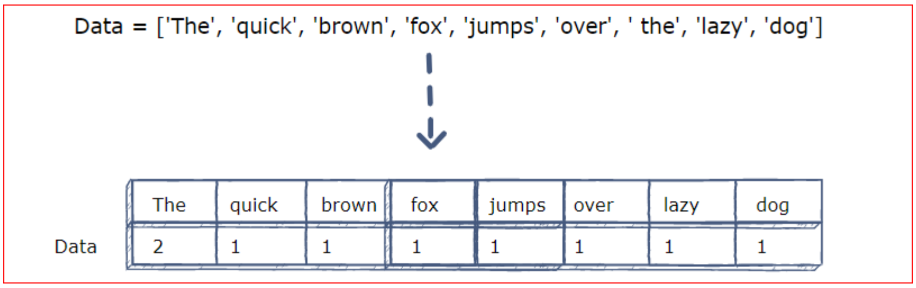
Name:-Harshvardhan Arvind Singh
Email:-harsh3166@gmail.com

SUMMARY OF MAJOR PROJECT

We Import all the necessary packages and modules required in our project , and read the dataset into a dataframe with name df.
As we had to build our model taking gender as the dependent variable , we cleaned the dataset and chose only those rows where gender contained the values “male” or “female” and had its gender confidence more than 0.99.
In the obtained dataframe we checked for rows with empty or null values in the description column and discarded them.
We normalised the text, and description columns of our dataframe(i.e remove all the non ASCII characters, URLS,Special characters and double spaces from all the rows) and created two new columns ‘text_norm’ and ‘description_norm’ and saved the results .
We concatenate normalised text and description columns and create a new column “all_features” and save our results.
To further make our dataset more clean we tokenize every row of column “all_features” , lemmatize (i.e convert the word into their root form) and remove stopwords and create a new column “corrected” and save the results.
Using Scikit-learn’s CountVectorizer we convert “corrected” column of df to a vector of term/token counts which returns a sparse matrix (i.e as many columns as unique words in our “corrected” column with the no of time the word has occurred in a row as its values) which is our feature set for prediction.

For example



Data = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', ' the', 'lazy', 'dog']

| | The | quick | brown | fox | jumps | over | lazy | dog |
|---|---|---|---|---|---|---|---|---|
| Data | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Using label encoder convert “gender” column from categorical into a numeric machinelearning usable form.

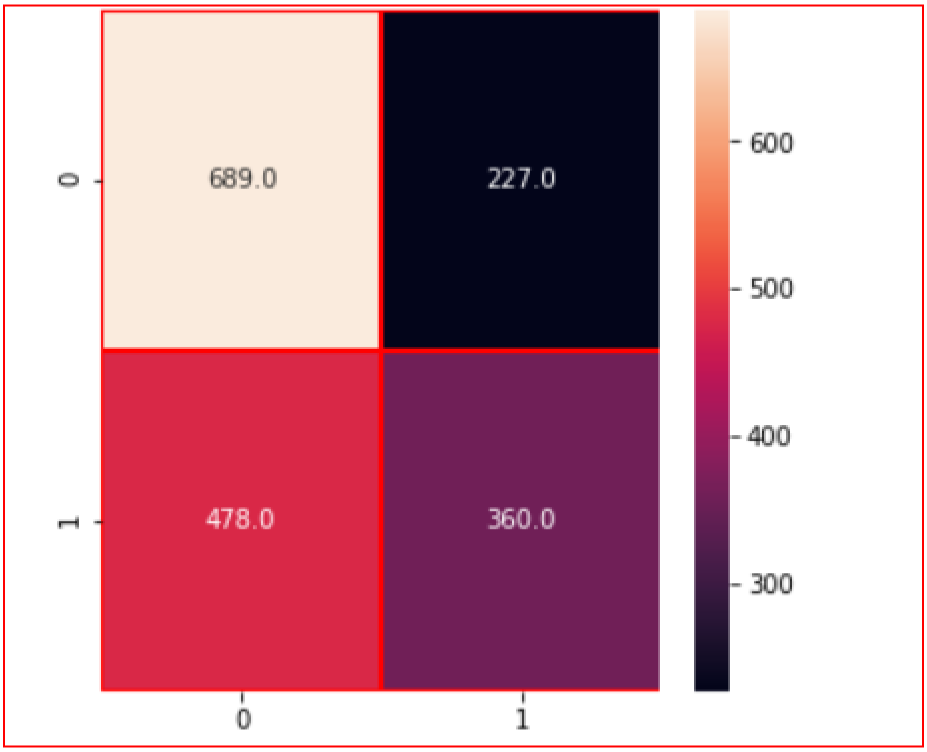
| Labels | Male | Female |
|---|---|---|
| Labels after encoding | 0 | 1 |

Using Scikit-learn’s train_test_split we Split the data into 80% training data and 20% testing data.
We create four different models i.e Naïve Bayes Classifier , Logistic regression, Support Vector Classifier, Random Forest Classifier and trained our models on the training data and checked their accuracy on testing data.We found that the the most accurate model for this classification is naïve bayes with highest accuracy among all.
We chose Logistic regression, Support Vector Classifier, Random Forest Classifier for building an ensemble model (i.e which takes predictions from all the models and returns the value with maximum vote) and implemented this by creating an np array “final_pred” which stores the mode of all the three predictions.

