Sentiment Analysis of Tweets to Predict Covid Cases

Report 1

EECS 4080
June 1, 2022
Taswar K. & *Hassan K.*

Table of Contents

Description

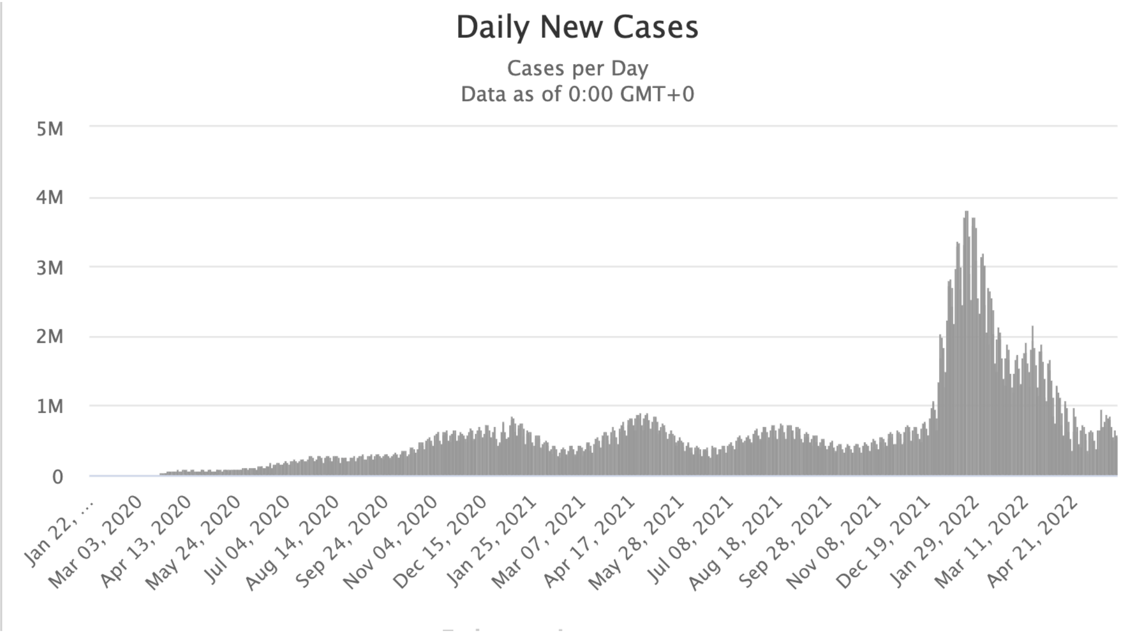
Our report summarises the findings on the provided literature reviews and mentions a timeline which states what duration of data (tweets) is collected from using Twitter. It also provides a summary of the disease we are choosing and the methods that will be applied to predict an outbreak of that disease

Abstract

This report will discuss our findings in the literature review and data collection process. The data required for our project are tweets from users related to covid-19 and its symptoms. We know that covid-19 was first reported in Wuhan, China on the 31st of December 2019. However, it was soon declared a “global emergency” on 30th January, 2020 by the World Health Organisation (WHO). Tweets were collected during various stages of covid-19 spikes ranging from **January 2020-September 2021.**This collection of tweets will be used for testing data once our model is done training. For training data, we collected a relatively large collection of tweets, each labelled with a sentiment of positive or negative. We also collected another dataset of labelled tweets and extracted the tweets labelled as neutral to give more balance to our training data. A wave of time-series analysis is conducted to further classify our data and reach a conclusion. A breakdown of timeline is shown below

Timeline

Our dataset corresponds prior to WHO declaring covid-19 as a “global emergency”. In our analysis, we consider spikes, rates of change and deviations (shown on fig.1) as well as some global facts such as the beginning of covid-19 and its trends. Our analysis is limited to a time frame of 18 months.



This graph was provided by: https://www.worldometers.info/coronavirus/

Data Collection

Kaggle

The data we will be using will come from pre collected tweets and sentiment in csv files. The first is Covid-19 tweets with the sentiment provided (url). It contains tweets 5k training tweets and 2.4k validation tweets. The tweets are matched with the sentiment ranging from "Optimistic (0), Thankful (1), Empathetic (2), Pessimistic (3), Anxious (4), Sad (5), Annoyed (6), Denial (7), Surprise (8), Official report (9), Joking (10).". The second data set we are using is from openicpsr (url). It contains tweets from 28th on January 2020 to 1st January 2021. The third data set is also from openicpsr (url) which contain tweets and sentiment from 28th of January 2020 to September 1st of 2021.

Methods

Based on provided literature reviews we have chosen to adapt some of the common useful methods to extract and extrapolate tweets data in order to predict a covid-19 outbreak. We have seen that most previous methods used to detect disease outbreaks using twitter were deployed via a pipelined-architecture which usually involved steps including but not limited to *Initial Selections, Duplicate removals* and various *classification* techniques applied to a monitoring algorithm that computes the expected duration between times of posting for consecutive tweets.These time-between events corresponds to the duration between consecutive events in a time-series. Furthermore, in addition to the methods described we have decided to approach these methods in a unique way by introducing a self-identification classification which helps us identify if a certain tweet is self-reported or just a fact-check.

To predict the date of an outbreak or significant spike in covid-19 we will first collect covid cases on the time periods we have selected with OpenCovid API. We will then match those covid cases with tweets from the same time frames containing information about covid. Those tweets will range from self reporting, questions, data about covid, and more. Then we will clean the tweets with data preprocessing by removing parts of each tweet that do not give us information about sentiment (i.e hyperlinks, non english characters,) We will use BERT and S-BERT to embed those tweets and then use multiple classifiers (MLP, KNN, ..) to train and label the sentiment of the tweets. Finally after storing the tweets, their sentiment, and other labels containing runny nose, fever, and so on. The information of the tweets will be stored on a day to day basis. Then we will use time series analysis to predict the date we believe to expect an outbreak or a significant spike in covid-19 in North America.