Team Signature Assignment

Northeastern University

College of Professional Studies

COVID-19: epidemiology, hidden factors, and communication policy

COURSE: ITC6040 81288 Informatics Capstone

INSTRUCTOR: Dr. Xiaomu Zhou

INVESTIGATOR: Mikhail Oet

GROUP-INFO: Team 2

Karishma Pathan

Shivani Raina

DATE: 26th June 2020

Team Signature Assignment

1) Project requirement, methods, and goals

The project COVID-19: epidemiology, hidden factors, and communication policy aims to focus on the analysis of information from the COVID-19 pandemic to facilitate the deployment of effective public safety policy. This project is sponsored by FNA which is a deep technology analytics company and also supervises complex financial networks. Currently, the FNA's goal is to acquire &depict the COVID-19 hidden factors which has affected the multiple countries and also analyze and compare the sentiment of government channel and other public sentiment channels describing emotions about COVID-19. In this project, we are working closely with Mikhail Oet, a principal investigator of our project. The requirements were to develop the SmartPLS model which will have standardized Twitter Sentiment result, The GDELT project sentiment result, RavenPack data, Mobility and economic uncertainties as input data. Initially, we had the storage issue as GDELT data is large in size, but later we migrated to the Northeastern Google Drive which has lot of space.

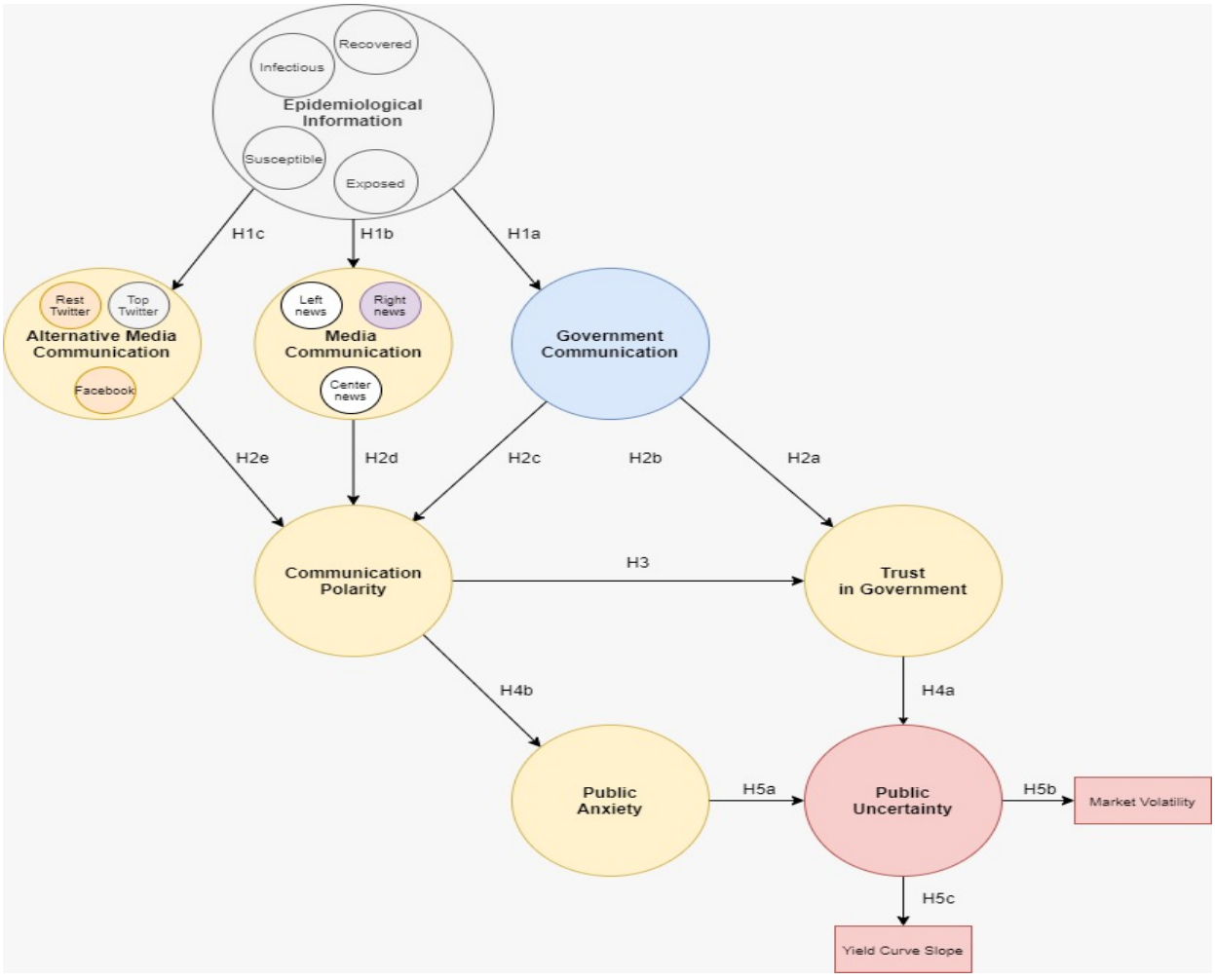


Figure 1

We have used sentiment analysis technique for opinion and text mining of sentiments of people regarding COVID-19 pandemic. We have focused on four developing countries such as India, Iran, Brazil and South Africa for this analysis. In public opinion, we have looked the answer of what people think about the COVID-19 pandemic through media (GDELT) and alternative media communication (Twitter, RavenPack). We have created bag of words for positive and negative words to measure the panic/anxiety in public. We have translated every word in around 15 different languages to get the maximum out of the data as we worked on four diverse countries with multiple spoken langueges.

Data Sources and Data Collection:

We have used the web scrapping technique to exctract the data from the GDELT, Twitter and RavenPack and the data of epidemiology, policy uncertinity etc. we have downloaded from the respective websites. The GDELT project is our main media communication resource as it has the largest database with data in 100 differet languages. However, Twitter is alternative media communication resource and we exctracted the data using twitter IDs and saved it in line oriented json format. On the other hand, we have used web scrapping graph method to scrape the data from RavenPack website. Below are some of the libraries/packages we have used to extract the data:

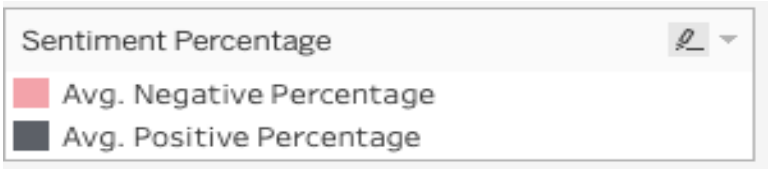
- lxml: Used to do the web scrapping
- gdelt: Used to access and analyze GDELT global database
- html2text: Used to create html files from some text file
- cleantext: Used to clean raw data
- tqdm: Used to create the progress bars and estimate the time to completion for the function
- twarc: Used to archive twitter json data and it handles the twitter APT rate limit.
- json: It parses the JSON data into python dictionary or list.
- urllib.request: It fetches the URLs using different protocols and offers urlopen function.

Team Signature Assignment

2)Project Outcome

To identify the sentiment of people in four countries which are India, Iran, South Africa and Brazil, we took data from three different resources RavenPack, The GDELT Project, Twitter. The Twitter dataset has the tweets of the people, The GDELT has the different articles and the RavenPack has the different reactions, like Media Hype, Panic, Infodemic. The other factors which is too be considered are the mobility which includes the parks, recreation center, workplace, grocery store.

Below is the time Series graph of the Twitter, which depicts the count of positive and negative sentiment across four countries from Jan 2020 – June 2020



Twitter_Brazil:

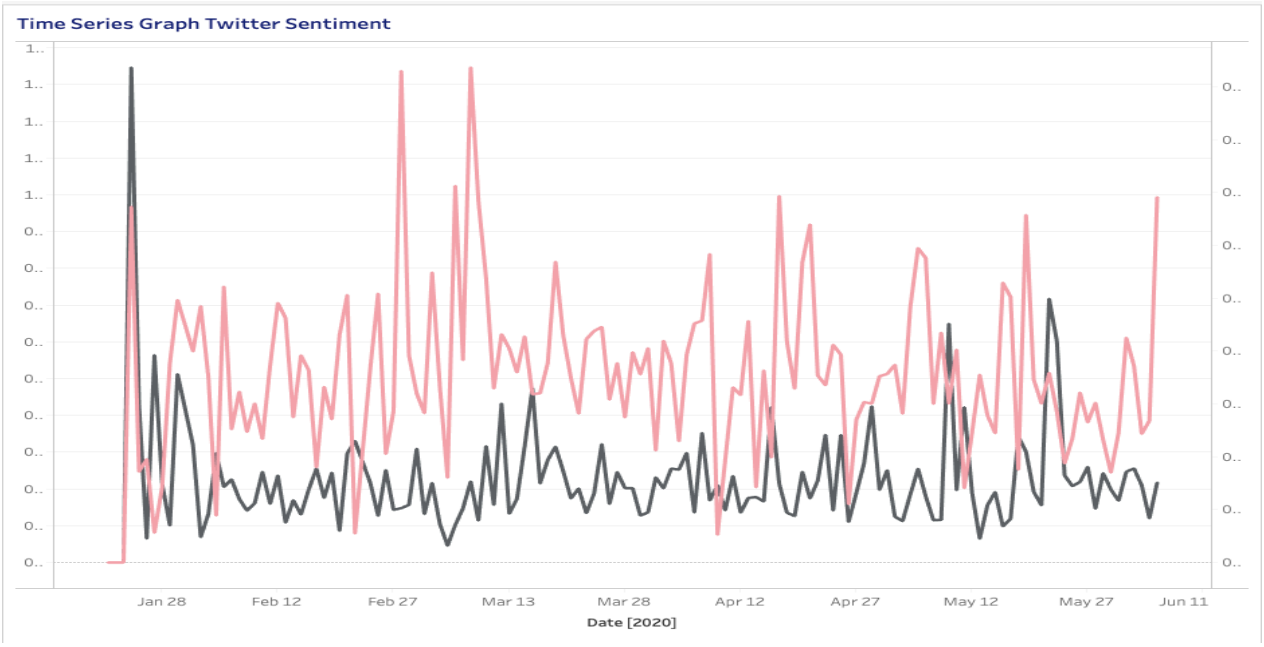


Figure 2

In the above graph (Fig. 2) of Twitter_Brazil, we can notice that the graph of negative sentiment is mostly higher than the positive sentiment, mostly during the month of Feb, however, in the month of May, the positive sentiment is increasing

