Project Outline

 COVID: Controlling the Outbreak with a Virtual Interactive Dashboard

**The Challenge**

**COVID is a collaborative response on the current coronavirus outbreak. COVID to fight COVID-19.**

With the proliferation of social media use in the Philippines, the availability and consumption of information by the general population has never been so instantaneous. Just as quickly, social media data can also serve as a general indicator of the overall mood of a citizenry, particularly in times of crisis such as the global Covid-19 Pandemic.

Government communication should not only be credible, accurate, and consistent, but should also be timely and relevant to the general populace. The function of communication in a public health crisis is to (1) provide information about risks, symptoms, or treatment of a disease; (2) enable the public to evaluate risks and deal with the threats; and (3) encourage the public to take appropriate precautions and increase compliance with health-care and social interventions. Crisis communication must not, however, create panic. Therefore, the primary challenge is to “raise awareness and concern without inducing irrational behavior.” (Rossmann, 2018)

In order for public health communications to craft timely and relevant messages related to global Covid-19 pandemic, these practitioners must have access to not only the latest available data from vetted sources, but also be able to gauge the “social temperature” of their audience.  By having all the relevant data available in a specific dashboard, policymakers and public communication officers should be able to build clear and credible communication plans to control the overall narrative and in turn lessen feelings of panic and unrest.

**Team**

*Who are involved in this project, what are their project roles and skill sets and affiliations*

Miguel Oscar Castelo – Strategist

Investment Management Strategist

Martin Francisco A. Tinio  - Data Analyst

Environment, Social and Governance (ESG) Research Senior Associate at MSCI Inc.

Marianne Vitug - Technical Lead

Predictive Modeler

Kristel Joyce Zapata – Project Coordinator

Financial market operations background focused predominantly on trade capture, verification, settlements, reconciliations, funding, reporting, process migrations and improvements

**The Problem**

With a current death toll at over 5,800  people and more than 156,000 confirmed cases across more than 100 countries, the coronavirus of 2019 (now known as COVID-19) was announced by the World Health Organization (WHO) to be a global pandemic threatening people’s lives. The virus is notable due to its unpredictable nature and high infection rate. Statista reported that little is known about the virus apart from its likely incubation period of 2-14 days and approximately 2.2 percent mortality rate.

Given the abundance of global data available, COVID-19 reporting by public health officials can be presented in an accurate and technical manner, but will an abundance of facts help assuage the general public?

**Our Solution**

In this work, we propose the creation of a sentiment analysis dashboard, which captures all pertinent information on COVID-19 including people’s sentiment on past or current trends. This dashboard is aimed at enabling the pertinent decision-making authorities to make any necessary adjustments, announcements or actions regarding the current situation, as well as craft an overall plan to manage the narrative by applying any data-derived insights into future communication efforts.

**Related Literature**

We have conducted research and cite the following prior work that provide the foundation for this solution:

In Adikhari et al (2018), *Sentiment Classifier and Analysis for Epidemic Prediction*, researchers proposed using sentiment analysis as an method to predict epidemics by utilizing social media data gathered from twitter and facebook and utilized machine learning to provide a probability score for an area to become infected during an epidemic.

In Rossmann et al (2018), *The Mediated Amplification of a Crisis: Communicating the A/H1N1 Pandemic in Press Releases and Press Coveragein Europe,* researchers conducted a quantitative content analysis on press releases disseminated by government and public health authorities, as well as tabloid news articles during the recent A/H1N1 Pandemic.

In Kim, et al (2017) *Resilience in risk communication networks: Following the 2015 MERS response in South Korea,* researchers conducted an analysis of inter-agency crisis communication practices during the 2015 MERS epidemic in South Korea.

In Regan et al (2014),  *Risk communication and social media during food safety crises: A study of stakeholders’ opinions in Ireland,* researchers analyzed the effectiveness of social media as a tool for information dissemination of food safety alerts from the perspective of the general populace.

