Project Report - COMP9313

(Swati Bharti – z5277828)

1.Stacking Model:

Stacking is an ensemble method used in machine learning to improve the results by combining several regression and classification models. The stacking model learns how to combine predictions from multiple base models. The model formed by combining output from multiple base models is called **meta-model** and is used for predictions on test set.

Task 1.1: Build a preprocessing pipeline

Preprocessing of the input is done to clean the data and to make it easy for analysis.

The reviews are fed as input, a bag of words is built. From the bag of words, the feature is extracted and feature vector built which contains the words(feature) and the frequency of the word. The categories are converted to numerical values between 0 to 2 to represent different categories (FOOD-0, PAS-1, MISC-2).

tokenizer converts the input to lowercase and removes all the white spaces from the input and produces a bag of words. **countVectorizer** extracts unique words and calculates its frequency which is used as features. **Stringindexer** maps categorical variables to numerical values and has a range [0, number of labels).

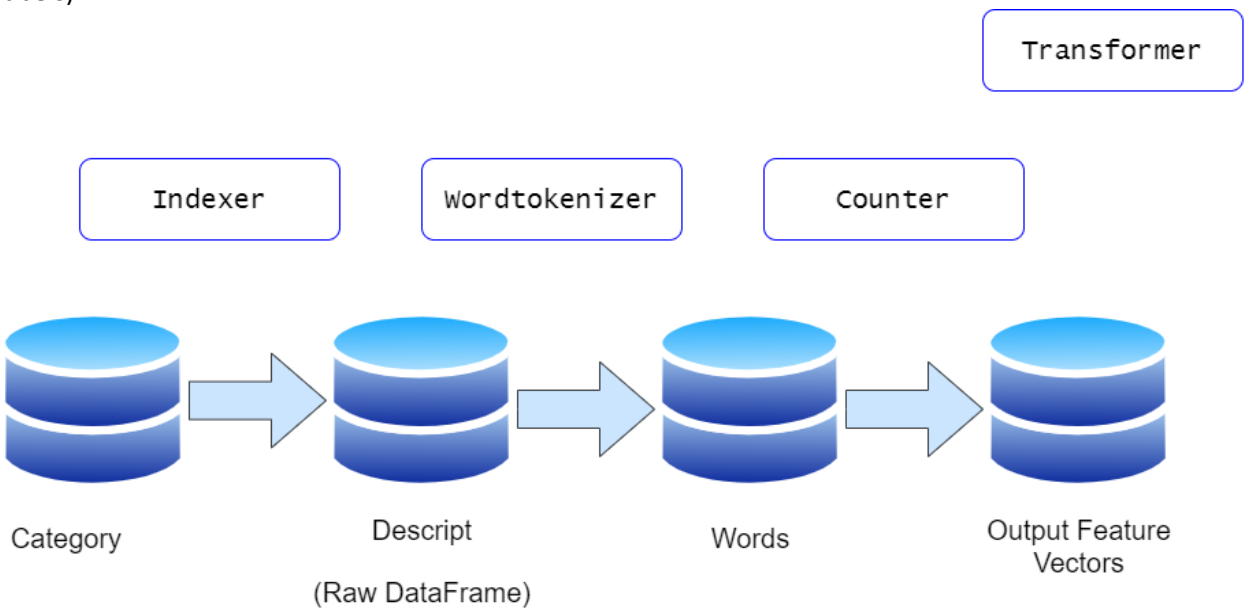


Fig 1.1.1: Preprocessing Pipeline

Task 1.2: Generate Meta Features for Training

The training data is divided into 5 different groups, using k- fold cross validation then each group is trained using two classifiers i.e. SVM and Naïve Bayes for each class label separately and the model is built using **fit** which is used to train the classifier and for this both the features and the category labels are provided. The model is then tested on the first group to get the prediction value using **transform**. The final prediction

value of the model is the **union** of the result of the prediction value of the first test group with the other test groups. The joint prediction value for the model is predicted by using the values of the prediction of both the SVM and Naïve Bayes Classifier. Using both the joint prediction and the prediction result from each base classifier the meta-features are generated.

Task 1.3: Obtain the prediction for the test data

The test dataset is passed through the preprocessing pipeline model, the generate_base_feature _pipeline model which is used to generate predictions from the base classifier. The predictions are used to build the meta-features, which is fed into the **meta_classifier** which uses Logistic Regression classification to predict the category labels.

