MAJOR PROJECT- ML062B1

In this Machine Learning project we took a twitter dataset.This twitter data set gives us information about various aspects of a user tweets like how many times a user retweeted? What gender is the user? And various other such information.

First we cleaned the dataset such that as much noise is minimized and then we performed Exploratory data Analysis (EDA) on this dataset to gain better understanding.Some questions are as follows

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

1.Gender count in this dataset:

Male : 5469
Female :5725

So from this we get to understand that there are 5469 male users and 5725 female users in our dataset

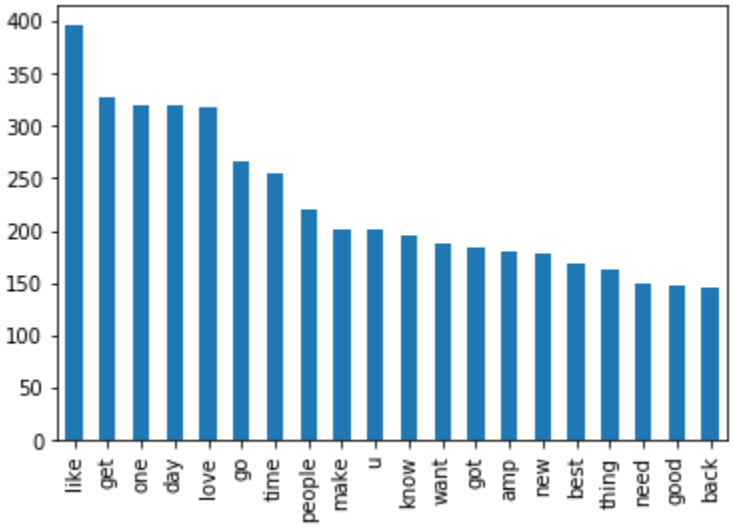
2.What are the most common emotions/words used by Males and Females?

Words used by female(count) :

like	396
get	328
one	320
day	319
love	317
go	266
time	254
people	221
make	202
u	201
know	196
want	188
got	184
amp	181
new	178
best	168
thing	162
need	150
good	148
back	146

These are the count of words used by female users

Graphical representation of this data is as follows :

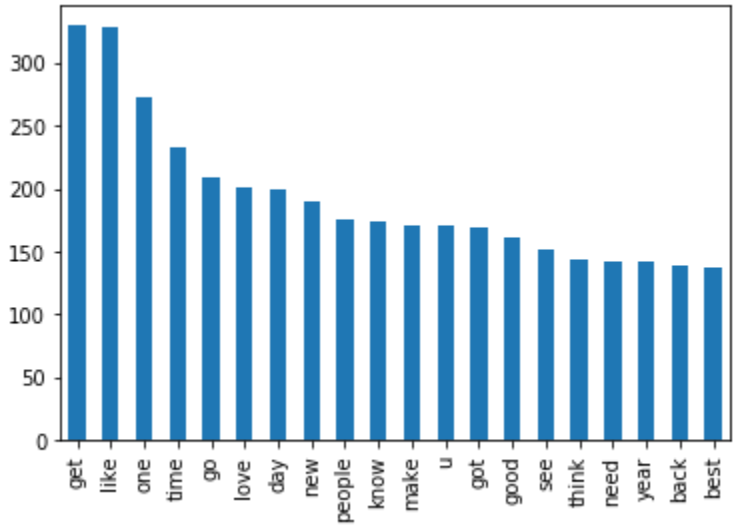


Words used by male(count) :

get	330
like	328
one	272
time	233
go	209
love	201
day	200
new	190
people	176
know	174
make	171
u	170
got	169
good	162
see	152
think	144
need	142
year	142
back	139
best	138

These are the count of words used by male users

Graphical representation of this data is as follows :



3.Which gender makes more typos in their tweets?

Male

We arrived at this conclusion by using TextBlob and counting each gender typo individually

Next for visualisation of data we used WordCloud

WordCloud representation for text is as follows ;

WordCloud representation for description is as follows :

In the next step we took 4 classification algorithms which are

1.MULTINOMIALNB

2.DECISION TREE CLASSIFIER

3.RANDOM FOREST

4.SUPPORT VECTOR CLASSIFIER

From these algorithms we built 4 respective machine learning models from these models.In all algorithms we took GENDER as a dependent variable.

In the first case we predicted gender by only using text column.

Accuracies for each algorithm in this case is as follows :

MULTINOMIALNB CLASSIFICATION ALGORITHM : 0.5874643874643874
DECISION TREE CLASSIFIER CLASSIFICATION ALGORITHM : 0.5247863247863248
RANDOM FOREST CLASSIFICATION ALGORITHM : 0.5601139601139601
SUPPORT VECTOR CLASSIFIER : 0.5863247863247864

From this above information we can suggest that MULTINOMIALNB CLASSIFICATION ALGORITHM gives us best possible accuracy to predict gender when only text column is taken.It is also noteworthy that SUPPORT VECTOR CLASSIFIER also gives us good accuracy

In the second case to get slightly more accuracy we took two columns(text and description) to predict gender

Accuracies for each algorithm in this case is as follows :

MULTINOMIALNB CLASSIFICATION ALGORITHM : 0.6957264957264957
DECISION TREE CLASSIFIER CLASSIFICATION ALGORITHM : 0.611965811965812
RANDOM FOREST CLASSIFICATION ALGORITHM : 0.674074074074074
SUPPORT VECTOR CLASSIFIER : 0.6564102564102564

Again in this case MULTINOMIALNB CLASSIFICATION ALGORITHM gives us good accuracy closely followed by RANDOM FOREST CLASSIFICATION ALGORITHM

Finally we can say that though each algorithm performs good in a certain case that is having their own advantage and disadvantage in our case of twitter dataset to predict gender we suggest using MULTINOMIALNB CLASSIFICATION ALGORITHM