Machine Learning Project

Verzeo Major Project Batch : ML063B2

Topic: Ensemble three classification algorithm and find which one is more accurate in predicting the result for the given dataset

Dataset : Tweet dataset

Library used: Pandas, Sklearn, NumPy, Matplotlib and Seaborn

Firstly , the dataset is imported to the jupyter notebook with help of pandas library. Initial stage is cleaning the dataset. We use inbuilt function call df.info() to find any missing values then we clean the dataset by removing all the unwanted values. The dataset is then divided into three - male, female, df as three datasets.

The “male” and “female” dataset are divided based on the gender. And the main dataset“df”contains both male and female gender and we ignore the other two gender categories such as “brand” and “unknown” because compared to other gender it is very low.

We did some feature selection for the main dataset “df” and did some exploratory data analysis. Using seaborn libraries, we plot correlation of dataset and the dataset which does not show good correlation. so we expect the machine learning algorithm will be 50- 55% accurate.

Male and female dataset is used to find some interesting facts from the dataset like common words and average words used by both genders. With the help of library and python programming language we found the answer such as -

Most common words used by both genders are:“and”, “the”, “to”.

Average number of words used by male is : 15

Average number of words used by female is : 16

We chose three classification algorithms :

1. KNN algorithm
2. Logistic Regression
3. Random forest

These are the algorithms we used. Before training we did feature selection. With the help of sklearn library it is easy to fit, train and test each algorithm.

Dependent variable is: gender (taken along Y axis)

Independent variables are: trusted judgements, tweet count, retweet count, fav number, profile confidence, gender confidence (taken along X axis)

Split the dataset into X\_train, X\_test, Y\_train, Y\_test. Then train each algorithm and predict each algorithm accuracy.

From the accuracy calculation , as we expected the algorithm accuracy lies somewhere around 50-55%.

Accuracy of algorithm :

1. KNN algorithm : 53%

2. Logistic Regression : 54%

3. RANDOM FOREST : 54%

So, the best fitting algorithm for the dataset is random forest algorithm and Logistic regression.