STEM Fellowship Big Data Challenge 2021

Visualizing Regional COVID-19 Vulnerability Based on Misinformation Prevalence

Sina Allen[1], Jessica Son[2], Tony Sun[3], and Matheus Zhu[4]

[1]University of British Columbia
[2]Trinity Western University
[3]University of Waterloo

June 1, 2021

Abstract

The overflow of correct and incorrect information has been affecting citizen's lives more than ever after the pandemic hit. As Abraham Lincoln said, "Let the people know the facts, and the country will be safe", the accuracy of the information being spread to the public takes high importance in the citizen's wellness and safety. To validate this assertion, this paper attempted to visualize the regional COVID-19 vulnerability with correlation to the regional misinformation prevalence. Using the datasets from various online sources and projects, the Fake News Per Capita and Cases Per Capita were calculated to represent the regional COVID-19 vulnerability. The prevalence of misinformation was calculated to compare with the Cases Per Capita values. As a result, the linear regression showed a positive correlation between the two values. Considering various factors that contribute to a region's infection rate, its partial prediction could be made using the predominance level of misinformation.

Keywords
COVID-19, COVID-19 Vulnerability, Misinformation, Predictive Model

1 Introduction

The importance of accurate news is emphasized with the infodemic that has emerged with the COVID-19 pandemic. However, is fake information really relevant to the citizen's lives? If so, what type of fake information would there be, and how would they be threatening our lives? Would the region with more false information lead to higher danger in its citizens wellness's? [5]

The far-reaching effects of the COVID-19 pandemic has individuals from every continent scrambling for information, often through outlets with variable quality control such as social media channels. The lack of highly skilled professionals, such as physicians trained in virology, immunology, and epidemiology, combined with the ease of dissemination, sets the stage for rapid exchange and evolution of poorly informed and misleading ideas and hypotheses. The flux of such unverified information could endanger the health and safety of many.

This study aims to investigate the extent to which a region's COVID-19 vulnerabilities are affected by the proliferation of misleading and poorly backed advice and statements.

2 Materials & Methods

2.1 Sourcing the Dataset

The COVID-19 cases data for each country is obtained from Our World in Data (OWID) through Johns Hopkins University. The university's Coronavirus Research Center collects data from more than 260 government agencies at various levels internationally. The total cases for each country until May 26th, 2021 was collected.[6] The population information was sourced from a publicly available dataset scraped from Worldometer, with a total of 235 countries.[7]

The FakeCovid project is the main source of global COVID-19 misinformation data, and it is the first multilingual dataset of 7623 fact-checked news articles 04/01/2020 to 01/07/2020 from 105 unique regions.[8]

Lastly, some data about populations of countries that needed supplementary information were retrieved from Macrotrends.net [1],Countrymeters.info [2], and Statistictimes.com [3], [4]. All of these website's data are sourced from United Nations Department of Economic and Social Affairs.

Using these datasets and a variety of data visualization tools, we will attempt to model and explain the underlying patterns betweeen variables.

2.2 Data Processing

For all of our data analysis we used Colaboratory from Google Research and Excel for formatting and some simple modifications of our data set's csv files. Python 3.7.10 was used to make our data visualizations come to life, using libraries such as pandas, matplotlib and seaborn. The collections library were used for data manipulation. Before running our data analysis however, we modified and combined values from the aforementioned datasets in order to gain insight into the relationship between the number of fake covid news and covid cases per country.

Using our population data, we converted fake news per country and covid cases per country values to be per capita. This was done through dividing the number of fake news and covid cases by the population of the respective country, receiving fake news per capita and covid cases per capita. It is worth mentioning that the values were scaled up for easier analysis. The equations used are

$$FakeNewsPerCapita = \frac{FakeNews}{Population} \times 10^8 \quad (1)$$

$$CasesPerCapita = \frac{Cases}{Population} \times 10^4 \quad (2)$$

Lastly, we combined our two newly generated values to make a new csv containing data for: population, fake news, covid cases, fake news per capita and cases per capita, all corresponding to their respective countries. From this dataset, two types of graphs were generated: bar graphs and regression plot graph. Bar graphs were selected to give a better understanding of the data. A regression plot was created to visualize the relationship between fake news per capita and cases per capita.

