MAJOR PROJECT

ML-MAJOR-NOV-ML11B2

ARUN PRAKASH

Libraries used:

Pandas, Numpy, Matplotlib, Seaborn, sklearn

Dataset Description:

This dataset contains twitter profile of male, female or a brand each with a profile image, date of account creation, their tweets, tweet and retweet count and even sidebar color.

The dataset contains 20050 rows and 26 columns before cleaning.

Data Cleaning:

Cleaning of data and dealing with missing values was done by using pandas and seaborn libraries during various stages.

By using isna().sum() we can find out how many missing values were there and by using sns.heatmap() for whole dataset we can see where all the missing values are located.

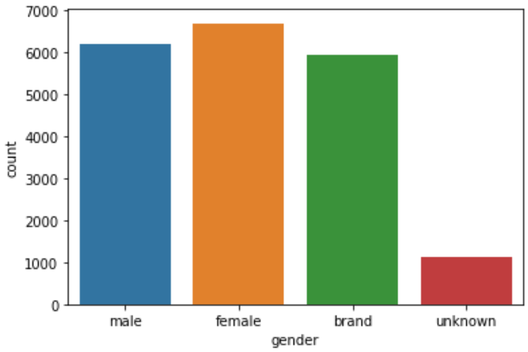
- At first gender_gold,profile_yn_gold ,tweet_coord columns were removed as they were largely empty
- Then tweet_location,user_timezone was removed because we cannot fill it
- description column was removed since it was text based data and we cannot fill and many of rows were empty
- Lastly gender column had 97 missing values and those missing values was removed by using dropna()

After cleaning the dataset , it contains now 19953 rows and 18 columns.

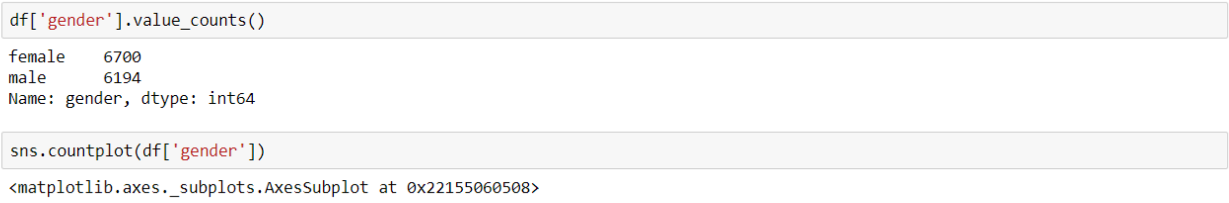
EDA:

Gender was considered the dependent value from the dataset

Distribution of gender:



Removing brand and unknown from gender because we don’t know whether they are male or female
Distribution of gender:

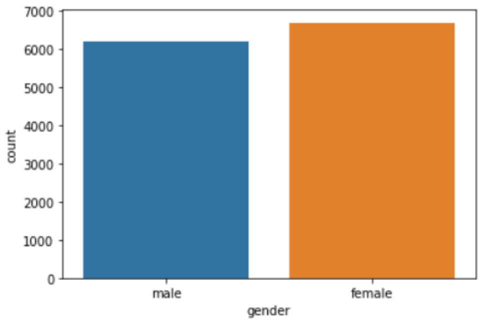


```
df['gender'].value_counts()
```

```
female    6700
male      6194
Name: gender, dtype: int64
```

