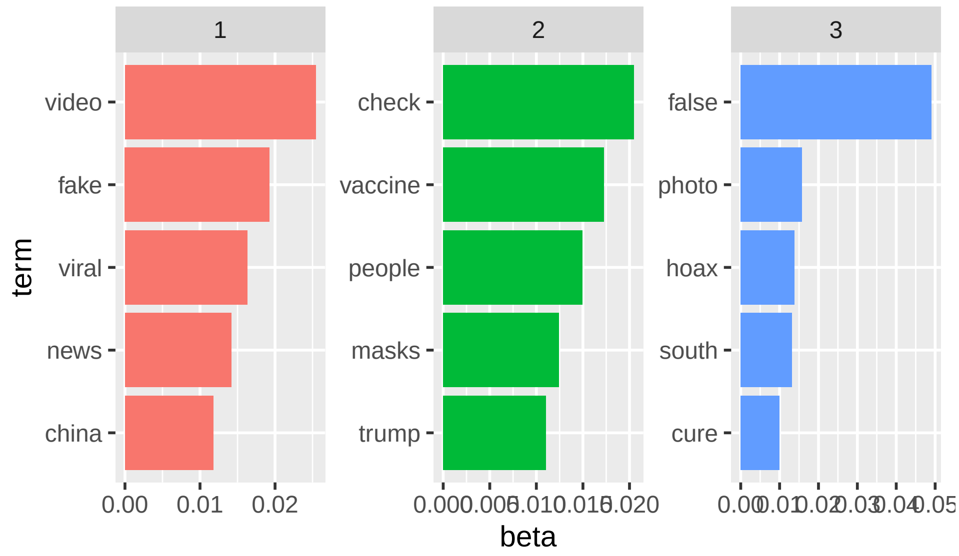
However, there was one more step to the pre-processing required for the Topic Analysis. This was to convert the data frame into a document-term matrix (DTM). A DTM represents the frequency of words in a mathematical matrix. DTM's are generally quintessential to any form of language processing, and especially useful in our case where to perform the LDA analysis using the topicmodels package we had to convert it to a DTM format (Grün and Hornik, 2011).

```
#Loading our wrangled dataset from Saved RDS Files

LDAAnalysisDataset <- readRDS("Saved RDS Files/LDATopicAnalysisData.rds")

#Tokenizing the text

tokenizedLDADataset <- LDAAnalysisDataset %>%
  unnest_tokens(output = word, input = Title) %>%
  select(Primary_Country, word)

# Filter out words like "COVID" and "19" and "coronavirus"
filterWords <- c("covid", "19", "coronavirus")

tokenizedLDADataset_filtered <- tokenizedLDADataset %>%
  filter(!word %in% filterWords) %>%
  select(Primary_Country, word)

#Getting rid of the stop words

data(stop_words)

tokenizedLDADataset_stopped <- tokenizedLDADataset_filtered %>%
  anti_join(stop_words, by = "word")

#Now we're trying to convert tokenized data to a Document-Term Matrix

dtm <- tokenizedLDADataset_stopped %>%
  count(Primary_Country, word) %>%
  cast_dtm(Primary_Country, word, n)
```

