

***Major Project***

(Documentation)

**Project Objective**:

To build a machine learning model that can analyse and predict whether a given *review* is “**Positive**” or “**Negative**” using “*count vectorizer*” and “*Support Vector Machine (SVM)*”.

**Resources**:

A tab separated value (.tsv) file named Restaurant\_Reviews.tsv, Pandas, *count vectorizer* and support vector machine (svm), *support vector classifier* (SVC) from ‘*sklearn*’, ‘*joblib*’ and ‘*pipeline*’ packages.

**Approach**:

* With Pandas, read ‘Restaurant\_Reviews.tsv’ file from the location it was stored in.
* The file has 2 columns named as ‘Review’ which contains the reviews given by different customers and ‘Liked’ which determines the nature of the reviews as 0s and 1s (0 – Negative and 1 - Positive).
* Using pandas utilities (.replace()), 0s and 1s are replaced with ‘Negative’ and ‘Positive’ respectively.
* Initially, the data was split into train set and test set as ‘x\_train and y\_train’, and ‘x\_test and y\_test’.
* ‘Count vectorizer’ is used to return the count of the words in the reviews in the form of an array. It returns a 1D array for every review.

* Support vector classifier (SVC()) takes the transformed x\_train and initial y\_train data to train the model. SVC is used to construct a hyperplane to classify the positive reviews and negative reviews.
* Before this, sklearn’s ‘*pipeline*’ is used to pipeline countVectorizer and SVC and resultant is assigned to an identifier (“sentiment\_analyzer” in the program) which in further process used as the model to predict.
* Using the obtained model object, ‘x\_test’ can be predicted.
* ‘*joblib*’ package has been used to dump the model as a file in the local machine and save it.

**Result**:

The model built with countVectorizer() and SVC() to predict the nature of restaurant reviews has obtained an accuracy of over *76%*.

Random reviews and few statements from social media have been successfully classified and predicted by the model as ‘Positive’ or ‘Negative’.