**Objectives**

*Detail the business objective, what question are you trying to answer, for whom, and why is this important?*

The objective of the project is to provide a dashboard as a monitoring tool that would be of help to public health authorities in making decisions and actions. The dashboard would follow a three-part structure: (1) descriptive, (2) sentiment analysis, and (3) topic clustering or analysis.

i.           The descriptive part would present the basic numbers on this pandemic including the number of cases, deaths and countries affected. The rate change on the numbers could be integrated as well. This would help show the current scenario or what is currently happening because of COVID-19.

ii.          The sentiment analysis part would measure the overall sentiment or outrage of the netizens on COVID-19. Moreover, this can be a great check on the populace’s panic level on the outbreak and also see whether people feel safe or not from the threat of not only COVID-19, but of future epidemics of similar ilk. This would be beneficial in knowing the right risk communication strategy to be utilized at the right time.

iii.         The topic clustering or analysis on the other hand would show the trending topics that public health could focus on or address especially if there happens to be misinformation.

Insights derived from these analyses will aid public health authorities have data-driven tools to analyze and determine if and when necessary actions, announcements or adjustments are vital to control the public situation on this health issue. This could also be useful in reducing misinformation and providing factual information.

In addition, the analysis of public sentiment on COVID-19 would also be of help in managing reputational risk, ensuring that public health authorities continue to operate in the best interest of public health.

**Data** - source, extraction, management, pre-processing (hashtag analysis, polarity generation

To come up with the dashboard, the required data are as follows:

* COVID-19 cases and deaths by date and country (source: [LINK](https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6))
* Tweets about COVID-19 (source: Twitter API)

**Tweet  Extraction**

To analyze the sentiments of netizens about COVID-19, tweets will be extracted using Twitter API and rtweet package. Using the keywords *“#Covid\_19PH”*, *“Philippines AND lockdown”*, *“ANO BA TALAGA”* and *“Duterte AND Covid Philippines”*, tweets were scraped on March 14, 2020 and the Twitter API returns tweets up to 7 days prior to the extraction date. The keywords were chosen based on the Twitter Philippines trends during the day of extraction in order to get relevant tweets as much as possible. Single extraction was done per keyword with a tweet limit of 3000 tweets. However, because of Twitter API’s limitation, there were only a total of 7839 tweets collected.

For polarity and emotions analysis, only english tweets were considered since R packages can only assign polarity and emotion scores for english words. This restriction can be indicated automatically upon extraction. Only 3610 tweets were extracted for this purpose.

**Data Analysis**

Before analyzing the extracted tweets, pre-processing was done using tidyverse, tidytext, tm, and textdata packages. This includes removing the stopwords, links, hashtags, punctuations, numbers, and whitespaces in the tweet. This is necessary to ensure that only relevant words are included in the analysis since most algorithms used depend on bag-of-word principle.

Package syuzhet was utilized for analyzing the polarity and emotions and polarity of the tweet. The NRC sentiment dictionary was used to calculate the presence of eight different emotions and their corresponding valence in a tweet. Polarity includes positive and negative while emotions includes trust, fear, anticipation, joy, anger, sadness, surprise, and disgust. The limitation of this analysis aside from only accommodating English tweets is it cannot identify sarcasm. So it is better to read some of the tweets that fall under the emotion and polarity categories to get an idea of the accuracy of the results.