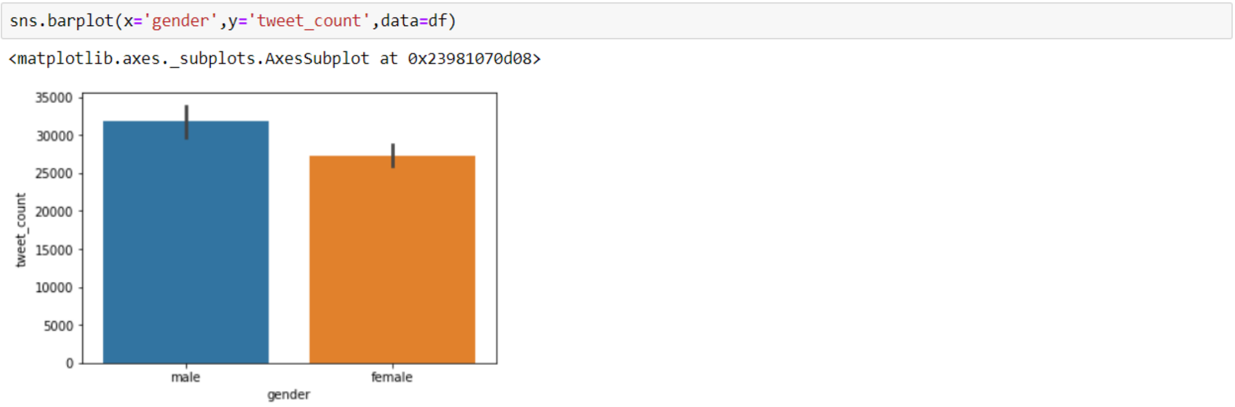
```
sns.countplot(df['gender'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x22155060508>
```



Problem Statements:

1. **Which gender has the third highest tweet_count and also visualization of total tweet_count of female and male?**

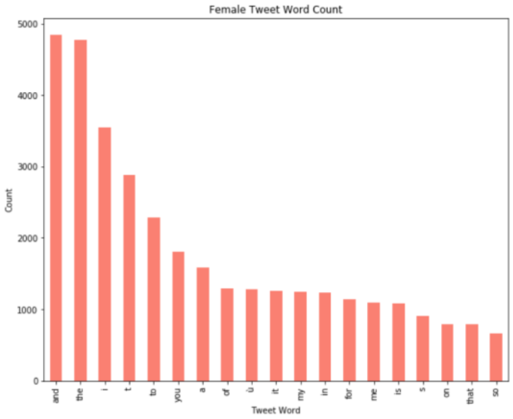


```
sns.barplot(x='gender',y='tweet_count',data=df)
```

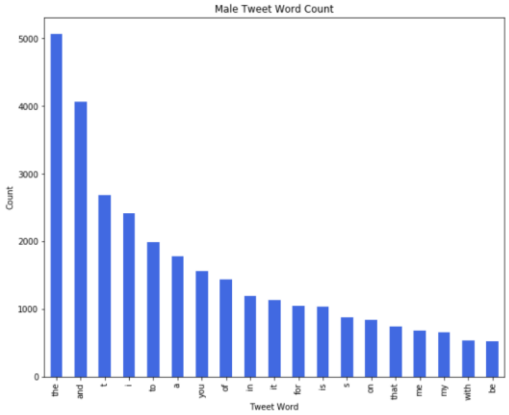
```
<matplotlib.axes._subplots.AxesSubplot at 0x23981070d08>
```

From the above plot we can see that male gender got the highest tweet count

2. **Most common words used by Male and Females**



From the above plot we can see that most words used by female is 'and'.



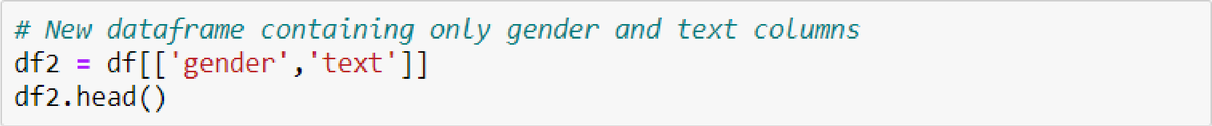
From the above plot we can see that most words used by male is 'the'.

Most common word used by female is ‘and’.

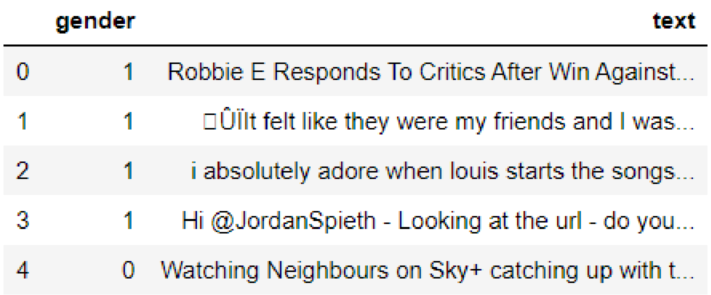
Most common word used by male is ‘the’.