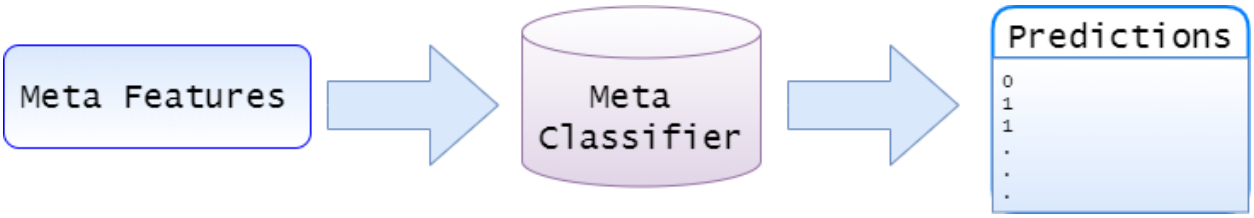
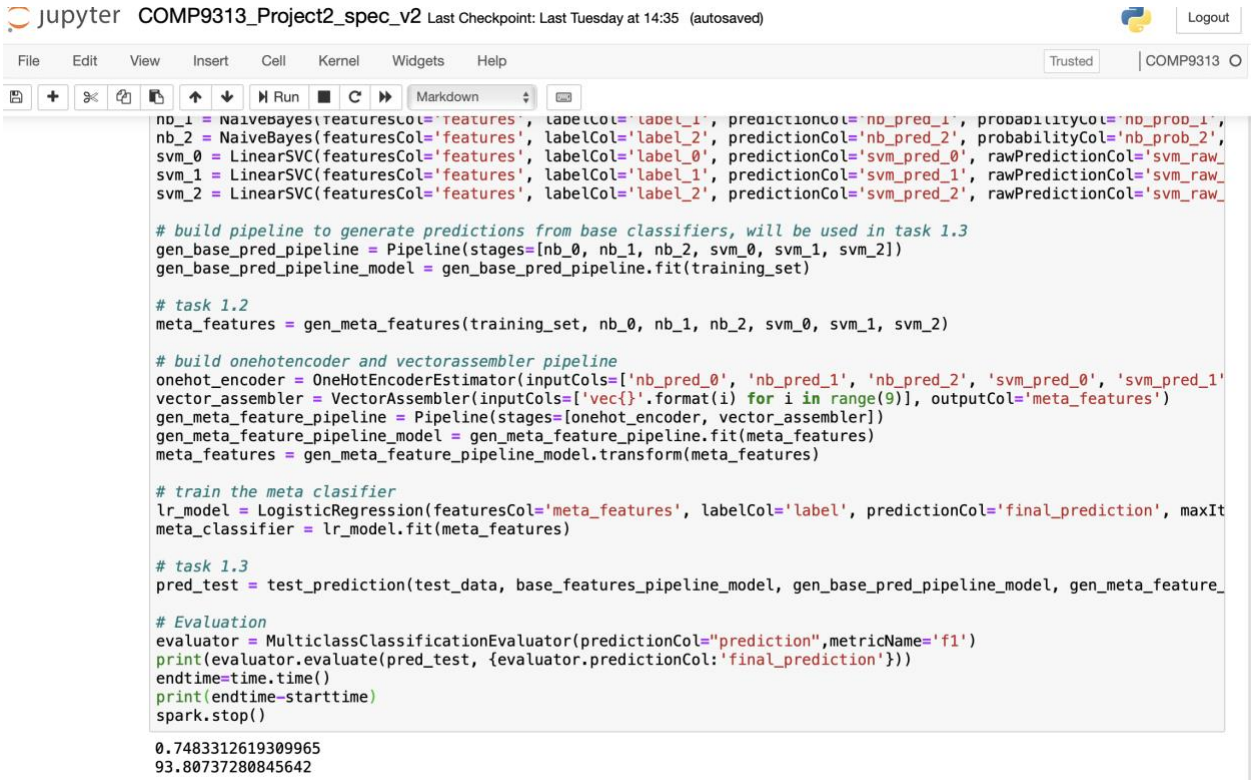


Fig1.3.1 The prediction model

2. Results:

Fig : Prediction of the result



3.Conclusion:

The model is tuned using **k fold cross validation** which selects a set of best parameters to fit the model. A **MulticlassClassificationEvaluator** is used to evaluate the model. The evaluator uses the **f1 score** as a metric for evaluation. The model built above, has an f1 score of **0.74833126** and the runtime is **93.807372** msec/sec