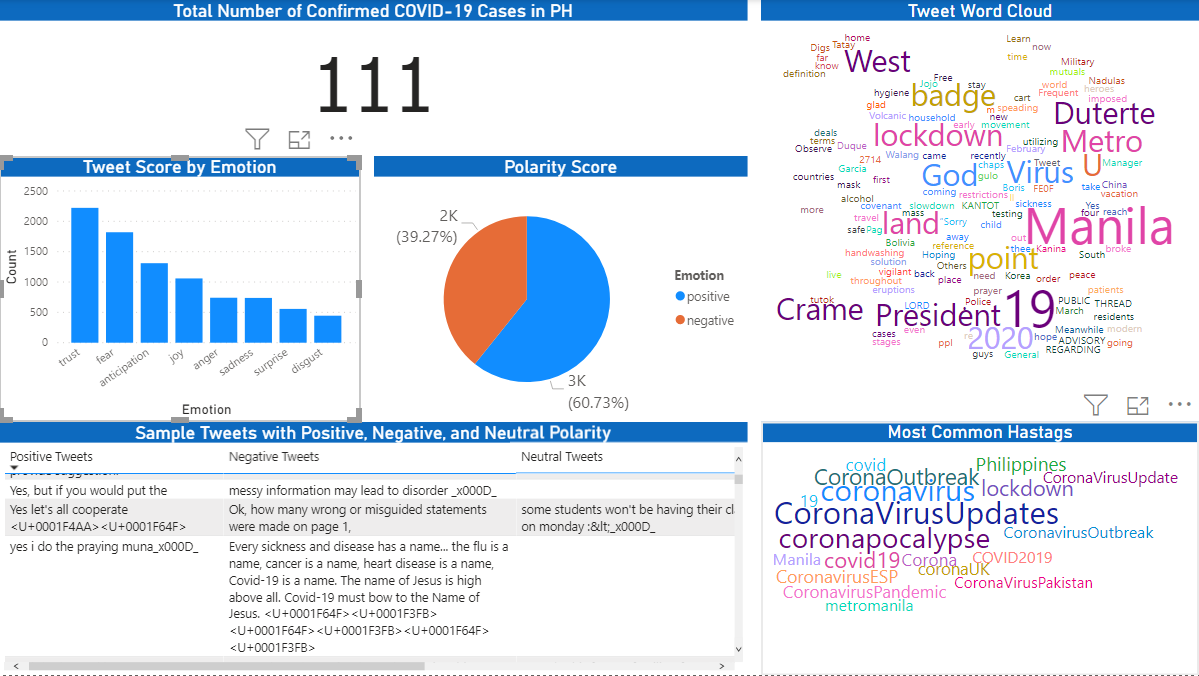
Finally, topic analysis using Latent Dirichlet Allocation was utilized to show in summary what people are talking about and further explain the sentiments/polarity results.

Data extraction, data pre-processing and data analysis were all done using RStudio.