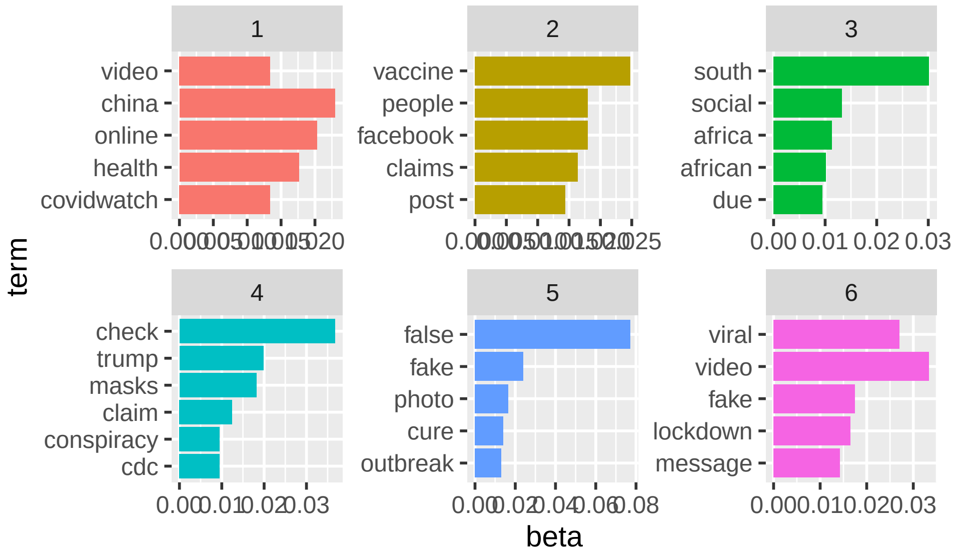
Finally, we can run the LDA text analysis using the topicmodels package as shown in the code chunk below. Notice how we run it twice. This is because the LDA function takes the number of topics as its second parameter. In my LDA Analysis file, I ran the code chunk multiple times to experiment with multiple different topics. I will discuss in my results further how I decided the number of topics to decide for my LDA analyses, which is primarily based off coherence of the terms with each other.

```
#The LDA function is the heart of our analysis

analysisLDA <- LDA(dtm, 3, method = "Gibbs", control = list(seed = 123))

topics <- tidy(analysisLDA, matrix = "beta")
```

```r
#Using GGPlot to visualise the data through means of a column bar chart

topics %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```



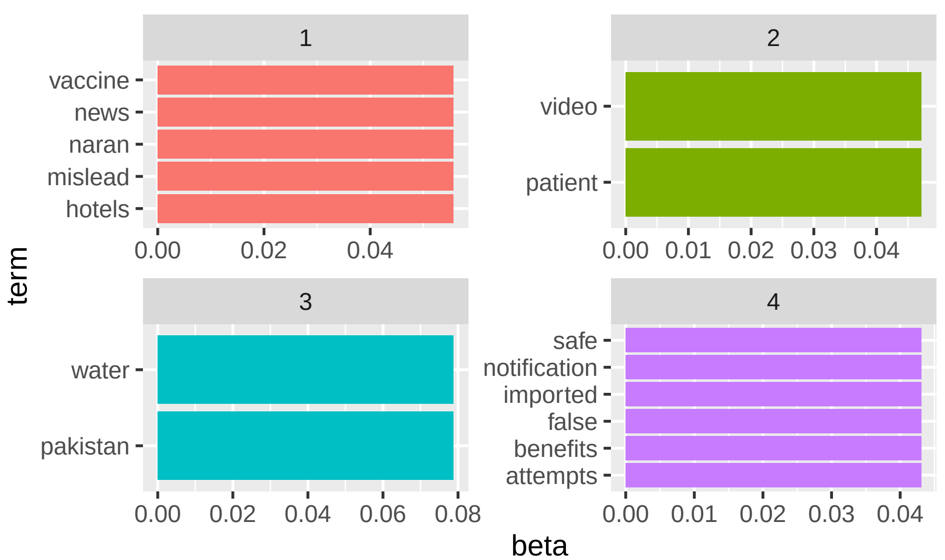

```r
analysisLDA <- LDA(dtm, 6, method = "Gibbs", control = list(seed = 123))

topics <- tidy(analysisLDA, matrix = "beta")