Twitter_India:

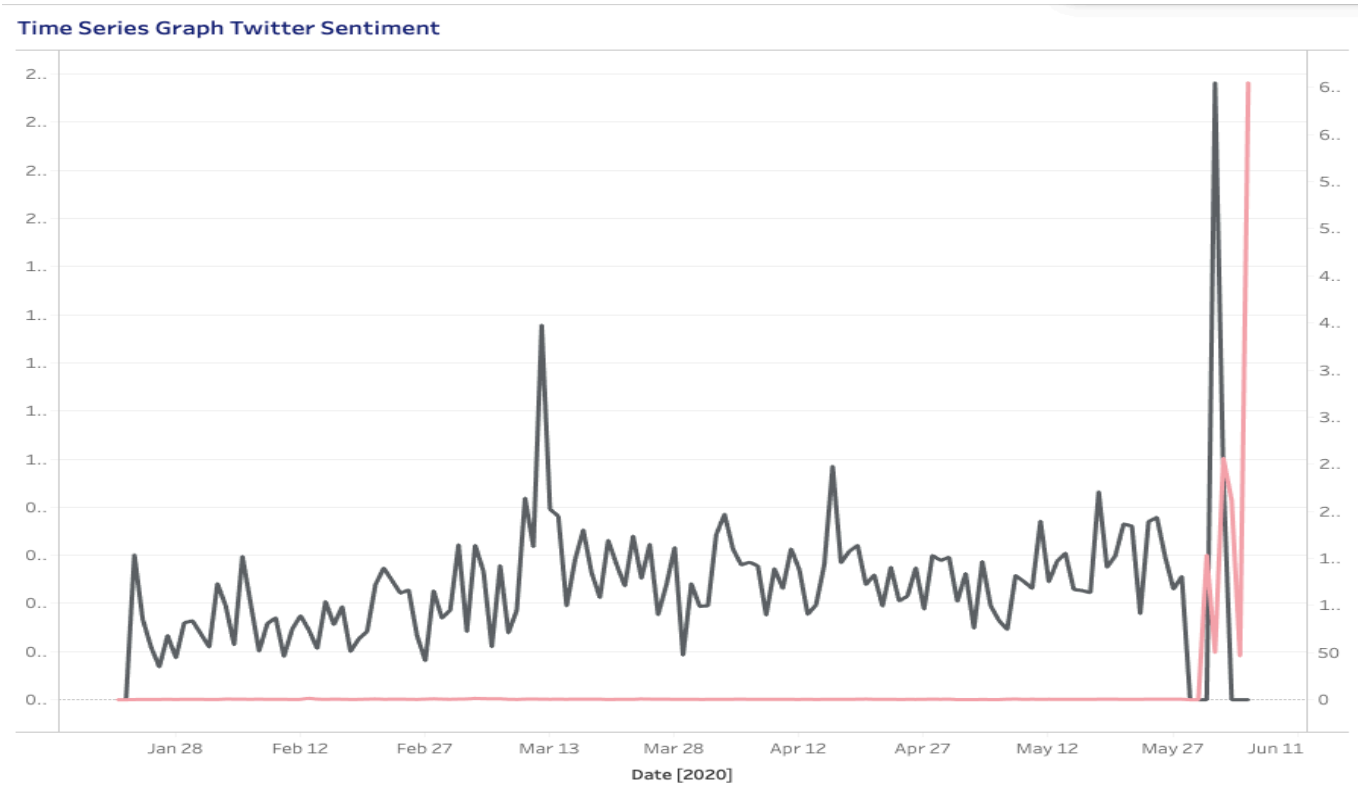


Figure 3

Similarly, while comparing the Twitter sentiment for India, we noticed that positive sentiment is shown more among the people, however, for the month of May, negative sentiment is increasing, may be the situation is not coming under control.

Team Signature Assignment

Twitter_Iran: Fig.4

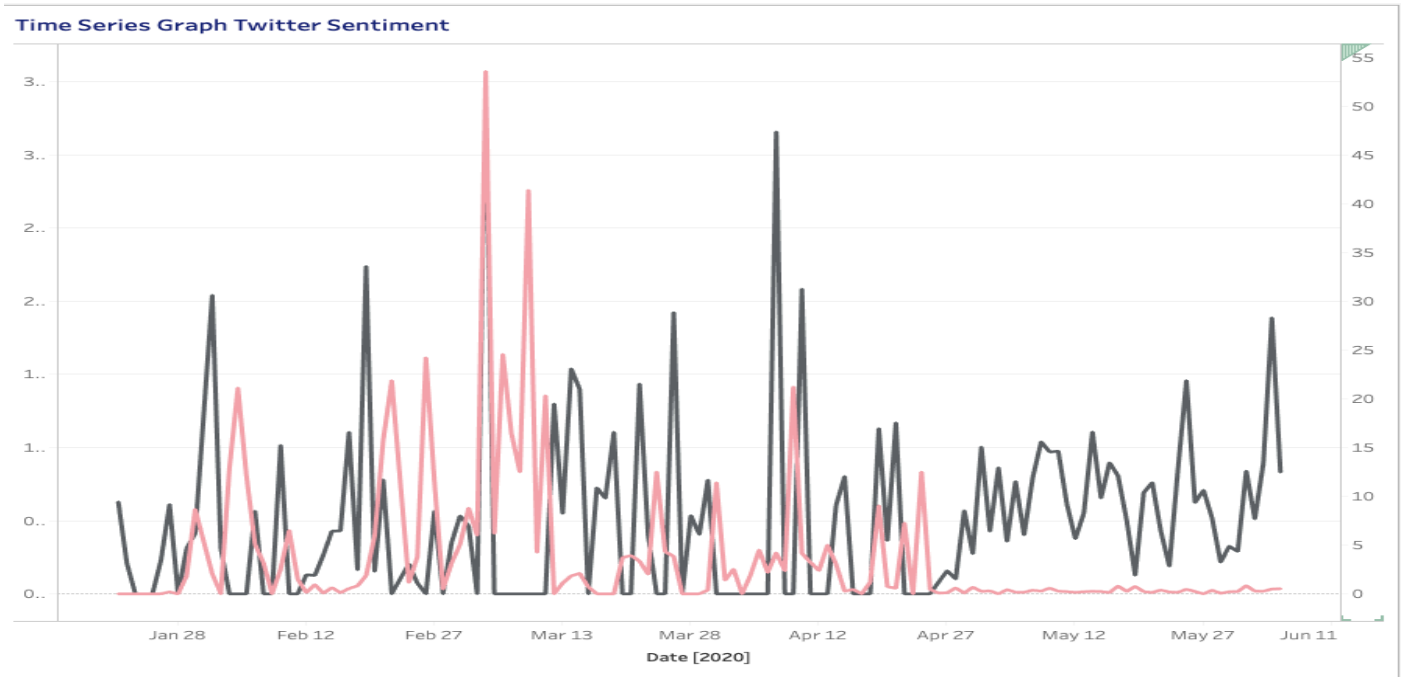


Figure 4

After, looking at the above two graphs (Fig.2, Fig. 3) it was certainly a curiosity to know the trend of Iran, coming up as a surprise, the Iran graph has the mostly the negative sentiment, throughout Jan to June

Twitter_South Africa:

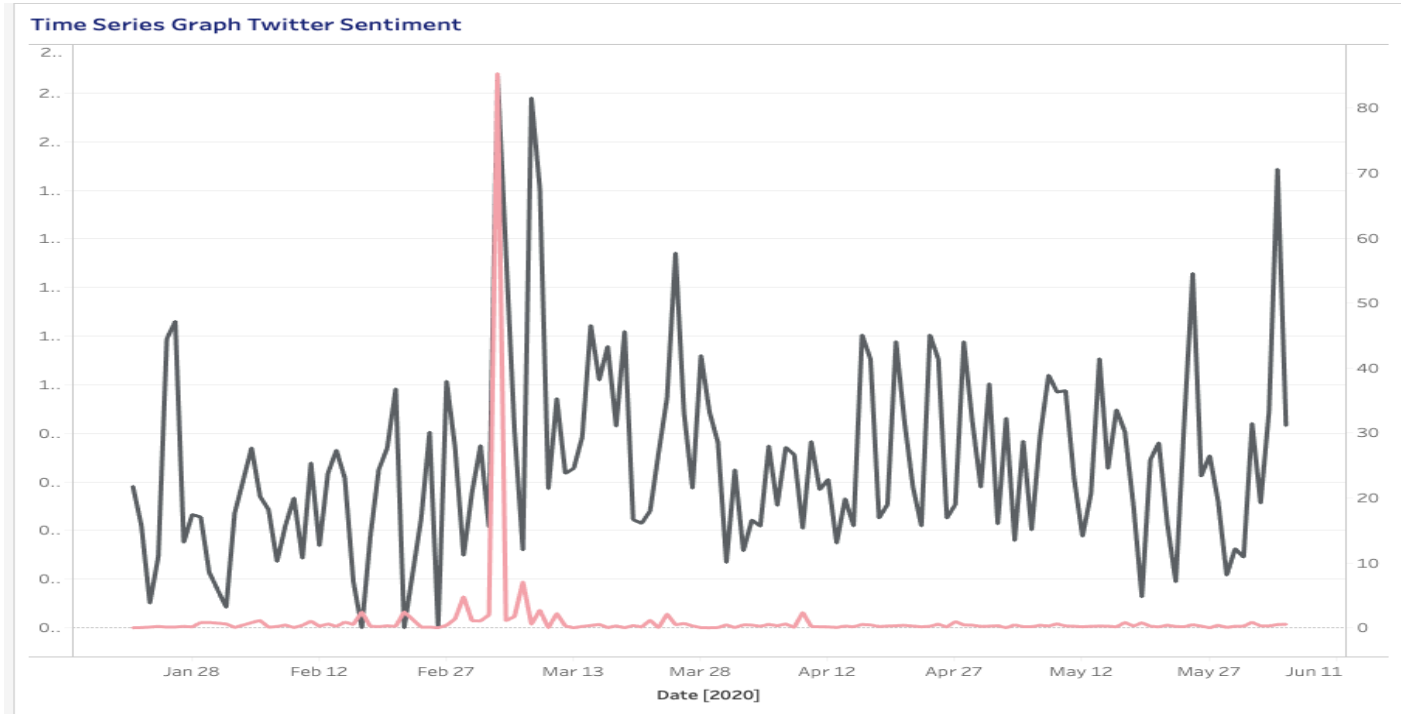
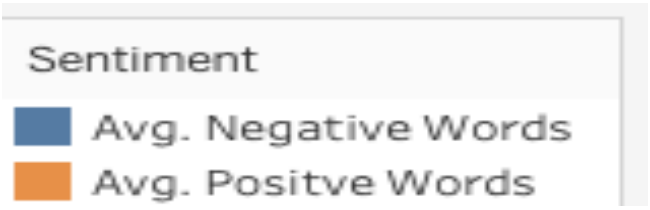


Figure 5

And also, the South Africa graph also took our special attention as it has drastic difference between the positive and negative sentiment. To have more information on the sentiments across these countries, we did the sentiment analysis on GDELT.



