VERZEO

MAJOR PROJECT

- **Project Name:**

Machine Learning March Major Project

- **Project Description:**

Problem statement: Create a classification model to predict the sentiment either (Positive or Negative) based on Covid Tweets

Context: 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.

Dataset:

https://drive.google.com/file/d/16UmG2L6RkaDoynNLAlw7aTzqru2djHCq/view?usp=sharing

Details of features:

The columns are described as follows:

1) UserName: UserName in encrypted numbers
2) ScreenName: ScreenName in encrypted numbers
3) Location: Country from where tweet was pulled from
4) TweetAt: Twee time
5) OriginalTweet: Tweet content
6) Sentiment: Positive, Negative, Neutral, Extremely Positive, Extremely Negative

Steps to consider:

1. Read the dataset with encoding parameter set to ‘latin1’
2. Remove handle null values (if any).
3. Preprocess the Covid tweets based on the following parameter:

a) Tokenizing words

b) Convert words to lower case

c) Removing Punctuations

d) Removing Stop words

e) Stemming or lemmatizing the words

4)Convert the 'Extremely Positive' and 'Extremely Negative' Sentiments to 'Positive' and 'Negative' sentiments respectively

5)Transform the words into vectors using

a)Count Vectorizer

OR

b)TF-IDF Vectorizer

6)Split data into training and test data.

7)Apply the following models on the training dataset and generate the predicted value for the test dataset

a) Multinomial Naïve Bayes Classification

b) RandomForest Classification

c) KNN Classification

8)Predict the Sentiment for test data

9)Compute Confusion matrix and classification report for each of these models

10)Report the model with the best accuracy.