topics %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```



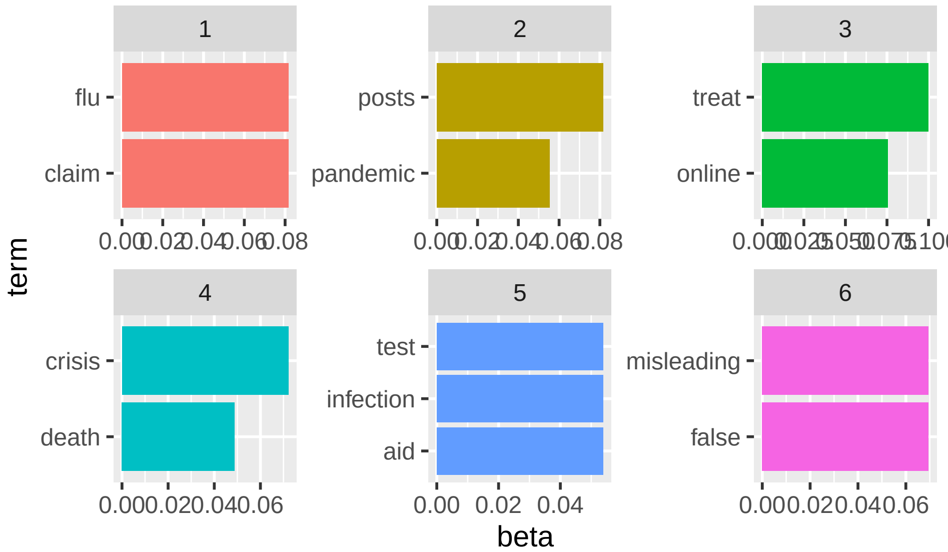
In this, the topics are labelled as “1”, “2”. . . , and they will be named later in the results and findings page. For each topic, terms are given based on what the model thinks the coherence is for each of these.

Next, I wanted to perform LDA Analyses on individual countries. Since I had already wrangled my dataset to include the countries from which each term was from, this was done by simply filtering the data with respect to each country I wanted to analyse. In particular, I looked at countries with a high number of COVID cases, which is why I explored Pakistan, India and the United States.

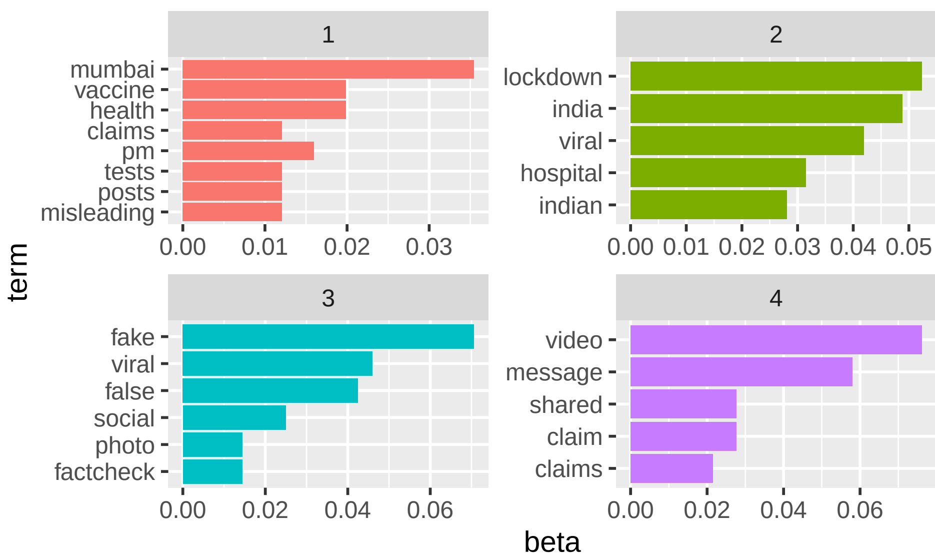
My first analysis was filtering the misinformation topics that were in Pakistan. I repeated the same process of exploring various topics between 2 and 8 like I did above. However, those visualizations are omitted for redundancy.