GDELT_Brazil

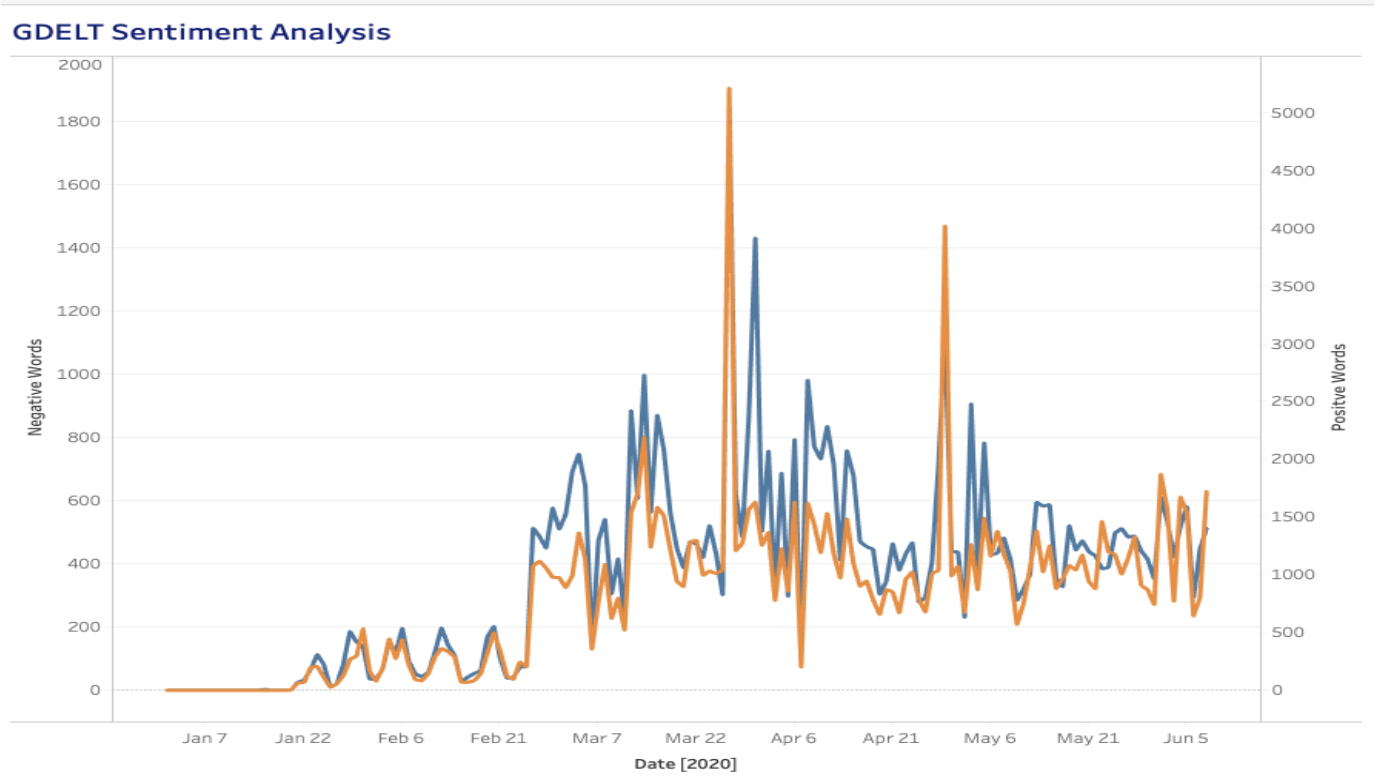


Figure 6

Team Signature Assignment

GDELT_India

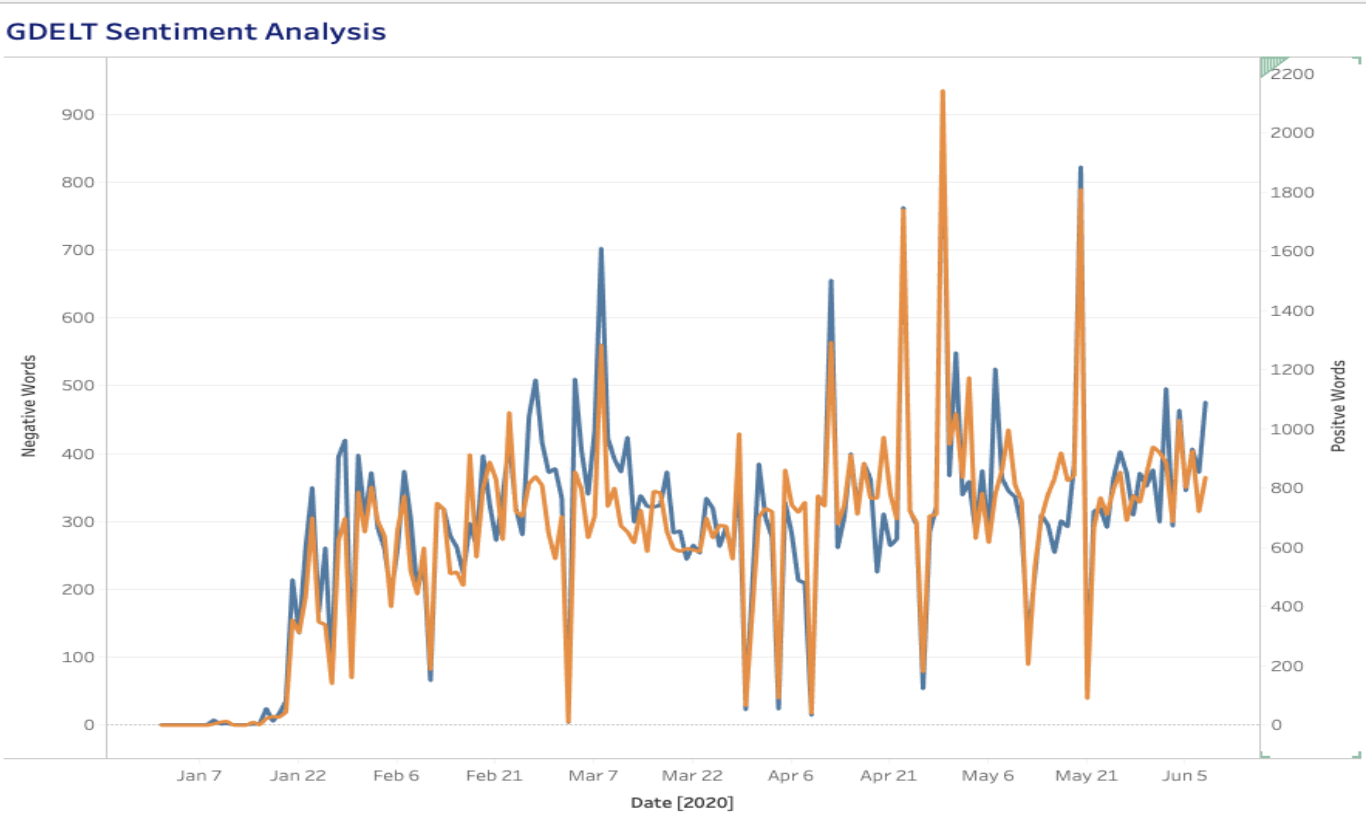


Figure 7

Team Signature Assignment

Iran_GDELT

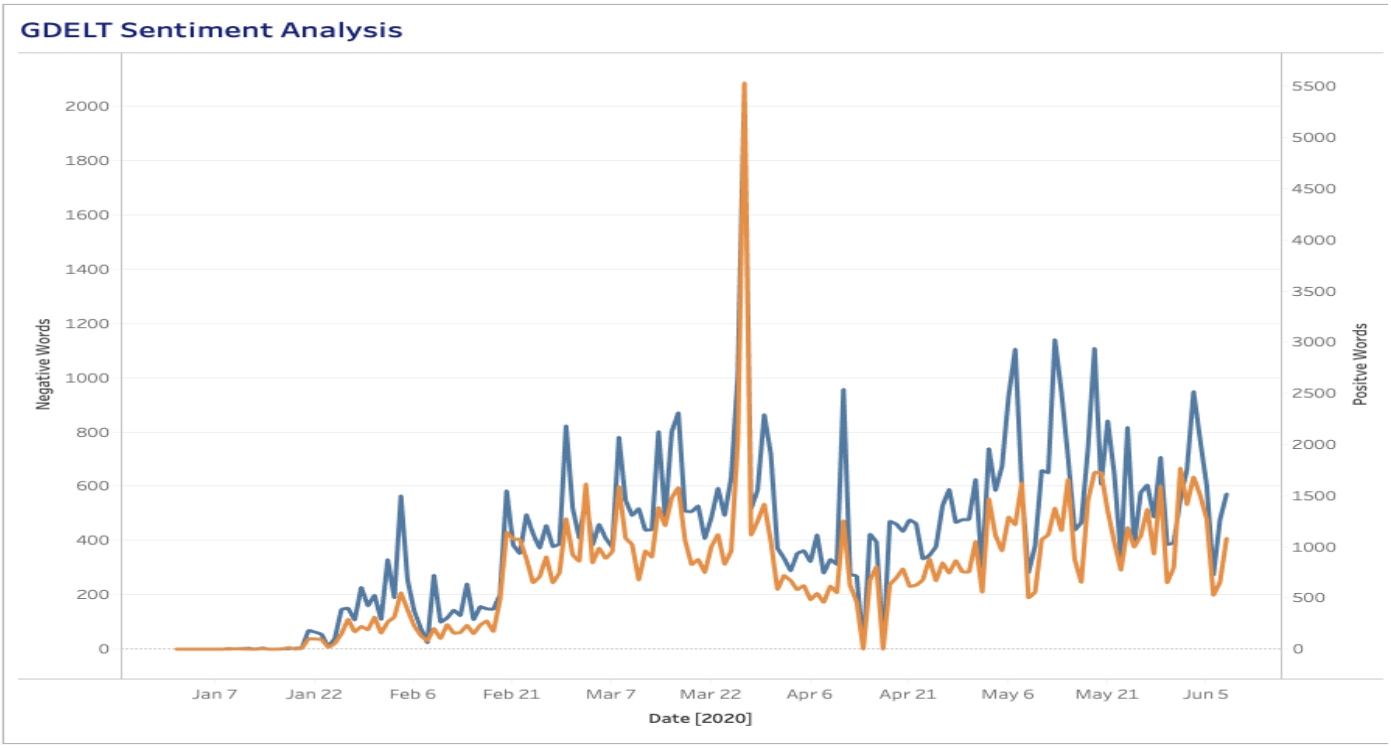


Figure 8

Team Signature Assignment

South_Africa_GDELT

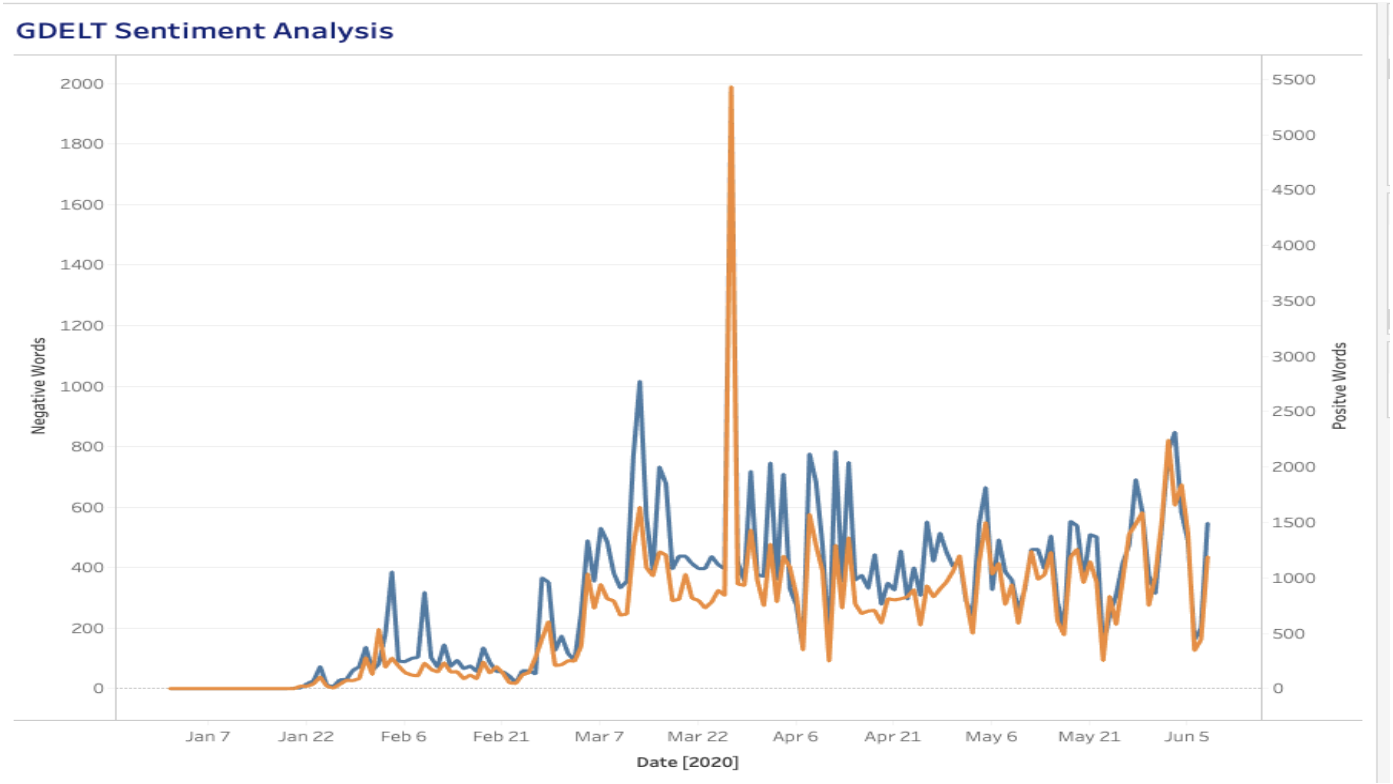


Figure 9

While also, comparing the Twitter results with GDELT for four countries, here also South Africa has the most positive sentiment, however, there Is increase of negative sentiment as well comparatively. Also, in GDELT analysis, Iran has positive sentiment at highest peak on March 27, where in Twitter (Fig. 4) it is shown April 12, a bit strange.

To dig deeper, we also did analysis of the RavenPack data taking three main factors, Infodemic, Panic, Hype.

Team Signature Assignment

RavenPack_Hype

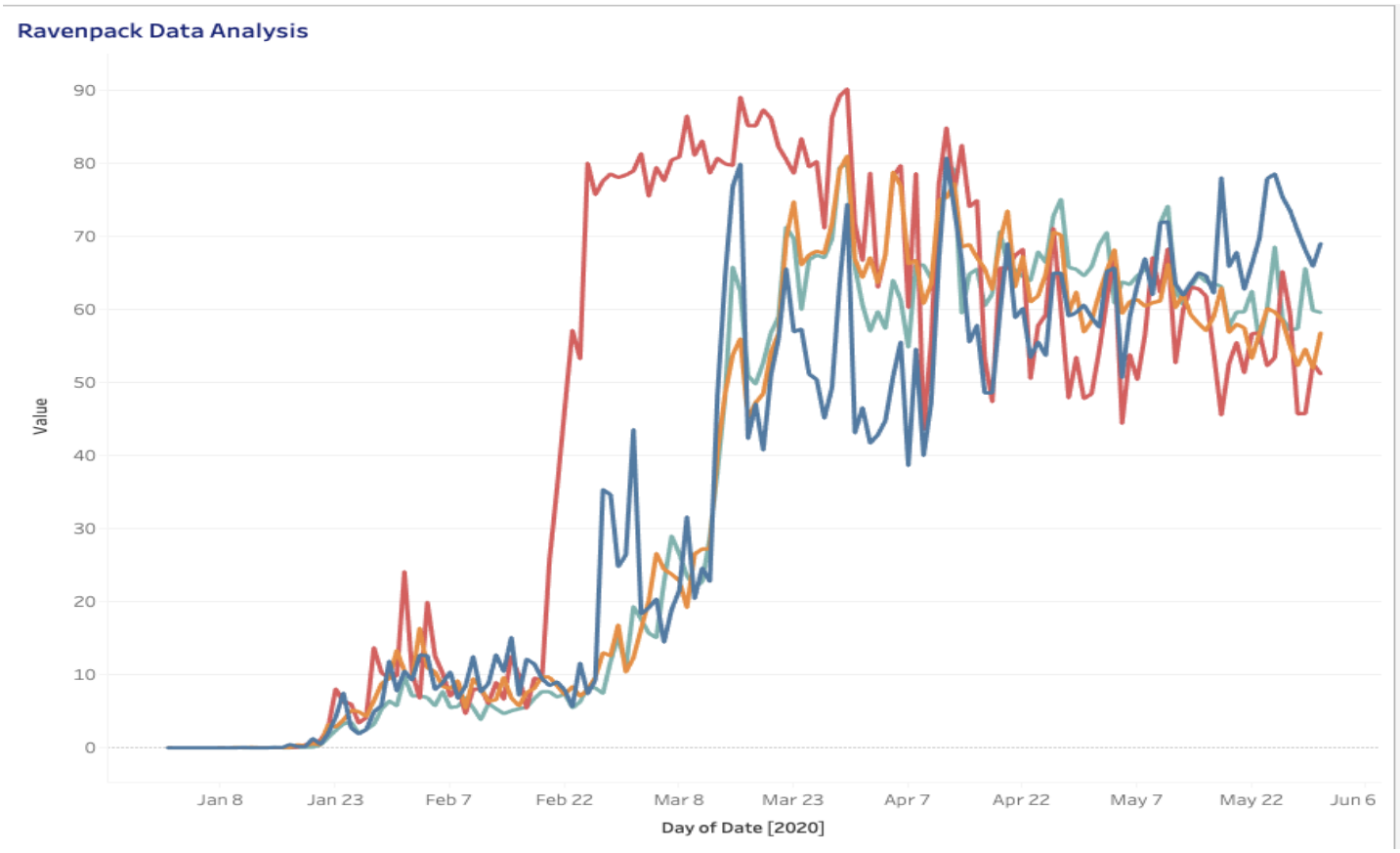


Figure 10

The graph clearly depicts the highest media hype in Iran from the month Jan- April and least media hype is shown in South Africa.

