Data Incubator's Data Science Fellowship- Feb 2021
Behnaz Jafari
MSc Biostatistics
Email: Behnaz.Jafari1@Ucalgary.ca

Investigating the effect of socioeconomic factors on the transmission dynamic of COVID-19 in a country: A tweeter-based surveillance.

Background:

COVID-19 has disrupted lives and livelihoods and caused widespread panic worldwide. Emerging reports suggest that people living in some poor/least developed countries are more susceptible to COVID-19. However, there is a lack of quantitative evidence that the speed of the spread of COVID-19 can shed light on whether residents of poorer countries are more likely to get exposed more to COVID-19 than residents richer countries and how much these socioeconomic factors affects the transmission dynamic of COVID-19 .

During this time, social media involvement and interaction have been increased dynamically and share one's health situation and point of view regarding the transmission dynamic of COVID . From user-generated content on social media, we can analyze the public's thoughts and sentiments on health status, concerns, panic, and awareness related to COVID-19 in different countries to find out the situation of speed of COVID transmission in the countries. Having the tweets data and socioeconomic situations of these countries can help us to understand the relationship between the spatiotemporal and socioeconomic factors on transmission dynamic of COVID-19 in a country. These Analyzing such information can help policy makers and health care organizations assess and predict the needs of their stakeholders and address them appropriately.

Method:
We choose countries from different economic level (Developed, Developing and underdeveloped) that we have Social and Economic indices (e.g. ESocial Progress Index (SPI), GDP) for them. Leveraging a set of tools (Twitter's search application programming interface (API) and Tweepy Python library and Natural Language Processing (NLP) Python library) and using a set of predefined search terms ("corona," "2019-nCov," and "COVID-19"), we extracted the text and metadata (user profile information including the country, number of followers) of public English language tweets from in a window of time that was in high pick of pandemic in all those countries using NLP tools. We analyzed the collected tweets using word frequencies of single (unigrams) and double words (bigrams). We can also perform sentiment analysis and extracted the mean number of retweets, likes, and followers for each topic and calculated the interaction rate per topic. We can build a model to check and predict the relationship between the indices and the transmission dynamic of the COVID.
The following graph explain the process schematically.

Investigating the effect of socioeconomic factors on the transmission dynamic of COVID-19 in a country: A Tweeter-based surveillance.

Twitter
[In a specified time frame]

Specify the number of tweets for each country

Remove non-English Tweets

Remove non-English Tweets

Remove retweets and irrelevant Tweets

Divide Data to training and testing based on Countries

Analyse tweets for COVID transmission dynamics based on the number of tweets about COVID

Build and test the model based on testing data