Gender Classifier by using Twitter dataset

T5 Bootcamp Data Science Project

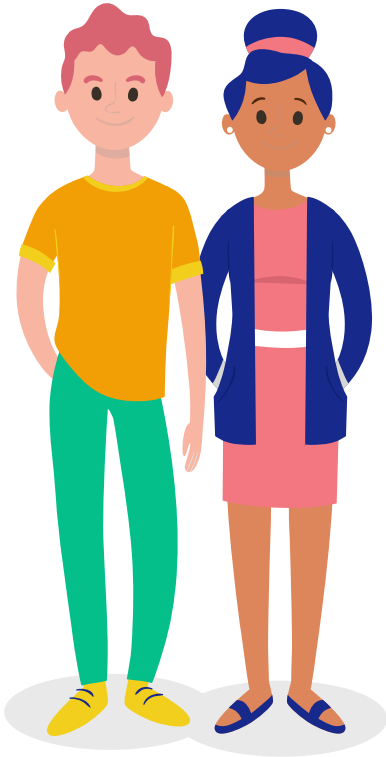
Aryaf Almusaiteer

Project Goal :

The goal of the project is to classify the gender based on random tweet and the bio. This will help to know more words for each gender or brand to know people with fake accounts causing incitement or insulting and criticizing political matters. It will make it easier to classify them and therefore they will be treated

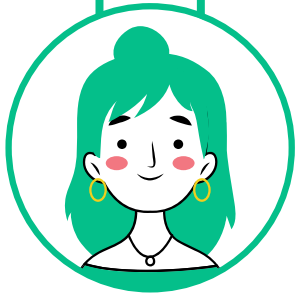
About the Data:

The Data has been extracted from Kaggle. The dataset consists of 20050 rows and 26 columns. Among 26 columns there are 25 predictor variables and 1 target variable which is gender in this case.



Process Data

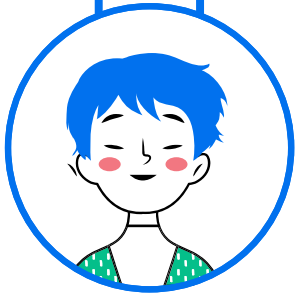
1



Cleaning

- Clean Special chars
-Convert to Lowercase
- Delete stop words
- Lemmatize the text