Team Signature Assignment

RavenPack_Infodemic

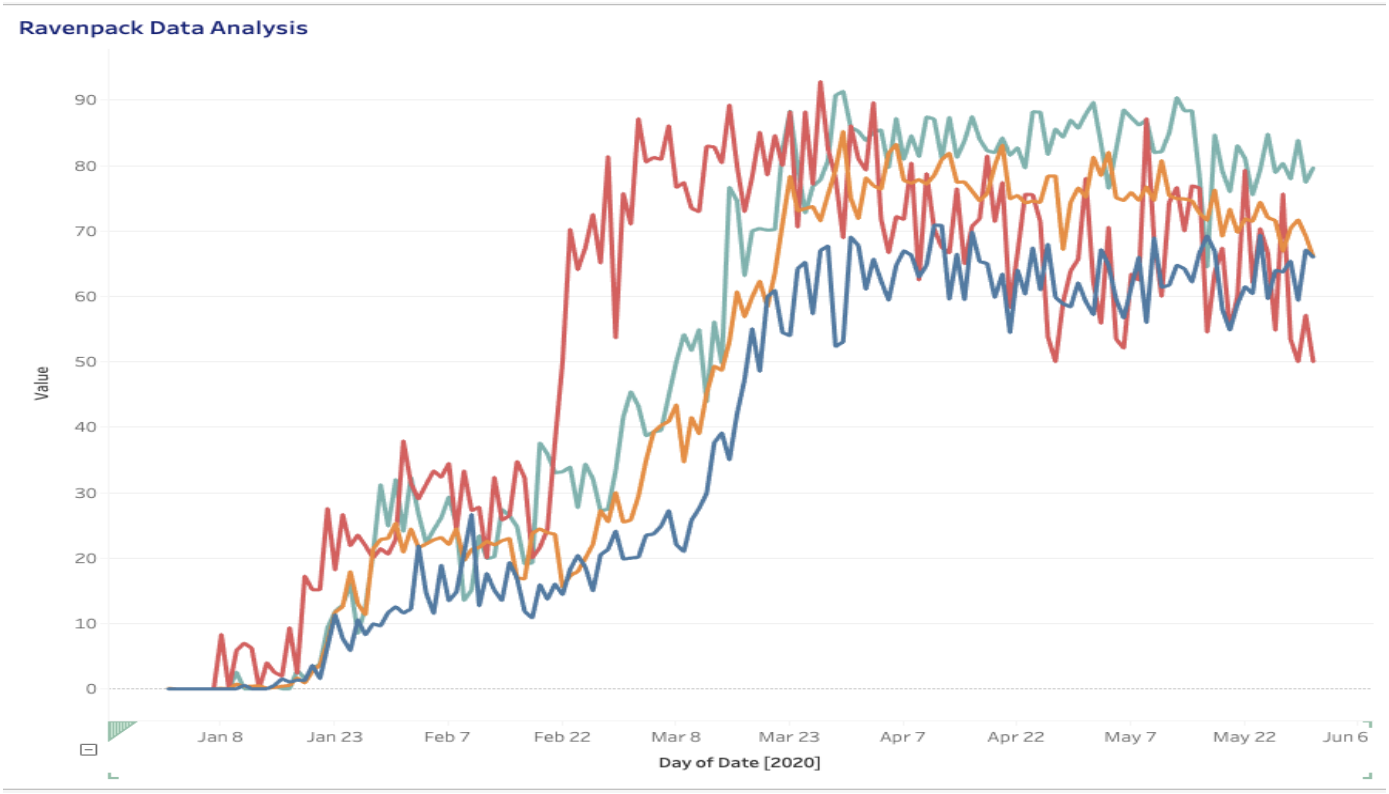


Figure 11

The infodemic means information being spread in pandemic, here also, we see an elevation in Iran and then decreasing after April. The information spread by media created most of the hype making it too often spread the misinformation which was mostly seen in Iran and India.

Team Signature Assignment

RavenPack_Panic

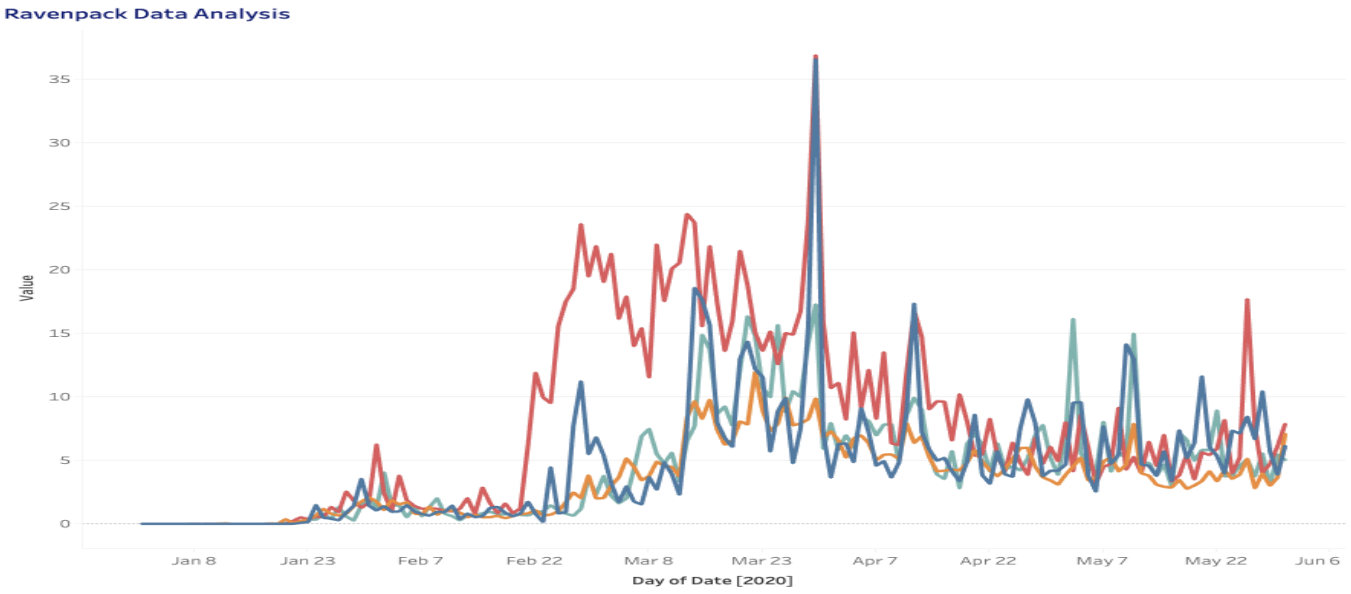


Figure 12

If we look at the Time Series graph of the Panic closely, Iranian people are panicking the most if compared with other three countries depicting the Media hype may have caused the panic among the people which made them to do panic shopping, believing the unrealistic facts and also developed several diseases including mental health.

Surely, to validate the point by analyzing the mobility of these four countries which includes, recreation center, parks, grocery store, residential

Mobility

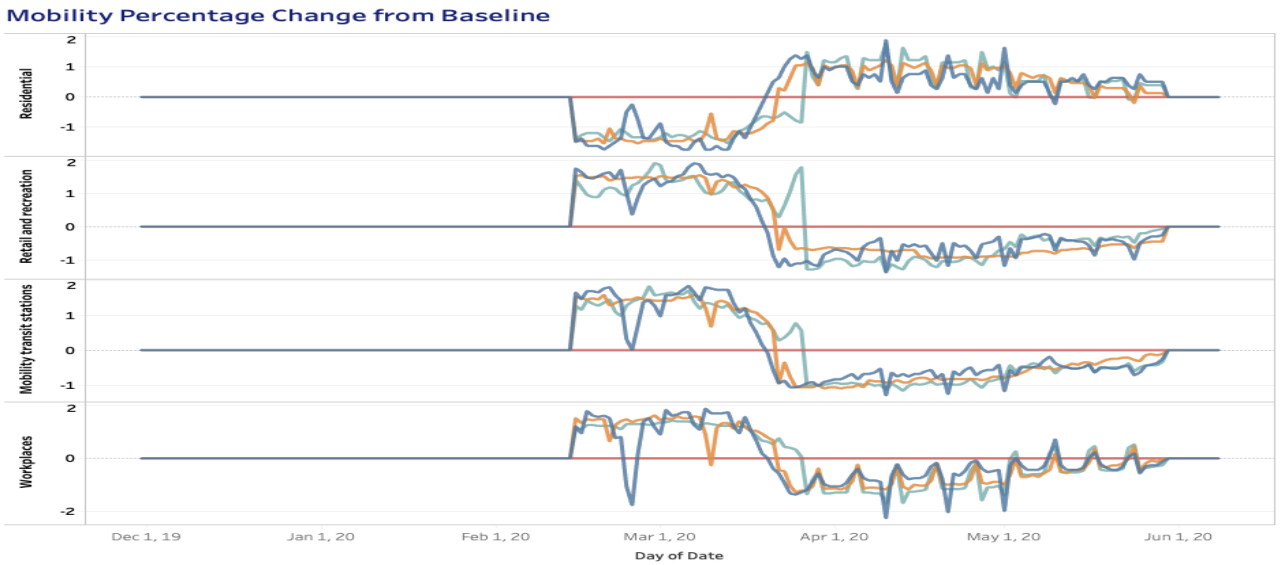


Figure 13

Here, after March 15 the mobility in retail and recreation, transit stations, workplaces are decreasing, however, most mobility is seen in Residential area, may be because of Lockdown. However, it's strange, for Iran, the mobility is less for every aspect. But if we look at the time series graph of new cases below (Fig. 14) it is increasing in Iran, and too from March onwards same when hype and infodemic is increasing making us to interpret that infodemic and hype can be behind the negative sentiment which has its most effect on Iran.