A similar process was followed for the United States:



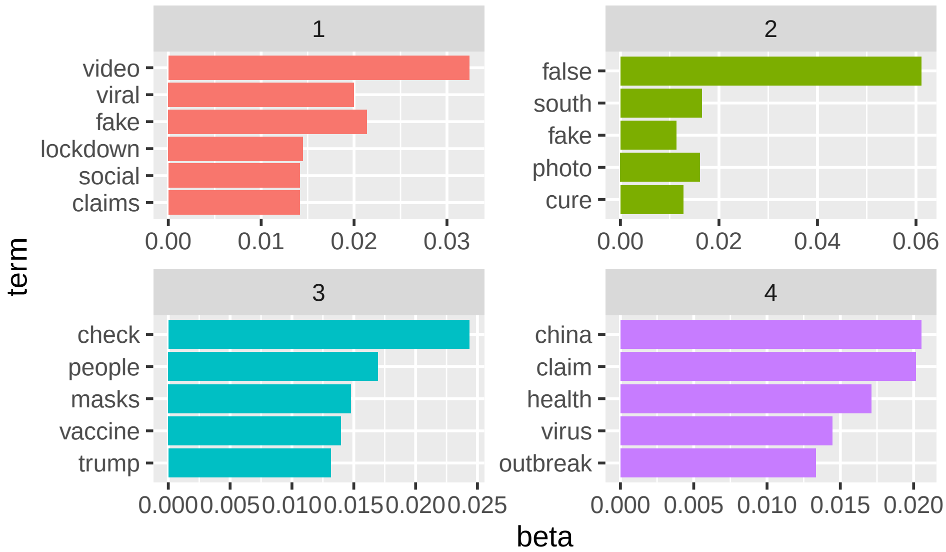
Finally, I did this process for India:



This summarizes the methods I had taken to perform my LDA Topic Analyses. To reiterate, each model for the countries that were investigated made sure that their topics were experimented with so that they are cohesive. Furthermore, as we will discuss in our results and findings, there are some topics that don't fit well with each other.

Results and Findings

So what does all of this mean? After looking at a visualization of the spread of COVID throughout the world and the misinformation that has been spread, how do we interpret these results? To reiterate, we were looking for misinformation to have an association with the cultures, geographies, number of COVID cases and development of the country. So let's interpret these results by first looking at the LDA Topic Analysis of the world.

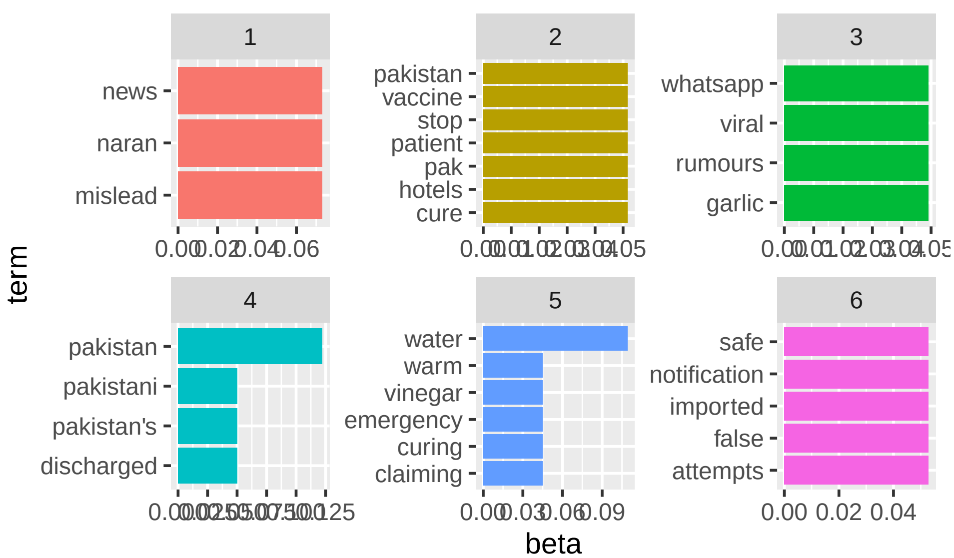


What I noticed when performing and LDA Analysis on the global misinformation was that there weren't many definitive "topics". This may be because there was just an overall use of words like "viral" and "false" etc. However, despite there not being an exact topic distinction, I settled on these 4 topics because they are the most distinct. The topics are described as follows:

- Topic 1 describes an overall topic of Social Media; videos going viral, fake news online.
- Topic 2 is similar to topic 1 with an overarching idea of false information.
- Topic 3 describes testing and vaccinations, and people's reaction to them.
- Topic 4 describes where the outbreak of the virus started and its health effects.

