Characterizing Anti-Asian Rhetoric During The COVID-19 Pandemic: A Sentiment Analysis Case Study on Twitter

Authors : Ramya Tekumalla, Zia Baig, Michelle Pan, Luis Alberto Robles Hernandez, Michael Wang, Juan M. Banda

This repository contains all the deliverables for the paper titled " Characterizing Anti-Asian Rhetoric During The COVID-19 Pandemic: A Sentiment Analysis Case Study on Twitter " which was accepted as a workshop paper in ICWSM 2022. This repository contains the following folders:

1) Data - This directory contains all the data used in this paper. To hydrate the tweets, you will require a Twitter developer account and obtain the keys from the developer account. After installing twarc (a python package), you can hydrate the tweet ids using the following command

   twarc hydrate tweet_id.txt > hydrated_tweet.json

   This command will hydrate the tweet id to a json format tweet which includes all the fields. Further if you would like to manipulate and reduce the steps to reproduce, you can use our Social Media Mining toolkit (https://github.com/thepanacealab/SMMT) which will help with transforming the tweet ids / data to required fields

2) Model - We include the best model from the paper to reproduce the results and also further use it for several downstream tasks such as obtaining predictions for an unseen set. We include fine tuned Covid-Twitter-BERT which can be loaded into the code included and use it for downstream tasks. Here are the steps to download and load the model.
   a) Download covidbert.tar.gz from the repository and untar the gz file using the following command
      tar -xf covidbert.tar.gz
   b) Store the model and all the files in a directory.
   c) Add the path of the directory to the code and run the code with the following command

      python3 predict_trans.py

3) **Libraries required**
   a) Simple Transformers == 0.61.13
   b) Twarc == 1.10.0
   c) Scikit Learn == 1.0.2
   d) Spacy == 2.3.0