For Example.

| Logistic regression | Support Vector Classifier | Random Forest Classifier | final_pred |
|---|---|---|---|
| 0 | 0 | 1 | 0 |

Further we plotted a heatmap for confusion matrix of our ensemble model to get a better idea of our models' accuracy.



The accuracy of our final ensemble model sums up to be around 60%.

https://drive.google.com/file/d/1f68MUU-2GaebTeui6JlVe0P4tRIfB5Jq/view?usp=sharing

Questions we asked on our dataset are

Q1.How well do stylistic factors (like link color and sidebar color) predict user gender?

We built a classification model with  link color and sidebar color as our features and gender as our target variable.

As link colour and sidebar colour are categorical features we used one hot encoding to convert them into machine learning usable numerical form(we chose one hot encoding cause we had only two features under consideration here and we wanted to increase the no of features count to increase accuracy of our model).

Using Naïve Bayes classification we got an accuracy of about 60% of our model.

https://drive.google.com/file/d/1hK4KtTAecff2iR0toKnQE1bW5INLIv0G/view?usp=sharing

Q2.What are the top 10 most used word by male,female and brand users respectively.

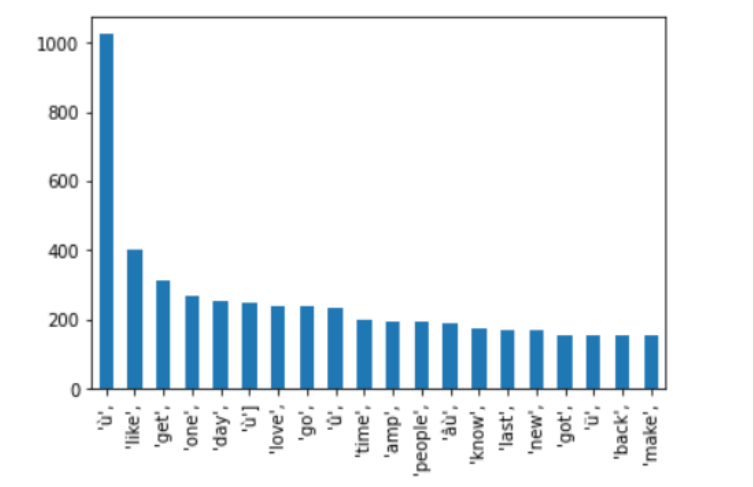
We cleaned the text column and normalised it. (i.e remove all the non ASCII characters, URLS,Special characters and double spaces from all the rows) and saved our results to a new column “Tweets”

We Split the dataset into male df(having all male rows ) female df (having all female rows) and brand df(having all the brand rows)

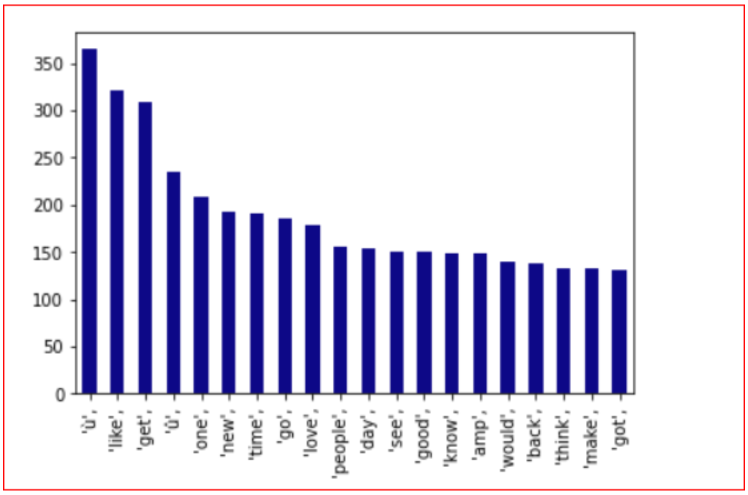
Typecast the Tweet column into string type , convert it to lower case and split it using split(“ “) , join every word with separation of “ “ and create Pandas Series of our result , now using value_counts() method of pandas series we found the 10 most used words by males, females and brands respectively.

Using matplotlib.pyplot we plotted bar graph to represent the most used  words and their frequency better.

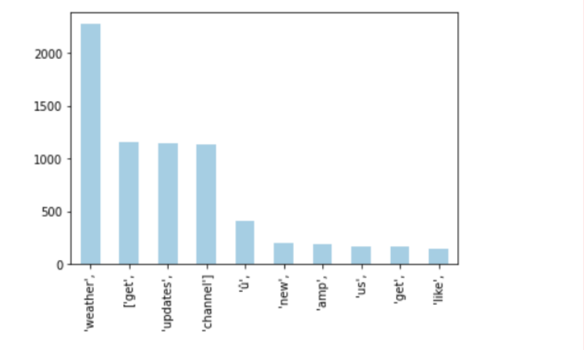
10 most used words by female



10 most used words by male



10 most used words by Brands



https://drive.google.com/file/d/1ZVNbmiC5uO1WqW85_1GCwSIGVx1nZK-O/view?usp=sharing