3 Results

3.1 Visualizing Misinformation

Compiling a collection of various types of misinformation, we are able to generate a graph of total amount of published misinformation per country for most of the counties in the world (see figure-1). This includes articles and widely circulated claims.

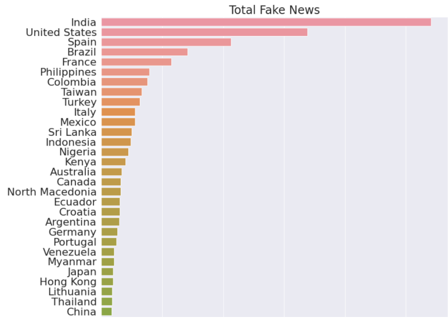

Figure 1: COVID-19 related misinformation amount for each country, showing top-30 values.

To standardize, we calculated the amount misinformation per capita of each country (see figure-2).

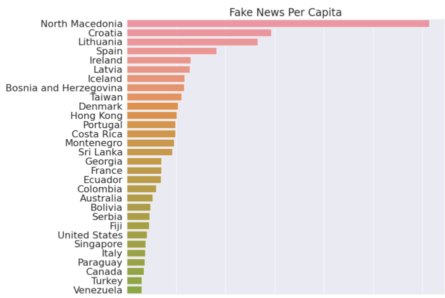

Figure 2: COVID-19 related misinformation amount for each country per capita, showing top-30 values.

Using our Covid-19 cases data set, we generated a bar graph of cases per capita for every country with data on amount of misinformation (see figure-3). Again with these graphs, we only care about the values relative to each other therefore the scaling does not have an effect on our overall results.

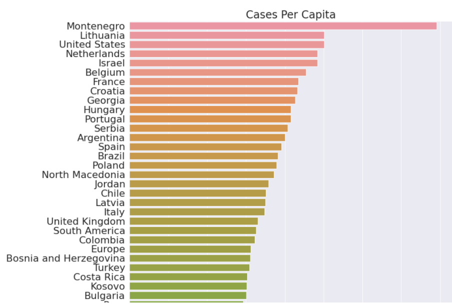

Figure 3: COVID-19 cases for each country per capita, as of May 2021, showing top-30 values.

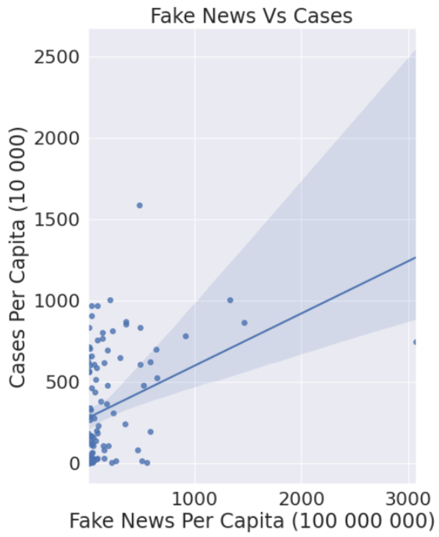

Figure 4: COVID-19 related misinformation amount per capita.

3.2 Interpretation

Using the values from the above two bar charts, we generated a regression plot of COVID-19 Misinformation news per capita against COVID-19 cases per capita of each country (see figure-4). The plot generated suggests a positive trend of the impact of misinformation on COVID-19 cases, shown by a positive slope. The coefficient of determination of the regression model is 0.129.

