Sentiment Analysis: Predicting sentiment of Covid 19 Tweets Dataset

Hritik Sharma Data science trainees, AlmaBetter, Bangalore.

Abstract:

Pandemics are a severe threat to lives in the universe and our universe encounters several pandemics till now. COVID-19 is one of them, which is a viral infectious disease that increased morbidity and mortality worldwide. This has a negative impact on countries' economies, as well as social and political concerns throughout the world. The growths of social media have witnessed much pandemic-related news and are shared by many groups of people. This social media news was also helpful to analyze the effects of the pandemic clearly. Twitter is one of the social media networks where people shared COVID-19 related news in a wider range. Meanwhile, several approaches have been proposed to analyze the COVID-19 related sentimental analysis. To enhance the accuracy of sentimental analysis, we have proposed a novel approach known as Sentimental Analysis of Twitter social media Data (SATD). Our proposed method is based on five different machine learning models such as Logistic Regression, Random Forest Classifier, Multinomial NB Classifier, Support Vector Machine.

Problem Statement:

This challenge asks you to build a classification model to predict the sentiment of COVID-19 tweets.The tweets have been pulled from Twitter and manual tagging has been done then.

The names and usernames have been given codes to avoid any privacy concerns.

You are given the following information:

1. Location
2. Tweet At
3. Original Tweet
4. Label

Introduction:

Coronavirus disease 2019 (COVID-19) was discovered in Wuhan, China after the virus had spread globally. It was announced by the World Health

Organisation to be a pandemic. This epidemic is now affecting a large number of individuals all across the world. At present, COVID-19 is a serious threat to human life all over the world, where several individuals developed symptoms such as pneumonia . It has a wide variety of effects on the human body, including extreme respiratory syndrome and multi-organ failure, which will potentially lead to death within a short period of time . When the globe has been fighting for in recent months and most people have been imprisoned, Twitter has become more important than ever. Even in the past, people have been using Twitter to communicate, express, and spread information relevant to the disaster, whether it be cyclones, ebola, flooding, or Zika . Twitter has been one of the platforms for millions to express their emotions regarding different issues.

MACHINE LEARNING ALGORITHMS USED :

In this research work, four different algorithms (classifiers) were used **1. Support Vector Machine (SVM)** - It is a supervised learning model with associated learning algorithms that gives analysis of data for classification. It represents the data as points in space. The classification into individual groups is achieved by discovering the best hyperplane that distinguishes the two classes in the optimal approach. Support Vector Machine separates positively labeled examples from the negatively labeled ones by finding the "hyperplane" that maximizes margin between the two classes which can be achieved by solving quadratic objective function Where b is the intercept and bias term of hyperplane equation.

2. Random Forest – Random Forest algorithm does the selection of observation and feature randomly in order to build several decision trees and then computes the average of the results. Random Forest algorithm creates random subset of the features and build smaller trees using the subset created. Furthermore, Random Forest produces high accuracy through cross validation, handles missing values and maintains the accuracy of large proportion of data. Random Forest classifiers don't allow over-fitting trees into the model in case there are no more trees

3. Naïve Bayes Classifier- The purpose of using a Naïve Bayes Classifier is to predict the likelihood that an event will occur with the assistance of evidence that is present in the data. A multinomial Naïve Bayes algorithm classifier was used because it is suitable and more efficient for features that describe discrete frequency counts which is similar to the features of the data present in the dataset obtained

4. K-Nearest Neighbor (KNN) – is a parametric method that is used for classification. An object is classified by plurality vote of its neighbor with the object being assigned to the class most common among its K nearest neighbors. A commonly used distance metric for continuous variables is Euclidean distance

Stochastic Gradient Descent - The word 'stochastic' means a system or process

linked with a random probability. Hence, in Stochastic Gradient Descent, a few samples are selected randomly instead of the whole data set for each iteration. In Gradient Descent, there is a term called “batch” which denotes the total number of samples from a dataset that is used for calculating the gradient for each iteration. In typical Gradient Descent optimization, like Batch Gradient Descent, the batch is taken to be the whole dataset.

XGBoost- is an implementation of Gradient Boosted decision trees. This library was written in C++. It is a type of Software library that was designed basically to improve speed and model performance. It has recently been dominating in applied machine learning. XGBoost models majorly dominate in many Kaggle Competitions. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and the variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

.

Observation

This section of the study presents the results or three keywords: “COVID,” “CORONAVIRUS,” and “COVID-19,” as well as emotive analysis on Twitter comments.It also includes a discussion of the implications of these findings. A Sentimental Analysis of the Keyword “COVID” is presented in Section A. We created a Word Cloud to represent the number of times the word “COVID” appeared in a Twitter dataset, as shown in . As a result, we can see the frequency of the keyword “COVID” by looking at the Word Cloud. Following that, we calculated the relationship between subjectivity and polarity for the same and displayed it with a scatter plot in When it comes to presenting the values for Cartesian coordinates are utilized for two constants in the data set In this case, subjectivity determines whether a word is subjective or objective. Polarity, on the other hand, teaches us about a person's positive and negative responses to a given keyword or phrase. The zero point is represented by the point “zero” in the polarity column As a result, everything to the left of zero denotes negative feedback, whereas everything to the right of zero denotes positive feedback. The percentage of neural tweets is larger than the percentage of positive and negative tweets in the emotional analysis for all three terms, which is not surprising. Even in these circumstances, People are maintaining positive as well as neutral attitudes in the face of chaotic illness spread scenarios as evidenced by the larger percentage of positive tweets compared to negative tweets.

Conclusion

Throughout the text, the importance of social network analysis is discussed. Twitter is a popular social media platform where people can express themselves and share their thoughts. This study employed over 370 tweets from Twitter to do emotional analysis for three key phrases connected to the COVID-19 pandemic (COVID, CORONA VIRUS, and COVID – 19).” The results were presented in this research. Positive tweets account for approximately 31% of total tweets, whereas negative tweets account for approximately 19% of total tweets. This means that half of all neutral tweets on Twitter, or half of all tweets based on these respective terms, are neutral in their attitudes. In the COVID situation, neutral sentiments outweighed both positive and non-positive sentiments, according to the polarity analysis.