Preliminary Inquiry

Work Group 1

November 6, 2020

1 Question of Inquiry

Throughout the COVID-19 epidemic, there have been changes in people's attitudes. Lots of people posted their thoughts on Twitter. In this project, we want to study how the growth and trend of the pandemic influence people's responses on Twitter. For example, we want to see whether the number of negative tweets would increase if the total number of confirmed cases goes up. Narrowing down to a statistical question, we want to study the impact of the severity of the disease on the tweets on Twitter.

2 Dataset

For this project, we look at the tweets with #covid19 hashtag on Twitter. The dataset we found contains tweets from July 25,2020 to August 30,2020. The columns include usernames, locations(countries) and texts etc.. There are 181061 well-formatted observations. For the first step, we want to analyze the texts and extract the messages indicating people's attitudes. This step can be problemetic because it's hard to decode the emotions in the texts without accually reading them. Our contemporary thought is to search for keywords such as "mad", "happy" etc.. Once we find a way to categorize the tweets, we can set it as the response variable. For the next step, we look for the total number of confirmed cases and total number of deaths in the users' regions and set them as the explanatory variables. Eventually, we can build linear models based on these variables.

In addition, there are some other interesting information in this dataset, such as the followers, friends, favorites of each user and whether the tweet is retweeted. These variables can also be added to the model once we found a way to quantify them.

One of the advantages of this dataset is the magnitude of its size. It contains about 80% of all countries in the world, which allows us to study the geographical distributions of the disease. The major limitation of the dataset is the lack of time span - all the data are ranged within a month. Therefore, we might not be able to get a general result. Another limitation of this dataset is that it doesn't have the total number of confirmed cases, which we need to find in other datasets.

3 Model

We construct the preliminary model as the following:

$$S \sim D + I \tag{1}$$

where S is the multilevel categorical sentiment representation of the tweets, D is the total number of deaths from the user's region and I is the cumulative total number of infections from user's region.

Explicitly, the model to produce label S_i for tweets T_i would be a function of Tweets, with the following form:

$$S_i \sim g(T_i) \tag{2}$$

To find the best function $g(T)$, we combine multiple existing methods, including Bag of Words, Word Vectors, Support Vector Machines. We start with performing information extraction with models such as Bag of Words and then proceed with more advanced models in order to capture the sentiment information from the tweets.

In the context of the problem, the model will be able to verify the relation between the mood of the tweets and the spread of COVID-19. Optimisticly, it will be able to show whether such a relation exists and what would be the nature of this relation. However, the applicability of this model lies on the accuracy for sentiment labelling. It is widely believed that the labelling of the sentiment data is somewhat unstable at this stage. According to the research of Aproov Agarwal, Boyi Xie, "Sentiment Analysis on Twitter Data"[1], the highest accuracy they could achieve after comparing multiple models is around 76% for a binary task. Consider the cases where the function $g(T_i)$ produces unaccurate labelling prediction S_i for the tweet T_i. Such a mislabel may affect the accuracy of the final model, $S \sim D + I$. Unfortunately at this stage we are unsure of the applicability of our model.

4 Interpretation & Critisism

In the preliminary model, we treat multilevel categorical sentiment representation of the tweets as our response and the local cumulative number of local deaths and infections as our covariates. Multilevel categorical sentiment representation of the tweets, extracted from the original texts, reflects the attitude and mood of the publishers under the public health event. The daily updated number of local deaths and infections seems to be the most accessible and direct information for people to learn about the COVID-19 condition. It is reasonable to view the number of local deaths and infections as the input and the extracted information from the tweets as the output, as the publisher, to build the model.

We concede that there are some limitations to our model. The first one is technical, namely the validation of our extracted information. Since the sentiment analysis is not very mature, we doubt the accuracy of the results. Besides, the local policy, medical resources, supply of masks also impact their mood, but we don't consider these in our simple model.

[1] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." *Proceedings of the workshop on language in social media (LSM 2011)*. 2011.

5 References

1. Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." *Proceedings of the workshop on language in social media (LSM 2011)*. 2011.