New Cases

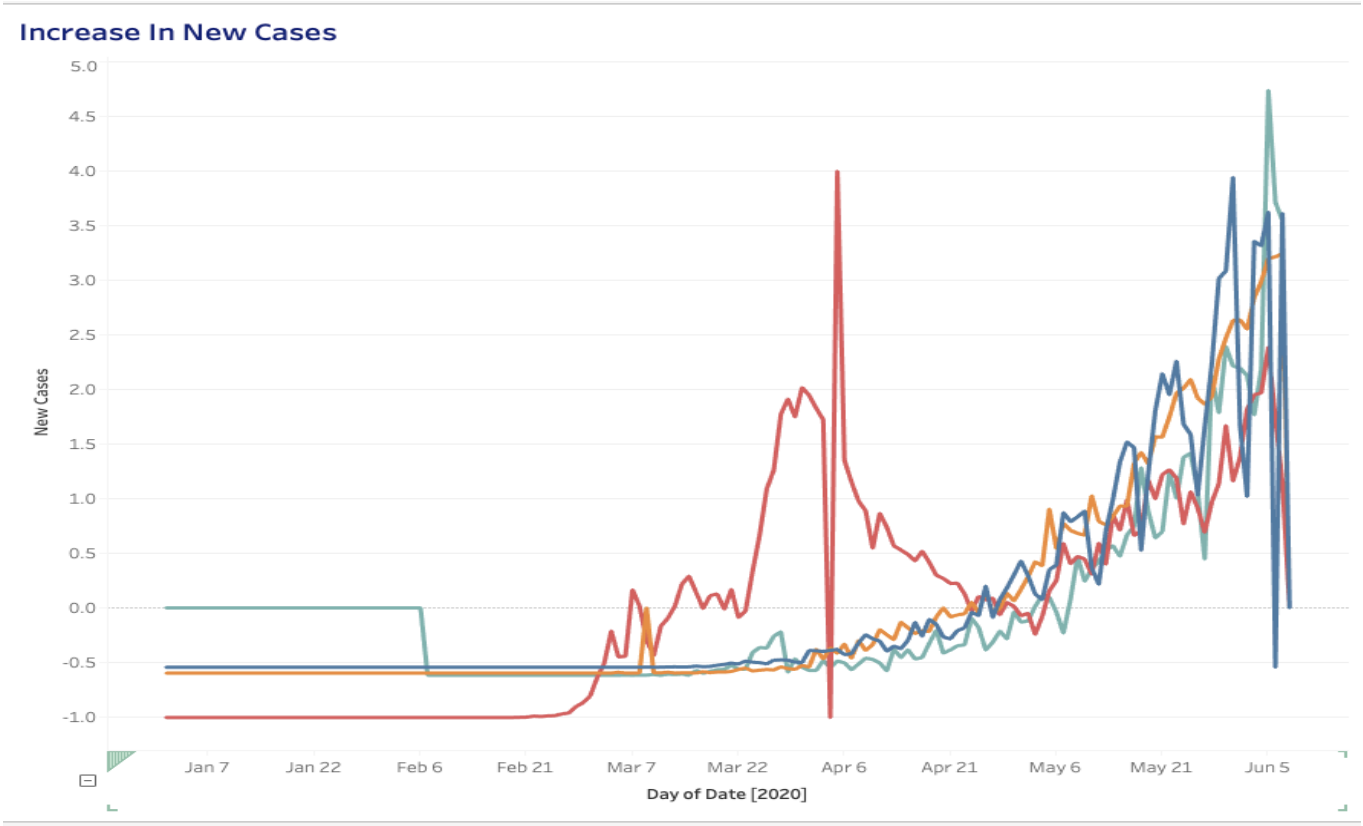


Figure14

Team Signature Assignment

SmartPLS

To implement the below SmartPLS models, we have provided the standardized data as an input and created around nine latent variables from the data. Below models are the reflective models as measures are the indicators (yellow rectangles) of the respective latent variable (blue ellipses) and eliminating one of the indicators will not affect much as other indicators are also the representative.

Epidemiological Data: The latent varable Epidemiological Data contains indicators associated with health related states and events such as new and total cases, deaths, number of testings on COVID-19 patients and recovered cases.

Alternative Media Communication: For analyzing the results of Alternative Media Communication, we have taken the indicators Fake News and Infodemic from the RavenPack.

Media Communication: The latent variable Media Communications has the indicator GDELT Sentiment

Lockdown Measures: The latent variable Lockdown Measures has the indicator stringency index which means strict measures taken by the countries during the pandemic.

Public Anxiety: The public Anxiety latent variable has the indicator panic which defines anxiety the caused during the pandemic.

Mobility: The mobility latent variable has the indicators like parks, retail and recreation, transit stations

Economic policy Uncertainty: The latent variable Economic policy Uncertainty has the indicator newspaper coverage of policy-related economic uncertainty.

Market Volatility: The latent variable Market Volatility represents the risk in the current market due to the pandemic. Higher the volatility, the risker the market.

Country Sentiment: The latent variable Country Sentiment has the sentiment of particular country which is taken from RavenPack.

Team Signature Assignment

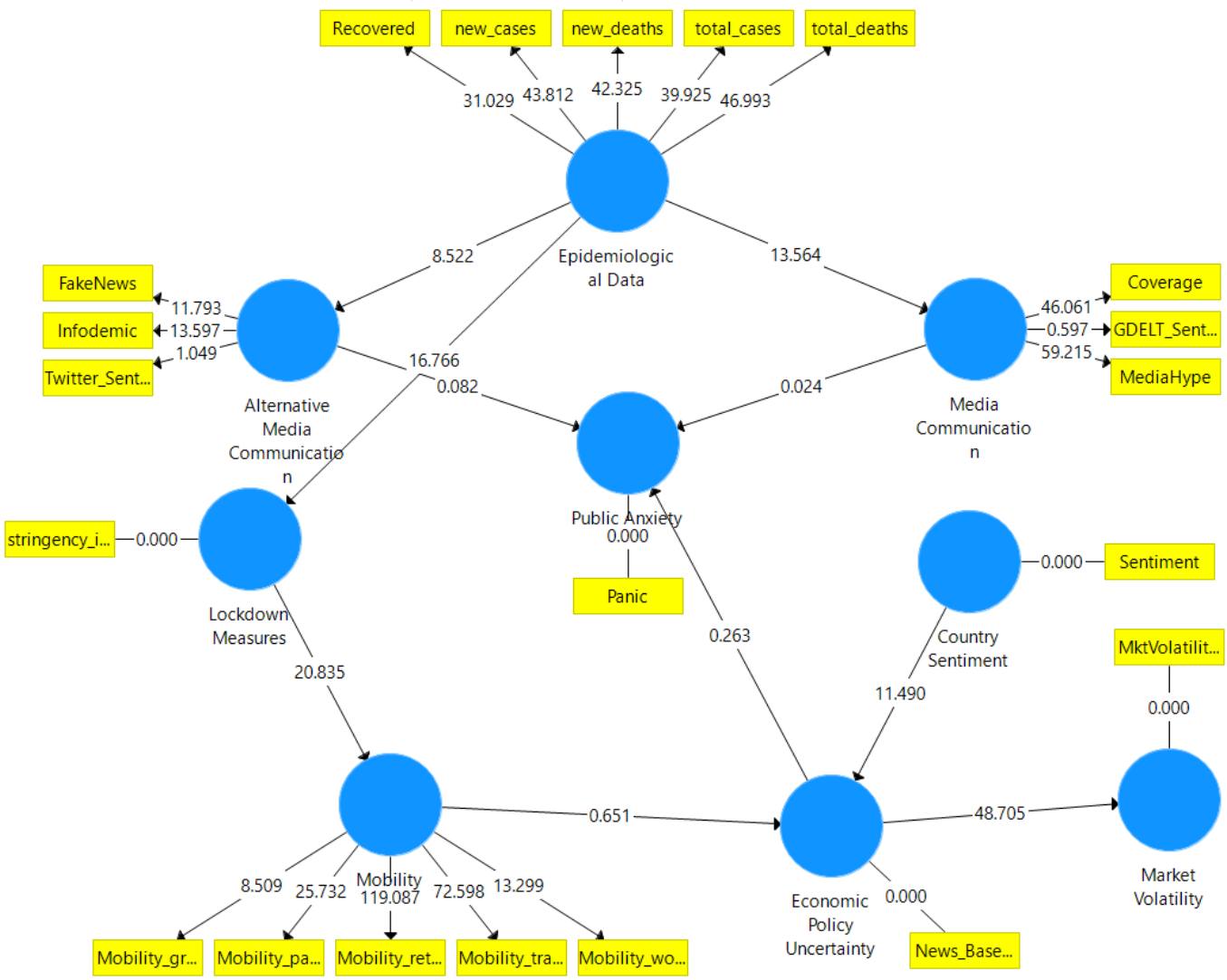


Figure 15 Brazil

In the Fig. 15 above, the path coefficient value between Economic Policy Uncertainty and Market Volatility is 48.705. That means in Brazil, the Economic Policy Uncertainty has around 50% direct effect on the Market Volatility. The Brazilian Institute of Geography and Statistics (IBGE) has conducted the National Household Survey (PNAD) and it is showing that approx. 38 million people in Brazil [1] are self-employed or work in the informal sector. Around 50% of employees from 11 states out of 27 (Fig. 15) in Brazil are in the informal sector hence, people are restricted to work in pandemic and that will result in risking their jobs which leads to recession in the market.