ML Models and Data Preprocessing:

New data frame containing only gender and text column was created for modelling.



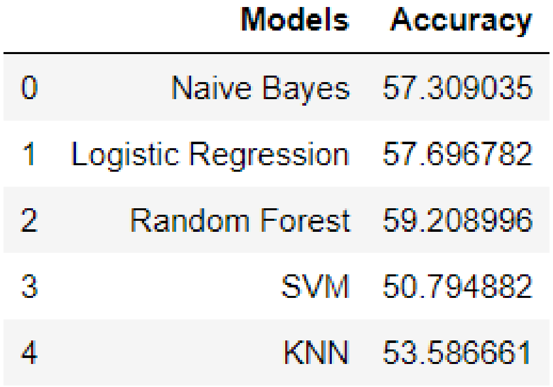
```
# New dataframe containing only gender and text columns
df2 = df[['gender','text']]
df2.head()
```



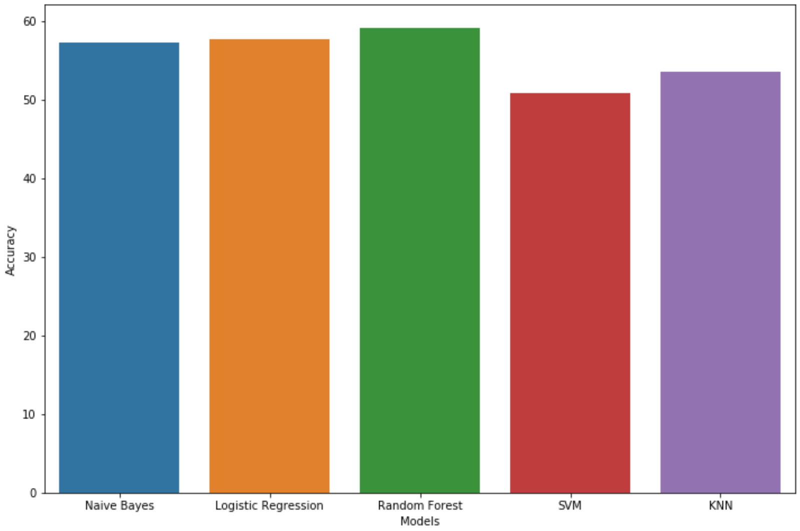
	gender	text
0	1	Robbie E Responds To Critics After Win Against...
1	1	□ÛÏIt felt like they were my friends and I was...
2	1	i absolutely adore when louis starts the songs...
3	1	Hi @JordanSpieth - Looking at the url - do you...
4	0	Watching Neighbours on Sky+ catching up with t...

Machine Learning models used:

1. **Naive Bayes:**
   - Independent Attributes: Sparse Matrix of Text Column
   - Dependent Attributes: gender
   - Accuracy of the model: **57.03%**
2. **Logistic Regression:**
   - Independent Attributes: Sparse Matrix of Text Column
   - Dependent Attributes: gender
   - Accuracy of the model: **57.69%**
3. **Random Forest:**
   - Independent Attributes: Sparse Matrix of Text Column
   - Dependent Attributes: gender
   - Accuracy of the model: **59.20%**
4. **SVM Classification:**
   - Independent Attributes: Sparse Matrix of Text Column
   - Dependent Attributes: gender
   - Accuracy of the model: **50.79%**
5. **K Nearest Neighbor:**
   - Independent Attributes: Sparse Matrix of Text Column
   - Dependent Attributes: gender
   - Accuracy of the model: **53.58%**



	Models	Accuracy
0	Naive Bayes	57.309035
1	Logistic Regression	57.696782
2	Random Forest	59.208996
3	SVM	50.794882
4	KNN	53.586661



Conclusion:

We can see that best machine learning model was **Random Forest Classification** which had an accuracy of 59.20% compared to other models.