Predicting COVID-19 Cases Leveraging Sentiment Analysis From Twitter Data

University of Toronto
MIE1516: Structural Learning and Inference
Final Project Proposal
Armando Ordorica

March 2019

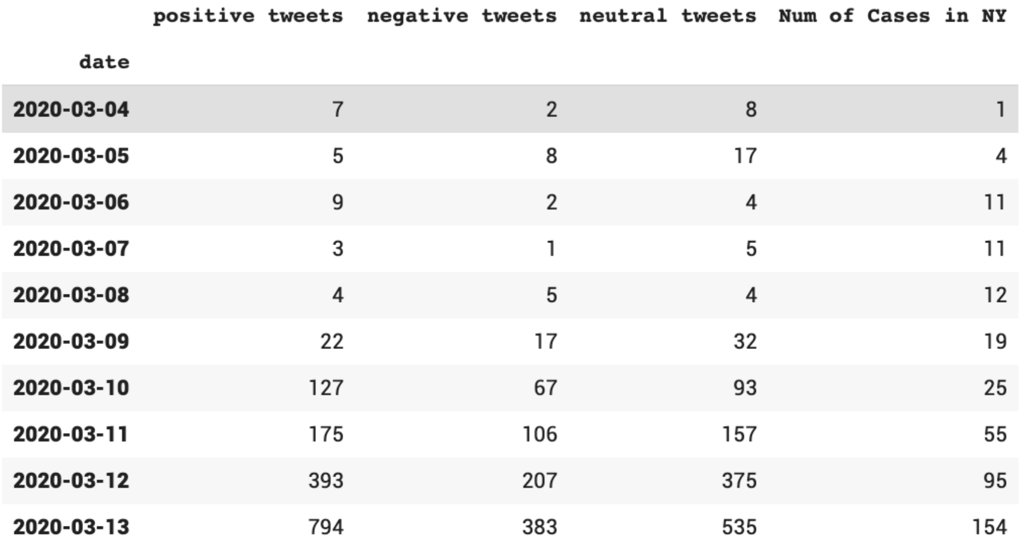
Twitter has become the main social platform for major influential actors to share their emotions in real time. Given that there is an inherent lag between the time of infection and the time of detection of a positive case of COVID-19, we hypothesize that Twitter data related to COVID-19 can be indicative of the total number of cases in a specific region. As shown on the figure below, the number of Tweets related to COVID-19 have increased exponentially with similar rates to daily confirmed cases.

This analysis will focus on New York City area only and aims to answer the question of whether people's behaviour on Twitter can improve the prediction of number of daily cases in NYC. The end objective is to establish a framework that could help health officials and policy makers in predicting the upcoming number of cases in a given region.

First, an autoregressive model such as ARIMA will be used to forecast number of daily cases in New York City. Second, time series to show the evolution of Tweets regarding COVID-19 will be generated. Finally, methods will be explored to see whether Twitter sentiment analysis data is valuable in helping to predict number of daily cases of COVID-10 in NYC area.

References

[1] Kaggle Covid-19 Twitter dataset *Coronavirus (covid19) Tweets*



	positive tweets	negative tweets	neutral tweets	Num of Cases in NY
date				
2020-03-04	7	2	8	1
2020-03-05	5	8	17	4
2020-03-06	9	2	4	11
2020-03-07	3	1	5	11
2020-03-08	4	5	4	12
2020-03-09	22	17	32	19
2020-03-10	127	67	93	25
2020-03-11	175	106	157	55
2020-03-12	393	207	375	95
2020-03-13	794	383	535	154

Figure 1: Comparison of number of daily COVID-19 cases with number of COVID-19 related Tweets by sentiment in the New York City Area