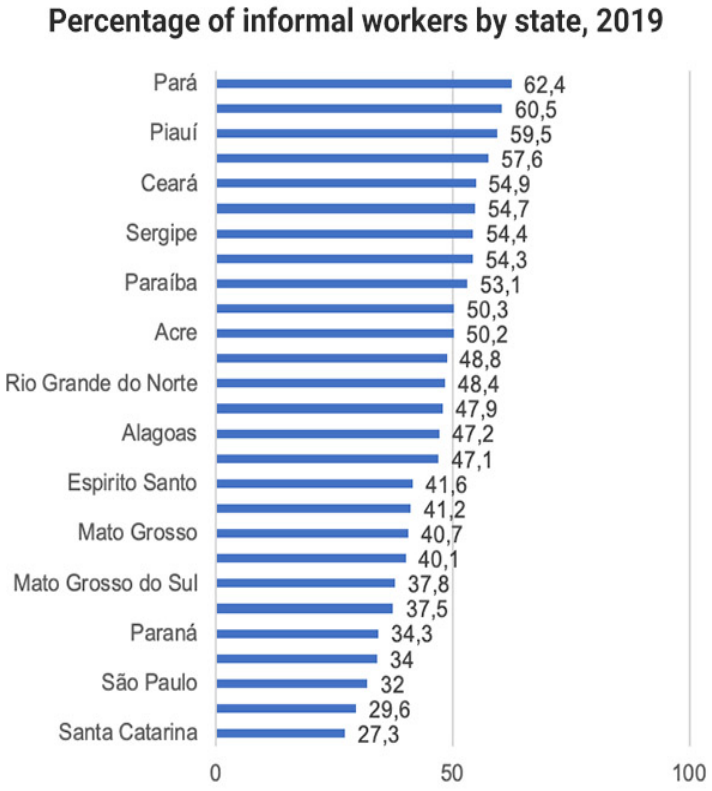

Percentage of informal workers by state, 2019
Pará 62,4
Piauí 60,5
Ceará 59,5
Sergipe 57,6
Paraíba 54,9
Acre 54,7
Rio Grande do Norte 54,4
Alagoas 54,3
Espirito Santo 53,1
Mato Grosso 50,3
Mato Grosso do Sul 50,2
Paraná 48,8
São Paulo 48,4
Santa Catarina 47,9
47,2
47,1
41,6
41,2
40,7
40,1
37,8
37,5
34,3
34
32
29,6
27,3
0
50
100

Figure 16

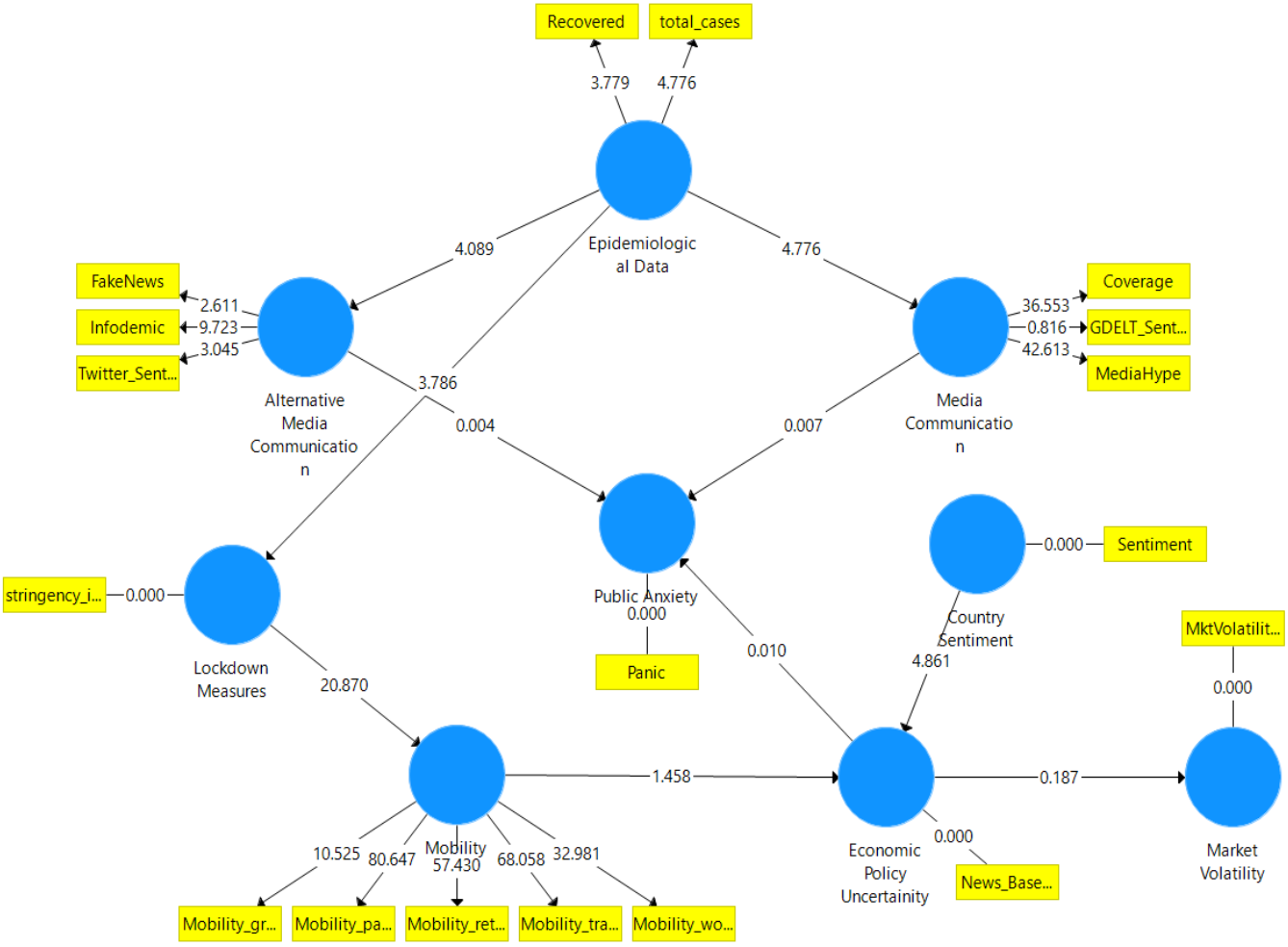


Figure 17 South Africa

We can see in Fig.17, lockdown measures taken during this pandemic affect the mobility most in South Africa. Lockdown measures have around 21% impact on Mobility and if we observe the outer loadings of the mobility, we can see that mobility in parks, restaurants and on transit stations has more than 50% relation with the mobility. According to Wikipedia on lockdown in South Africa [3], all gatherings were restricted and people were only allowed to step out of their houses to seek health services and purchase food/goods. On the other hand, Media communication has a very low impact on Public anxiety in South Africa according to the model. So, the panic caused in South Africa is mostly due to lack of knowledge about the virus and poor communication between government and public [4].