4 Discussion

Given that the linear regression, which highlights a positive correlation, has an accuracy of 12.87%, we can make the following conclusion. First, we conclude that case infections can be partially predicted by misinformation prevalence. Although the confidence interval is large, which indicates that numerical predictions might be inaccurate, the upward trend visible in Figure 4 showcases the general relationship between the two parameters achieves the goal of this study. Due to the restrictions in number of geo-tagged datasets on COVID-19 misinformation, performing regression on a scatterplot was the most feasible. The low degree of accuracy is to be expected, as COVID-19 cases depends on a complex web of socioeconomic, political, and cultural factors, of which misinformation is only a part of. Additionally, with various media regimes and standards in different regions, it is impossible to determine the true misinformation count. A classic example would be the case of North Korea, which has only one piece of false article. On the other hand, the varying standards also affect COVID-19 case count, which might not be accurate even for developed countries, much less for rural regions. China is a good example in this aspect. These errors are fundamental in the analysis of this paper and contributes to the error. Although the accuracy of 12.87% might not be useful in deriving accurate predictions going forward, the relationship determined in this paper despite uncontrollable error reaffirms the importance of information accuracy in saving lives.

Conclusions

As waves of COVID-19 peak and trough, news outlets such as social media play an evermore important role in our [COMPLETE THIS SENTENCE].This paper aimed to use big data to gain insight into how COVID-19 misinformation could impact the number of COVID-19 cases in a given country. Using simple data analysis methods and data visualization tools with Python we were able to generate results that although were not accurate, achieved the goal of our study. Our findings suggested that there is a positive correlation between amount of covid fake news and covid cases in a country. By how much exactly is unknown at this stage, but we believe that this is something worth further investigation for more accurate results. We suggest for future studies to be conducted that perhaps use a complete data set of amount of COVID related misinformation from a specific regions of a country such as the United States and analyzes how different regions with different exposures to COVID-19 misinformation behaved in regards to the pandemic. Furthermore, analysis of how the impacts of COVID-19 change as the rate of misinformation changes through out periods of time could further verify the significance of infodemics on public health.

Acknowledgements

We would like to thank STEM Fellowship for this amazing event and the support and for making resources available to us.

References

[1] United Nations, Department of Economic and Social Affairs. *2019 Revision of World Population Prospects*. 2019. URL: https://www.macrotrends.net/countries/AFR/africa/population.

[2] United Nations, Department of Economic and Social Affairs. *2019 Revision of World Population Prospects*. 2019. URL: https://countrymeters.info/en/Kosovo#:~:text=Demographics%20of%20Kosovo%202020%20As%20of%201%20January,births%20exceeded%20the%20number%20of%20deaths%20by%204%2C240.

[3] United Nations, Department of Economic and Social Affairs. *2019 Revision of World Population Prospects*. 2019. URL: https://statisticstimes.com/demographics/south-america-population.php#:~:text=The%20population%20of%20South%20America%20is%20projected%20to,to%20427%20million%20or%20427%2C199%2C446%20people%20in%202019.

[4] United Nations, Department of Economic and Social Affairs. *2019 Revision of World Population Prospects*. 2019. URL: https://statisticstimes.com/demographics/europe-population.php#:~:text=As%20of%201%20July%202020%2C%20the%20population%20of,the%20third-most%20populous%20continent%20behind%20Asia%20and%20Africa.

[5] Hasell, J., Mathieu, E., Beltekian, D. et al. *A cross-country database of COVID-19 testing*. 2020. DOI: https://doi.org/10.1038/s41597-020-00688-8.

[6] Hasell, J., Mathieu, E., Beltekian, D. et al. "A cross-country database of COVID-19 testing." In: 345 (2020). DOI: https://doi.org/10.1038/s41597-020-00688-8.

[7] Prabhu, Tanu N. *Population by Country - 2020, Version 1*. 2020. URL: https://www.kaggle.com/tanuprabhu/population-by-country-2020.

[8] Shahi, Gautam Kishore and Nandini, Durgesh. *FakeCovid-A Multilingual Cross-domain Fact Check News Dataset for COVID-19*. 2020.