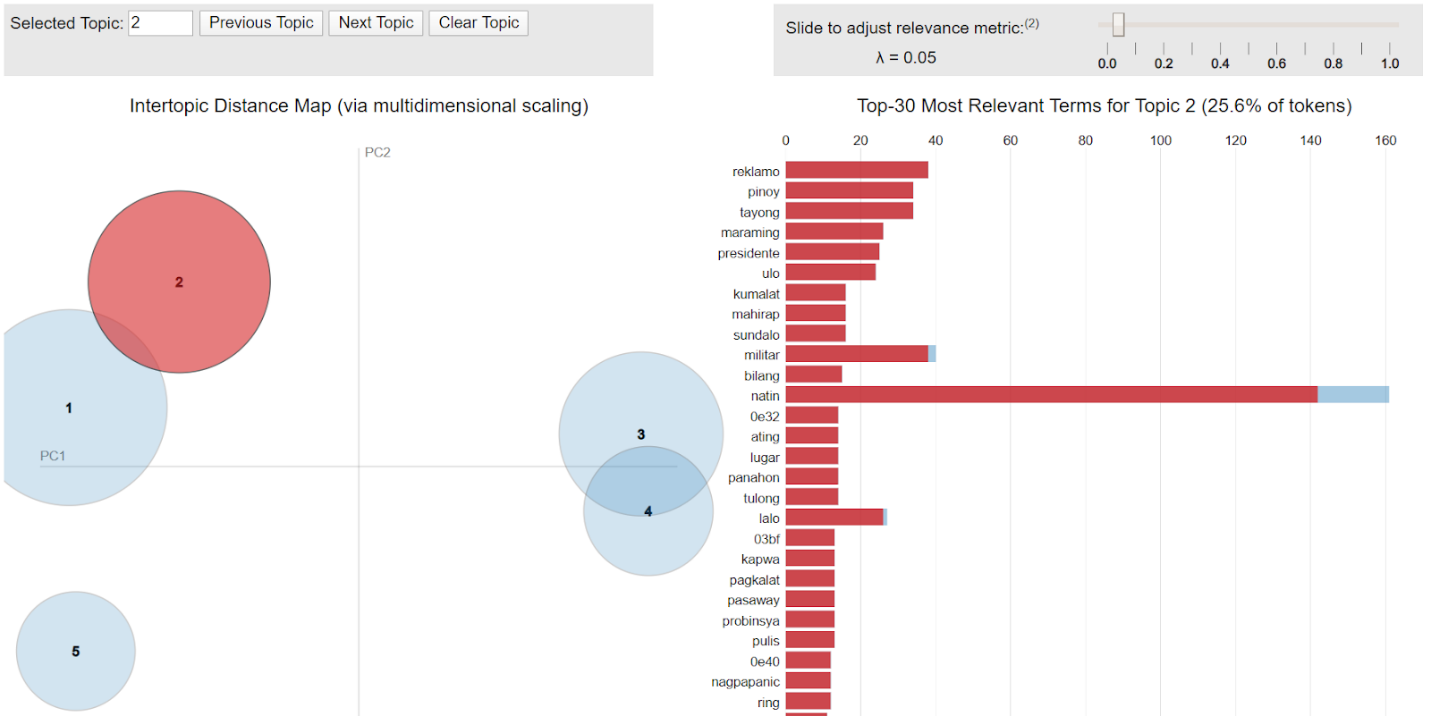
**Visualization**

Power BI was used for the dashboard that showcases the counts of COVID-19 confirmed cases, word cloud for tweets and hashtags, list of positive, negative, and neutral tweets, and count of emotion and polarity scores. Figure 1 shows how the dashboard looks like.

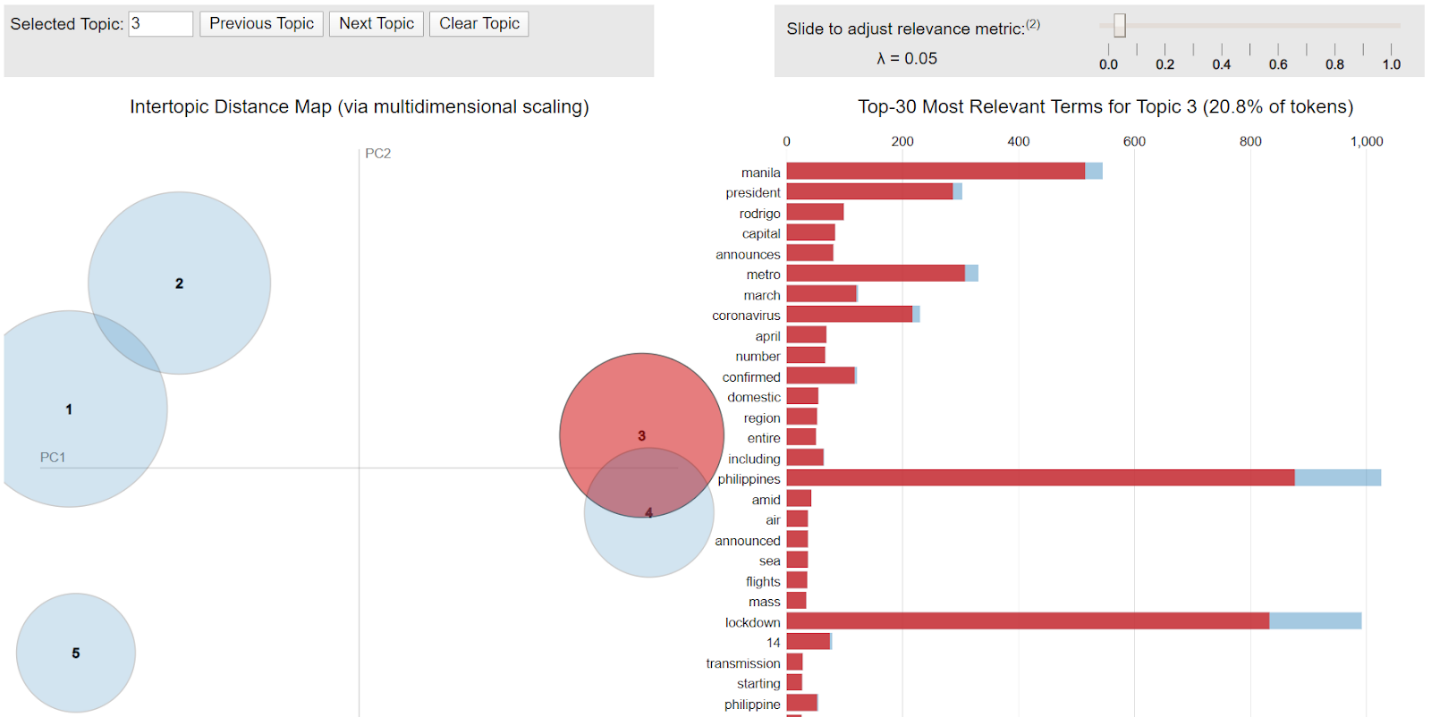
For topic clustering, LDAvis package was utilized and was separated from the dashboard. Figure 2 and 3 shows that topics 1 and 2 are overlapping, same is the case for topics 3 and 4. Upon examining the topics, it is best to interpret topics 2,3, and 5 based on the words that’s included within them. Reading some of the raw tweets helps in interpreting the topics. Topic 2 talks about the problems that we experience during this time including words like mahirap, reklamo, militar, tulong, probinsya, and nagpapanic. Topic 3 is about the lockdown announcement that was recently made as reflected by words like manila, president, rodrigo, announces, air, sea, flights, and lockdown. Finally, topic 5 shows panic tweets which include words like gulo, confused, labo, mood, puta, fake, putangina, and ano ba talaga.



