

## Abstract

This project aimed to use classification models to predict how a person feels about the coronavirus. Impacts of the pandemic, such as economic strife, and unprecedented restrictions on social contact have threatened people's mental health. By tweets data provided from Kaggle with a logistic regression model, the clinics will be able to identify those needing help and treatment by classifying their 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 Coronavirus which contain words such as Corona-19, Coronavirus, etc. I used "SentimentIntensityAnalyzer" to be labeling data. Machine learning techniques are useful in understanding the sentiment of the people about this virus. In this project I trained a model to classify tweets as either positive or negative, To help people detect their sentiment about the pandemic and contacting with a clinic to avoid risk 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, user verified, date, text, hashtags, source, is retweet). Although I used one feature text and I added Label column contain the tweet sentiment.

## Algorithms

### Feature Engineering

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

**Models** : Logistic regression, Support vector machine, Bernoulli naive bayes, and neural network classifiers were used and after training the logistics regression model got the highest accuracy.

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

The score for	logistic regression	svm	BernoulliNB	Neural network
Training	99.03%	96.22%	91.18%	86.68%
Test set	94.36%	93.96%	84.76%	86.15%

## 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