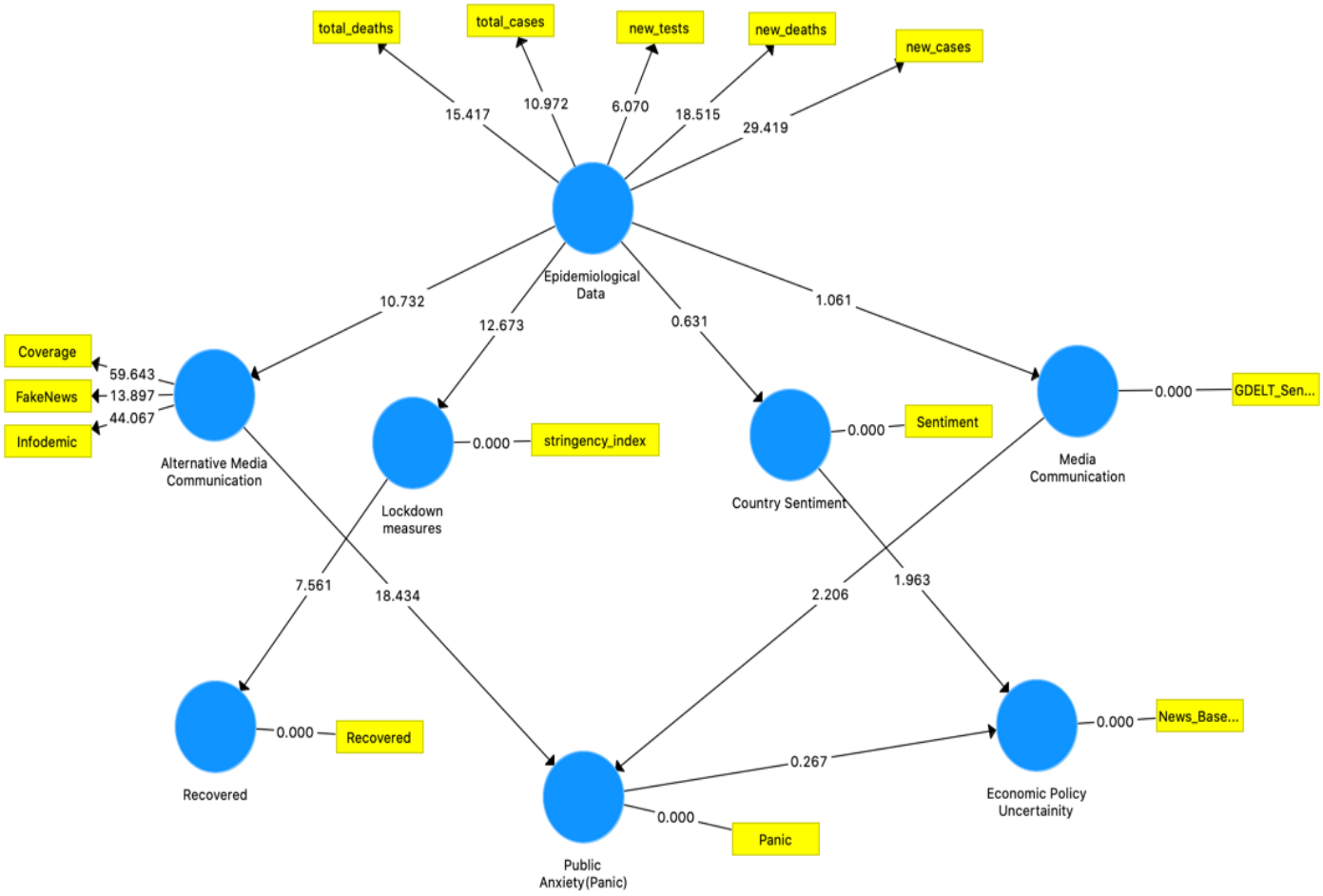


Figure 18 Iran

Above Fig.18 represents the Iran model developed in SmartPLS in which we can see the highest relation between the Epidemiological data and the Alternative Media Communication which contains Fake News, Infodemic, coverage and thus if we come down to the Public Anxiety variable, the increased was observed majorly caused by Fake News and Infodemic. The sentiment among the Iranian people due to COVID was mainly seen negative during the pandemic. Also, the newspaper coverage of policy-related economic uncertainty has major contribution by the Media communication and Public Anxiety.

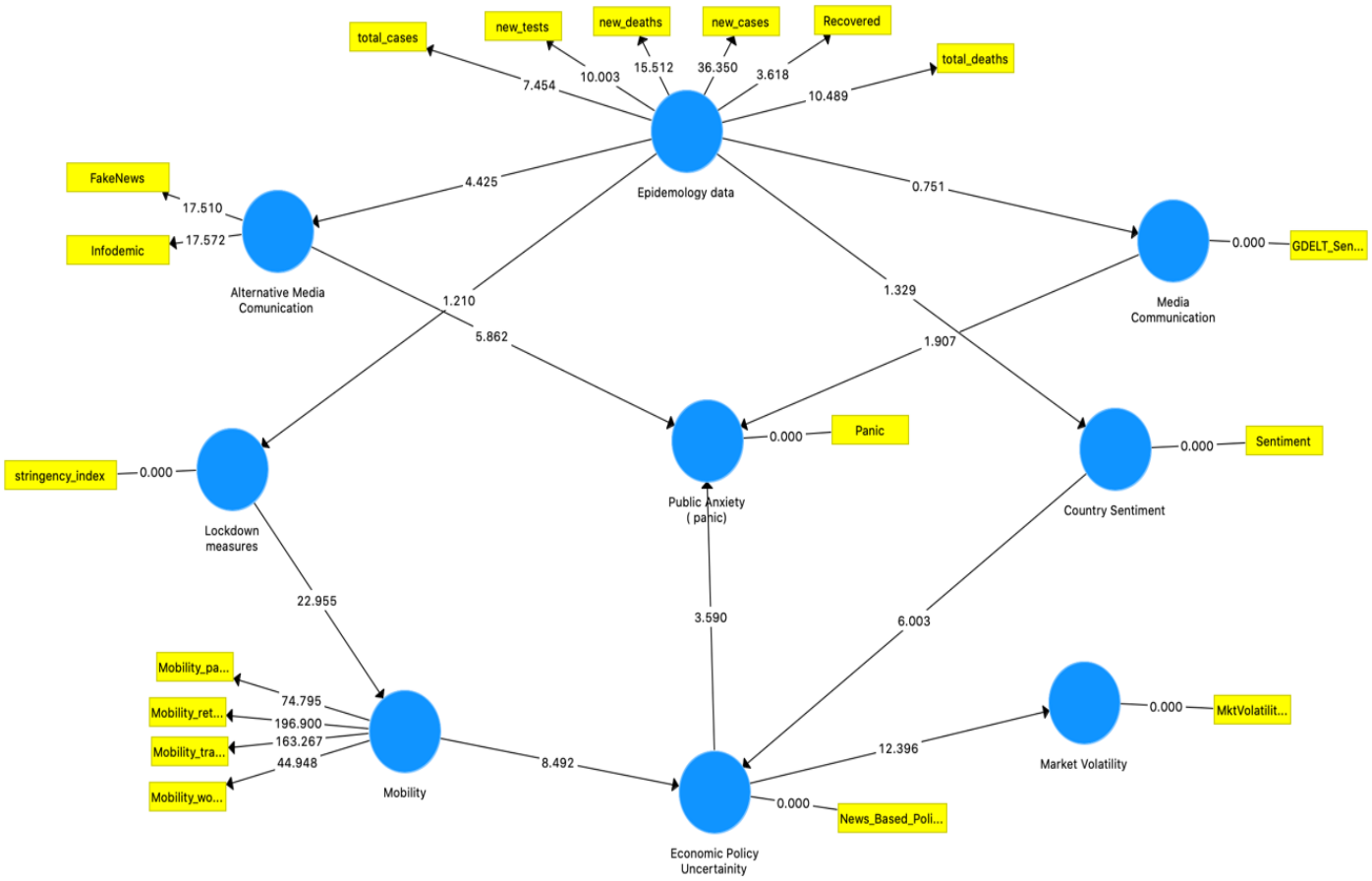


Figure 19 India

Above Fig.19 represents the India model developed in SmartPLS. The panic among the people is less if compared with Iran. The strongest part to notice is the relation between the lockdown and mobility which is mostly seen due to the transit and the retail. The newspaper coverage of policy-related economic uncertainty has impacted due to Mobility, lockdown and the country sentiment.

3) Tips to pass along to next team

As you look at the business model in section 1, we have completed the sentiment analysis of the alternative media communication which is Twitter and main media communication which is GDELT till June 6, 2020. We have worked on RavenPack data, epidemiology, policy uncertainty, mobility and market volatility datasets to draw the impact of COVID-19 pandemic on different aspects. The points we were not able to cover in this project due to the shortage of time are as follow:

Team Signature Assignment

- The classification of GDELT data in Left news, right news and centre news.
- The classification of Twitter data in top twitter and rest twitter.
- Sentiment analysis on government communication(Bag of words for trust in government sentiments)

If we got the opportunity to work on this project again from the start then we will still use the web scrapping technique for data gathering, sentiment analysis in python, data visualization in Tableau and data modelling in SmartPLS. But for data storage we might use Microsoft Azure or Amazon Web Services to make the process faster as data we dealt in the project was huge. We might want to try Flourish for the visualization as it is new and has creative features for COVID-19 data.

References:

[1] Home. (n.d.). Retrieved June 26, 2020, from https://www.ibge.gov.br/en/statistics/social/education/18083-annual-dissemination-pnadc3.html?edicao=27633

[2] Mara Nogueira (2020, June 03). The impact of COVID-19 on Brazil's precarious labour market calls for far-reaching policies like universal basic income: LSE Latin America and Caribbean. Retrieved June 26, 2020, from https://blogs.lse.ac.uk/latamcaribbean/2020/06/03/the-impact-of-covid-19-on-brazils-precarious-labour-market-calls-for-far-reaching-policies-like-universal-basic-income/

[3] COVID-19 pandemic in South Africa. (2020, June 23). Retrieved June 26, 2020, from https://en.wikipedia.org/wiki/COVID-19_pandemic_in_South_Africa

[4] (PDF) THE CAUSE OF PANIC AT THE OUTBREAK OF COVID-19 IN SOUTH AFRICA – A COMPARATIVE ANALYSIS WITH SIMILAR OUTBREAK IN CHINA AND NEW YORK. (n.d.). Retrieved June 26, 2020, from https://www.researchgate.net/publication/339834946_THE_CAUSE_OF_PANIC_AT_THE_OUTBREAK_OF_COVID19_IN_SOUTH_AFRICA_A_COMPARATIVE_ANALYSIS_WITH_SIMILAR_OUTBREAK_IN_CHINA_AND_NEW_YORK

[5] "The GDELT Project." (n.d.). Retrieved from *GDELT*, www.gdeltproject.org/.

Coronavirus News Monitor. (n.d.). Retrieved May 31, 2020, from https://coronavirus.ravenpack.com/india

[6] Introduction to Tweet JSON - Twitter Developers. (n.d.). Retrieved from https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json#:~:text=Tweet JSON formats,Polls metadata, and Exhanced URLs.

[7] echen102. (2020, May 25). echen102/COVID-19-TweetIDs. Retrieved from https://github.com/echen102/COVID-19-TweetIDs