Taif Alwafi

Abstract

This project aimed to use classification models to predict how a person feelsabout the coronavirus. Impacts of the pandemic, such as economic strife, andunprecedented restrictions on social contact have threatened people's mentalhealth. By tweets data provided from Kaggle with a logistic regression model.the clinics will be able to identify those needing help and treatment by classifyingtheir tweets.

Design

This project was developed during the online data science Bootcamp at SDAIA.I obtained the data from Kaggle, and it presents tweets about Coronaviruswhich contain words such as Corona-19, Coronavirus, etc. I used "SentimentIn-tensityAnalyzer" to be labeling data.Machine learning techniques are useful inunderstanding the sentiment of the people about a this virus. In this project Itrained a model to classify tweets as either positive or negative, To help peopledetect their sentiment about the pandemic and contacting with a clinic to avoidrisk mental health.

Data

The dataset contains 179108 tweets and 13 features : (user name, user location,user description, user created, user followers, user friends, user favourites, userverified, date, text, hashtags, source, is retweet). Although I used one featuretext and I added Label column contain the tweet sentiment.

Algorithms

Feature Engineering

- Prepossessing the text feature using the following NLP techniques : Convert-ing to lowercase, removing text in square brackets, removing links, Removingpunctuation, removing words containing numbers, removing stop words, and Lemmatization.
- labeling text feature using SentimentIntensityAnalyzer 0 fornegative and 1 for positive.
-

Models : Logistic regression,Support vector machine, Bernoulli naive bayes, and neuralnetwork classifiers were used and after training the logistics regression modelgot the highest accuracy.

Model Evaluation and Selection : All models were trained on 80/20 train vs test and based on the result of thisexperiment the logistics regression model was the best model according to accu-racy metric. after choosing the logistics regression I applied Repeated StratifiedKFold with 5-fold and gridsearch in order to tune the model. The final result :

The score for	logistic regression	svn	BernoulliNB	Neural network
Training	99.03%	96.21%	91.18%	91.23%
Test set	94.36%	94.17%	84.76%	84.40%

Tools

Pandas for data manipulation , Scikit-learn for modeling , re for clean data , tensorflow and keras for neural network model nltk for natural language processing and Matplotlib, Seaborn and plotly for plotting