COVID-19_Vaccines_Tweets

Manal Nawar Allahyani.

Abstract

This project aimed to understand people's emotions about the covid-19 vaccines by analyzing the tweets using machine learning models to help the safety of society, and know the most covid-19 vaccine that is comfortable for the people. The used data in this project is provided by Kaggle, the data is labeled using Sentiment Intensity Analyzer, with sklearn library random forest was trained and get 97% accuracy. The streamlit is used to build an interactive dashboard to visualize and communicate the final results.

Design

This project is one of the T5 Data Science BootCamp requirements. Data provided by Kaggle has been used in this project. The COVID-19 vaccines included in the data are Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V. Classifying how people feel positive/negative to covid-19 vaccines using machine learning algorithms would enable understanding most comfortable vaccine for people

Data

The dataset is provided in `.csv` format. It contains 19,3272 tweets, each tweet has 15 features. The most relevant feature to this project is the text which contains the tweet text. Some other features are extracted from other features such as country name is extracted from location, where it contains the user location at the time of the Tweet. Another important feature of this project is the label of the tweet where it extracted from tweet text using Sentiment Analysis tools.

Algorithms

Feature Engineering

- Cleaning the text feature by converting the chars to lower case, removing all non-chars, removing stop words, and applying lemmatization.

- Labeling the text using the `SentimentIntensityAnalyzer` and `Blobtext` to 0 for negative text and 1 for positive text
- Determining vaccine type of each text using words filters (this is not accurate because some texts are for more than one type)

Models

Naive bayes, and random forest were used to classify the tweets. The naive bayes get the higher accuracy with `Blobtext` labeling.

Model Evaluation and Selection

Naive Bayes, and random forest were used to classify the tweets. The naive Bayes get the higher accuracy with Blob text labeling. The models were trained on a 25/75 test vs. train. Each model train twice on sentiment analysis labels and bold text labels. The official metric was the accuracy of the model, where the model tested on the accuracy, precision, recall, and F1 score. The result of used model:

- Accuracy: 97%
- Precision: 95%
- Recall: 96%
- F1: 95%

Tools

- Pandas for data manipulation
- Scikit-learn for modeling
- re for clean data
- nltk for natural language processing
- Matplotlib for plotting
- streamlit for interactive visualizations

Communication

The slides are provided here, besides details are provided at the readme of the project. Feel free to any pull requests.