Topic Analysis of the world was only so revealing, so let's delve into individual country topic analysis.

Pakistan Misinformation LDA Analysis

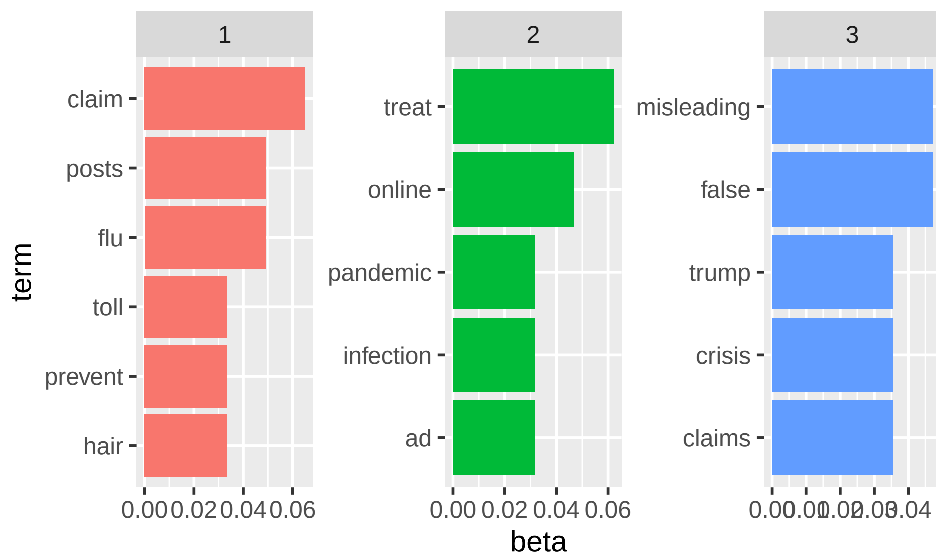


Being an international student in the US from Pakistan, this particularly stood out to me. Many of the topics here have a similar overarching idea as the world topic analysis we evaluated initially, but there are some distinct topics that show the misinformation in the nation.

- Topic 2 and 4 show a topic of COVID from a "Pakistani" perspective; misinformation relevant to people being discharged earlier than they should have etc.
- Topic 3 shows how misinformation spread in Pakistan. This was prominently done via Whatsapp and rumours, which are a big part of Pakistani Culture.
- Topic 5 shows Natural Ailments. It's interesting to see how in a developing country where the culture revolves around naturally healing, there was misinformation pertaining to healing methods which was curated towards Pakistan.
- Topic 1 refers to Naran, which is a local tourist destination. In 2020, there was a lot of misinformation around the COVID spread there and people still went to that area.

