LIBRARIES:

1) PANDAS: pandas is an open source python package that is most widely used for data science/data analysis
2) NUMPY: NumPy is a Python library used for working with arrays. Its full form is numerical python and used to work with numerical python.
3) RE: Its full form is regular expression. A regular expression (or RE) specifies a set of strings that matches it
4) NLTK: The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP)
5) MATPLOTLIB: Matplotlib is a data visualization and graphical plotting library for Python and its numerical extension NumPy
6) SEABORN: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
7) SKLEARN: The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction. Scikit-learn is probably the most useful library for machine learning in Python.
8) WORDCLOUD: Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.
9) GENSIM: Gensim is designed to handle large text collections using data streaming and incremental online algorithms
10) TQDM: tqdm is a Python library for adding progress bar. It lets you configure and display a progress bar with metrics you want to track.
11) TAGGED DOCUMENTS: A single document, made up of `words` (a list of unicode string tokens) and `tags` (a list of tokens). ... Replaces "sentence as a list of words" from Word2Vec.

SKLEARN ML LIBRARY FUNCTIONS:

1) TfidVectorizer: It is a feature engineering function , Tfid will systematically compute word counts using CountVectorizer and then compute the Inverse Document Frequency (IDF) values and only then compute the TF-IDF scores

2) TRAIN_TEST_SPLIT: train_test_split is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data.

3) LOGISTIC REGRESSION: Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

4) SVC: (support vector classifier) SVC is a nonparametric clustering algorithm that does not make any assumption on the number or shape of the clusters in the data.

5) LINEAR SVC: The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data.

6) COUNTVECTORIZER: CountVectorizer is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.

7) STANDARD SCALAR: StandardScaler will transform our data such that its distribution will have a mean value 0 and **standard** deviation of 1. In case of multivariate data, this is done feature-wise (in other words independently for each column of the data).

8) RANDOM FOREST CLASSIFIER: **random forest classifier** builds multiple **decision** trees and merges them together to get a more accurate and stable prediction

9) CONFUSION MATRIX: A **confusion matrix** is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset.

10) F1 SCORE : **F1-score**, is a measure of a model's accuracy on a dataset.