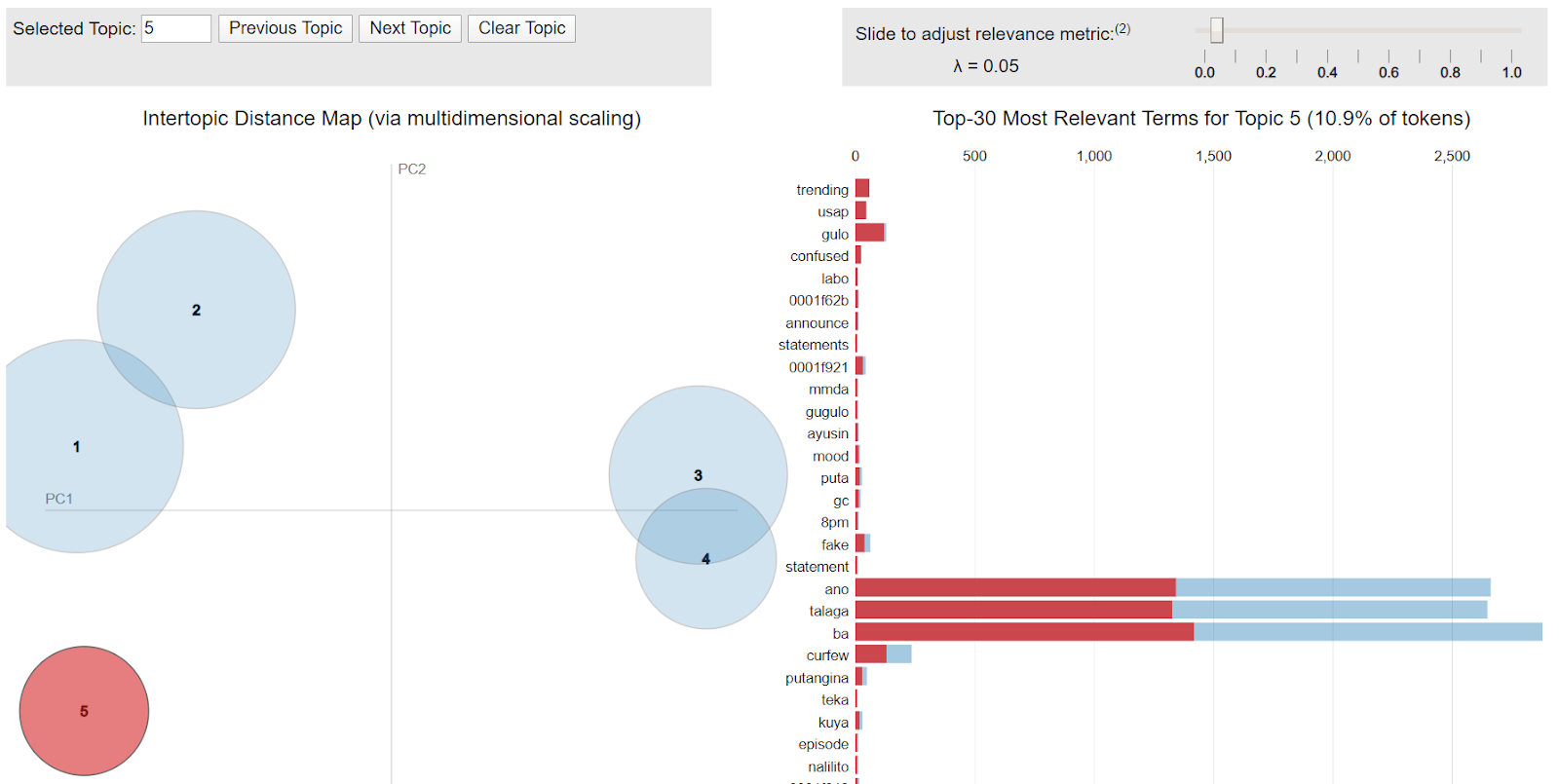
*Figure 1*: Dashboard showing the total cases, polarity, and emotion analysis results



