MSc Project Report

Custom NER Model For

Pandemic Outbreak Surveillance Using Twitter

Hansinie Madushika Jayathilake

2021

A report submitted as part of the requirements for the degree of MSc in Big Data Analytics at Robert Gordon University, Aberdeen, Scotland

Abstract

Pandemic outbreaks have become the most significant discussion all over the world at the time of writing this report. Pandemics can cause serious public health consequences if no action was taken in advance, since outbreaks are out of control once stepped out. Hence, getting early awareness about outbreaks is essential. Therefore, outbreak surveillance is highly important. Nowadays people are used to rely more on social media for receiving and reporting urgent public health matters especially pandemics outbreaks. Twitter is playing a huge role in this regard. Therefore, this study was conducted to develop a pandemic outbreak surveillance system with aid of natural language processing and deep learning techniques by using Twitter data.

In this study Twitter data was extracted and applied text preprocessing techniques for model training. Training and test data were annotated with label “Pandemic” by a custom annotator written by the author. This study trained two custom named entity recognition (NER) models to predict pandemic named entities using Python spaCy library which is based on deep neural network architecture. These models were trained with two optimizers such as Stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (Adam) algorithms. Confusion matrix was used to evaluate the model performance. Based on the results, Adam optimizer based NER model with 0.9691 precision was selected as the best model for the study. Using python web framework FastAPI, the model was deployed for a web-based surveillance dashboard to visualize the country wise results in map and chart views based on user input parameters.

Keywords: Pandemic, Outbreak Surveillance, Named Entity Recognition model, Deep learning, Optimizers, neural network, Confusion Matrix, Precision, Twitter, Stochastic Gradient Descent, Adam