Major Project July ML Batch 1

Name: Sanjay Marreddi

Google Classroom Code : ML07B1

sanjay.marreddi.19041@iitgoa.ac.in

Brief Summary of Project

Overview of Steps :

- Importing the Modules, Libraries, and Datasets
- Exploratory Data Analysis
- Data Cleaning
- Feature Selection and Feature Engineering
- Ensemble Machine Learning Modelling
  - Classification using Naïve Bayes Algorithm
  - Classification using Support Vector Machine Algorithm
  - Classification using Logistic Regression Algorithm
- Accuracy Comparison
- Asking and Solving Questions on Dataset.

Brief Description of steps

Importing : All the required libraries including *seaborn* for Visualisation, *Collections* for Counting , *ntlk* for Text Pre-processing are imported. Moreover, all the Algorithms, tools, metrics from *sklearn* are also imported.

EDA: The given dataset is analysed very clearly using the various attributes available in *pandas* for Data frames. Columns with *NaN* values are observed. Data Visualisation is also done using *seaborn* including heatmap for Correlation matrix.

Data Cleaning: Using stemmer from *nltk* and *re* few columns names *description* and *text* which has text values are cleaned and separate columns are created in the data frame. Rows with NaN values are dropped.

Feature Selection & Feature Engineering: Only few columns which are of high importance for training are selected as independent variables. Then only those rows with *gender:confidence* equal to *1* are chosen. Then the dependent variable column *gender* is label encoded.

Ensemble Machine Learning Modelling : Three Classification algorithms using *sklearn* library are implemented for predicting the *gender* based on the features chosen.

The three algorithms used are Naïve Bayes, Support Vector Machine and Logistic Regression. The common steps in all these three algorithms include :

- ✓ Splitting the dataset into training and testing sets.
- ✓ Using the *LabelEncoder* to fit and transform the features
- ✓ Then using the *Tfidf Vectorizer* for vector transforming the training and testing features.
- ✓ Then fitting the data followed by prediction.

Coming to specific Algorithms:

- In Naïve Bayes I used only default parameters during fitting.
- In Support Vector Machine, I used *linear* kernel, *degree* value as 2 and *C* as 2.
- In Logistic Regression , I used default parameters for everything except *max_iter* which was set to 1000.

Accuracy Comparison:

By using the *accuracy_score* metric from the *sklearn* library, accuracy is obtained in all the three cases.

1. Naive Bayes Algorithm is **71.42857142857143 %**
2. Support Vector Machine Algorithm is **68.0672268907563 %**
3. Logistic Regression Algorithm is **70.58823529411765 %**

So, For the given data set and the used hyperparameters, based on the *accuracy score* , we can say **Naive Bias Algorithm is best Classification Algorithm** in this case.

Questions on the dataset :

1. What are the most common emotions/words used by Males and Females?

Using the *Counter()* from *Collections* library, I counted the words in the *cleaned_text* column of the dataset and then sorted them in Descending order.

- The Most Common WORD used by **Males** is "I" which was used 1460 times
- The most Common EMOTION used by **Males** is "love" which was used 158 times
- The Most Common WORD used by **Females** is "I" which was used 2311 times
- The most Common EMOTION used by **Females** is "love" which was used 275 times

2. Which Gender has a greater number of tweet count ? On an average what is the tweet_count by each gender ?

I looped over all the rows and using the dictionaries I stored the *tweet_count* for each gender separately and then divided with number of users per gender.

- ✓ Gender "**brand**" has a greater number of tweet count

- ✓ The **average** number of tweets per Gender is as follows :-
  - 'male': 31796, 'female': 27287,
  - 'brand': 60147, 'unknown': 35361.

3. Which user_timezone has the maximum number of tweet_count ?

I used the *group_by* function to group the rows using the user_timezone and then used apply function to calculate the sum of *tweet_count* in each group. Then I *sorted* them after converting the data frame into a dictionary.

- ➢ Upon doing the entire calculation, it is found that **Eastern Time (US & Canada)** time zone has the maximum number of tweet Counts from the users.