**Approach:**

* Given dataset containting two files one is to train the model another is to test by predicting using trained model.
* Dataset contains many columns. i fetch those columns which contains tweets and sentiments (given in terms of numerical values).
* So tweets are independent variable and sentiments are dependent variable
* The given tweet datasets are comprised of very much unstructured tweets which should be preprocessed to make an NLP model. In this project, I tried out the following techniques of preprocessing the raw data. But the preprocessing techniques is not limited.
* Removal of punctuations.
* Removal of commonly used words (stopwords).
* Normalization of words(Stemming).
* After cleaning the tweets, as we can't use text directly to train the model so I used feature extraction technique Bag Of Words Model to extract features from texts.
* After this I found that the data is having a very high dimension. So I want to reduce the dimension of it using Dimensionality reduction technique KernelPCA
* After that I used Random Forest Classification Model. I tried other models as well but this gives me better result than other model. I used hyperparameter tuning as well to optimize the result and used the best one.
* Test set data was also not clean and needed preprocessing so I applied the same technique to clean it and tranformed it in the format which is needed to input into the model
* Then after getting pricted results, I converted the prediction into the submision format.
* Then then saved the data into csv file.

**Tools Used:**

* Used pandas and numpy for preprocessing.
* Used re and NLTK - Stopwods, PorterStemmer for Cleaning Data.
* Used Scikit-learn - CountVectorizer for feature extraction.
* Used Scikit-learn - KernelPCA for Dimensionality Reduction.
* Used Scikit-learn - RandomForestClassifier as Machine learning model