*Figure 2*: Visualization for Topic 2



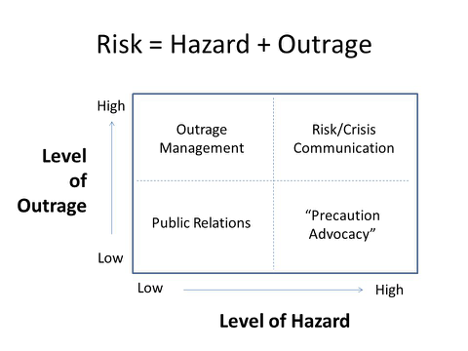
*Figure 3*: Visualization for Topic 3



*Figure 4*: Visualization for Topic 5

The team has interviewed a medical doctor in the communications research field, Dr. Jason Ligot, who shared that one aspect that public health agencies or authorities can focus on and improve is risk communication. Based on the World Health Organization (WHO)’s risk communication training module selecting the best risk communication strategy would utilize the below model created by Peter Sandman. In order to measure the right level of risk, one must take into consideration both hazard and outrage. Hazard pertains to how dangerous the disease is. What are the mortality and morbidity rates? On the other hand, outrage is measuring the emotional response of the population affected. Level for both factors is measured on a low to high scale. Not in the framework but also a great factor to consider is the novelty of the disease.

The team’s sentiment analysis then comes into picture in measuring the level of outrage. Given this as well as the mortality / morbidity rate provided by the epidemiologists, the public health authorities or agencies can then decide the appropriate decisions and actions that they can employ.



Based on the insights generated from the sentiment analysis, the tool is designed to inform its users on what exactly does the general populace think and feel about COVID-19? It can also help determine the effectiveness of government information dissemination. This in turn should help inform future actions of public health policy makers.

With this tool, public health authorities will have all the necessary information to craft the narrative around public sentiment. By utilizing only relevant, factual information, aimed at managing the narrative, the public can be better informed and educated on the proper preventive measures to contain and control the pandemic.

**Conclusion and Recommendation**

Recommendations for future work

Analysis can be improved in many ways. First, using Twitter API has restrictions in pulling out tweets so it is possible that the scraped tweets are not representative of the population being studied. Using a firehose or streaming API will provide more tweets and accurate analysis. Next, it is better to include the tweets written in Filipino for the emotion and polarity analysis. However, to do this, manual tagging of tweets is necessary since it is not yet available as an r function. We can also build a Filipino lexicon that incorporates an emotion tagging to words so this type of analysis can be automated later on.

The work done by the researchers were mostly social listening using Twitter. Future researchers can further develop the study by making it social analytics one, that is, incorporating the predictive or prescriptive side or putting more concrete actions on the insights generated. We envisioned a one-stop COVID-19 information portal that is real-time. The prescriptive part could possibly generate the right response that the public authorities could follow considering the weight of the current alert levels and the sentiment analysis. Moreover, a news aggregator could also be something to be looked at.

Git hub

<https://github.com/MarianneVitug09/Corona_Virus>