ML-MAJOR-JULY-ML07B9
PROJECT - REPORT

Name: - Ankit Raibole
College:- Indian institute of information technology, Nagpur

Problem Statement:- For a given dataset (problem) which is the best classification algorithm (as per accuracy).

In this project, we will be predicting the gender-based on their tweets. We used the dataset mentioned in the mail.
We will be using tweets text their tweeter description, user name as main independent features and gender will be our dependent variable.

Our first step was to clean the dataset. We selected only those data points which had "gender: confidence" of 1 as it means that they are 100% confident that the gender given in dataset is correct. Then we removed the "unknown" gender and "brand" gender from the data-frame.

Next, we cleaned the name column to get the actual name from the user-name. Cleaning involves splitting the words into name and surnames based on the camel cases, numbers or special character's for example "AnkitRaibole" will become "Ankit Raibole".And removing of special characters, single letters, numbers and stop-words.this gives us the actual name. Also correction for spelling mistakes was done.We converted the encoding to utf-8 as it covers all English words which helps to increase accuracy.
Next we combined the text and description column into one column as "context" and cleaning and spelling correction was done on context column.
Next, we cleaned the description and text column. cleaning text and description column were done as tokenizing the words, removing stop words, numbers, special characters, and also doing spelling correction later on after counting the typos for each gender classes and calculating the frequently used words.
We removed the words which were common for both males and females as it will help to distinguish better .
For spelling correction, we used "symspellpy" as it's very fast as compared to other packages.
Later we used countVectorizer to convert text to a matrix of token counts.
For training the model we used clean and corrected text columns for better accuracy.
Then we choose 3 model as Naive Bayes, logistic regression, and support vector classifier.among this 3, Naive Bayes performed best with an accuracy of 80.7297%, logistic regression with an accuracy of 73.888%, support vector classifier with an accuracy of 72.291%.

We used voting classifier to ensemble 3 models .ensembled model gives the accuracy of 79.2474%.

The two questions asked were answered as follows :-

Que1. What are the most common emotions/words used by Males and Females?
Ans:We combined the two columns text and description and counted the frequencies of the words using stack and value_counts functions and after that we found that:
The most common word used by men and woman were the same which was the word ″love″ which implies the emotion also.
For male's, the frequency was 452 time.
For female's, the frequency was 697 times.

Que2.Which gender makes more typos in their tweets?
Ans: we used symspellpy to check for spelling mistakes. For each word we checked if its a typo or not and then increases the count if its a typo.
For male's, the number of typing mistake was 7308 words in their description and 6099 words in there tweets which in total was 13407 words.
For female's the number of typing mistakes was 6602 words in their description and 5756 words in there tweets which in total was 12358 words.

Que3. What word are used in brands tweets and descriptions?
Ans: For the description most used words by brands were "news","official","business","service".
And for tweets most used words were "weather", "update", "job".

Que4. What are the most common names in dataset?
Ans: For the given dataset the names of individual account users are distinct but for brands we see common words begin used,mostly used words in brand names were "news","radio","wellness",""sport","health"

Summary

In this project, we created a machine learning model to predict gender from the person's tweets and description and name.
We have done cleaning on the text columns to increase the model accuracy.
We answered the questions asked on the dataset as mentioned earlier in the report.
We ensembled Naive Bayes, logistic regression and support vector classifier.
We got an accuracy of 79.2474% for the ensembled model.
From the 3 algorithms, Naive Bayes performs better with an accuracy of 80.7297%.