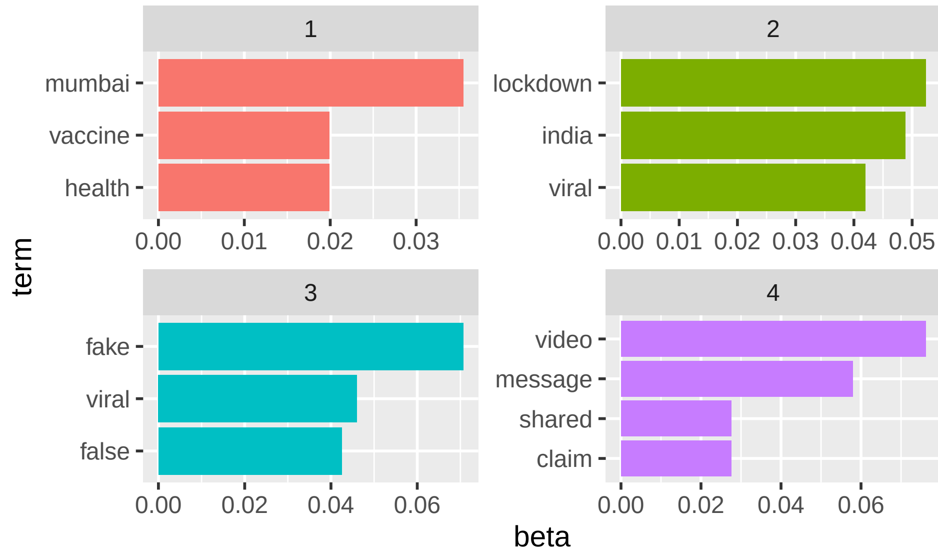
Overall, we can see that many of the misinformation topics revolved around culturally specific factors: the use of Whatsapp, natural ailments and local tourist destinations. There seems to be a high association with the topics of misinformation with the culturally and geographically specific factors that were mentioned before.

United States Misinformation LDA Analysis



When I looked at the topic analysis for the United States, it appeared to be very similar to that of the topic analysis of the world, so it was pretty inconclusive if it showed any indication of geography / culture affecting the analysis. One notable exception was topic 3, which seemed to be a more Political topic which catered to the U.S, especially considering how Donald Trump was president at the time and his handling of the COVID crisis in the United States. However, what's important to note is that United States culture and sentiments are generally resonated by other countries, which is why the United States doesn't necessarily have a distinct set of misinformation topics, while other countries like Pakistan and India (discussed below) do.

India Misinformation LDA Analysis



Analyzing the results for the topics in India also gave a few key insights:

- Topics 1 and 2 revealed more about the crisis in India. The first resonates around Mumbai, where there was a crisis with distributing vaccines and the health of the elderly. Furthermore, Topic 2 talked more about the migrant worker crisis, where many Indian migrant workers had to deal with a decrease in wages due to lockdown.
- We can also see a general trend where the topic analyses we found across the world were also found in these countries. Topic 3 and 4 represent the topic of viral fake news and the way misinformation was spread through videos, messages, etc.

Conclusion and Reflection

Our initial goal was to address whether misinformation topics across the globe and across countries were affected by cultural, geographic, socioeconomic and political factors and how the misinformation was associated with the spread of the coronavirus. Through our LDA topic analysis, we found that there was indeed an association of local events with misinformation that had been spread across countries, despite there also being repeated topics of misinformation that had been mutually shared between various countries. Our study did reveal a general understanding of misinformation throughout the globe, however the topics did overlap and were ambiguous. Our map visualization also helped illuminate the intensity of the spread of the virus, and it showed that there was an association of topics with 'outbreak' and high density cities like 'Mumbai' in areas where there was a higher risk of catching the virus.

Limitations

The LDA topic analysis proved to be useful in some circumstances but unreliable in others. Since the LDA analysis functions by grouping words together, it is very much possible that there were limitations in understanding the context of some words. If we were to perform an analysis on misinformation in the future, it would be better to use a text analysis that incorporated sentence context, which would then have made the types of misinformation much clearer. Furthermore, we were limited in the sense that our data only comprised of titles for each article. This made it harder to ascertain the frequencies of misinformation.

To improve, we could have scraped the individual documents from the dataset with their provided URL's and performed an LDA text analysis on those documents. Finally, the biggest limitation in our study was that the dataset was picked by individuals, and inherently can't be an objective metric. Since some forms of misinformation may not necessarily be as clear, it's difficult to determine what does and does not classify as misinformation, and this may have skewed our results in our topic analysis.

References

(Click on the text to be taken to the respective website)

1. Craig Polosky, KFF, 2021
2. European Center for Disease Prevention and Control, 2020 (Dataset for Daily Cases)
3. UNESCO, 2020 (Dataset for COVID misinformation)
4. Grün and Hornik, 2011 (Topicmodels R Package)