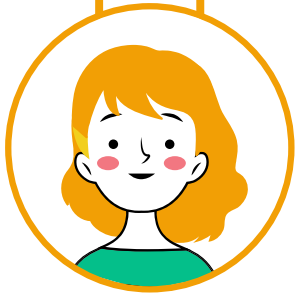
2



Labeling

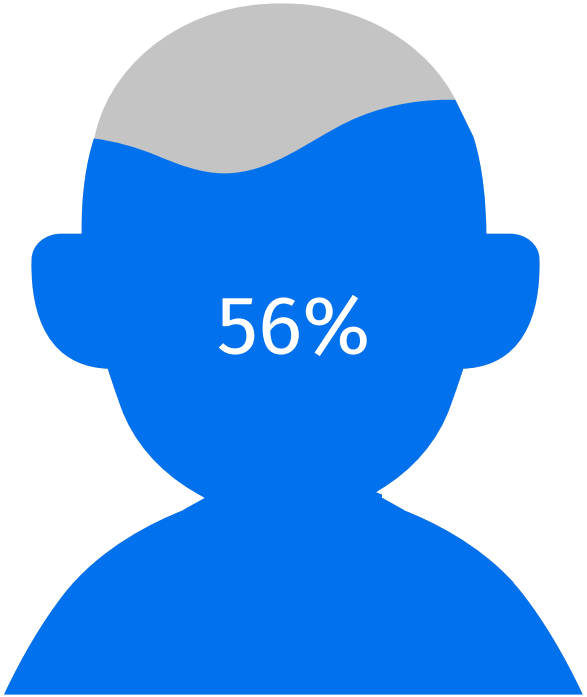
Labeling the text

3



Bag of word

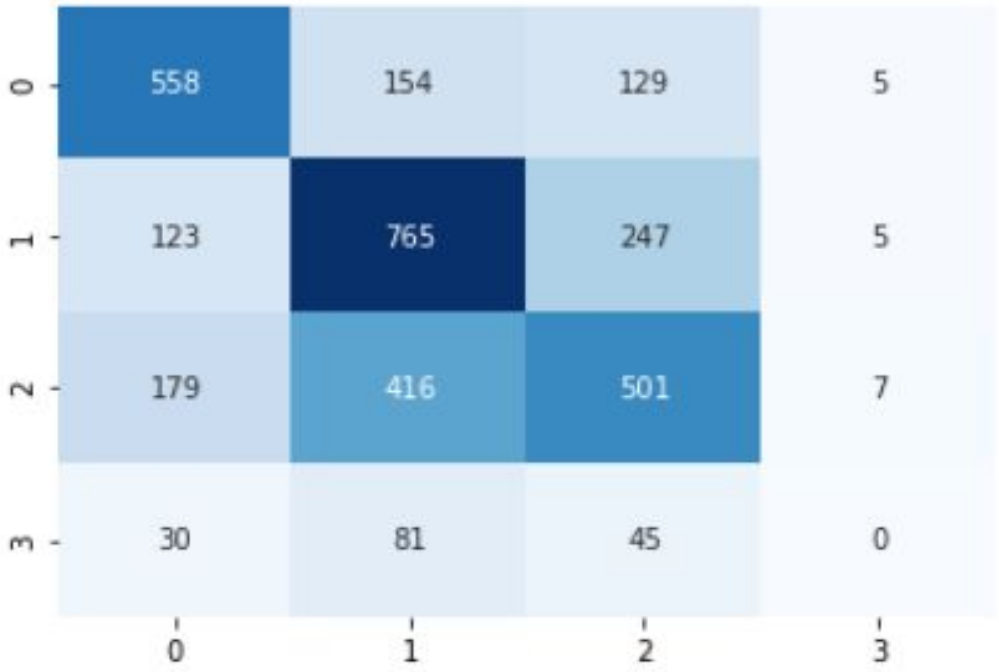
Create bag of words by using Vectorizer



Modeling :

Multinomial Naive Bayes Model was used to classify the tweets based on the gender with 56.20% accuracy

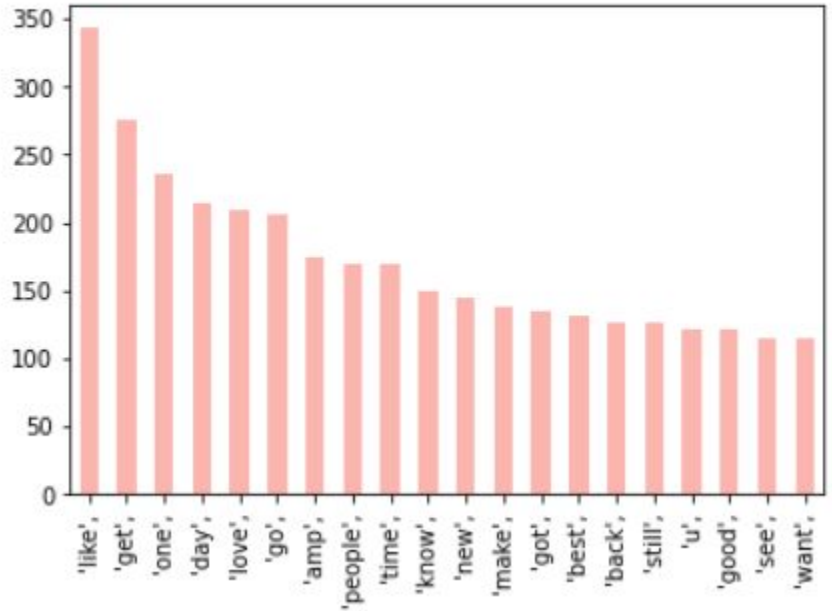
The result :



The result :

```
Female_Words.plot(kind='bar',stacked=True, colormap='Pastel1')
```

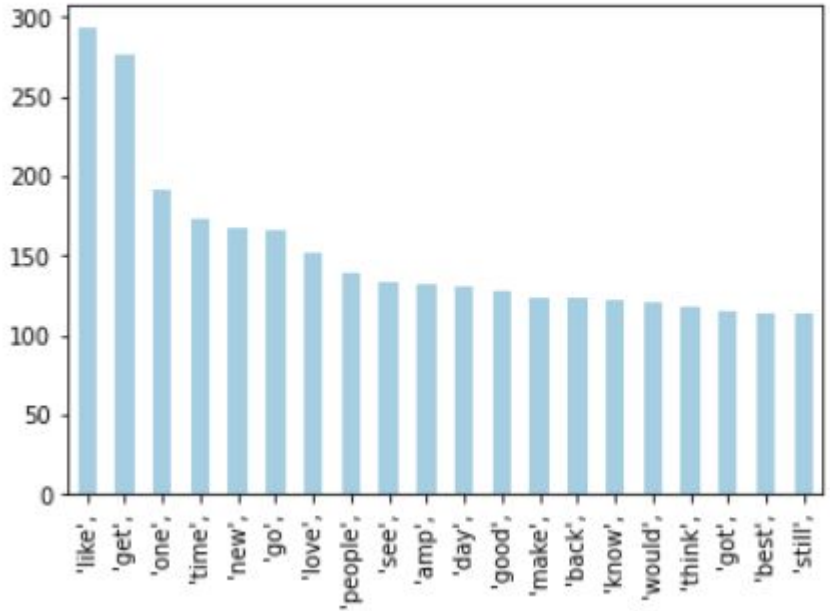
```
<AxesSubplot:>
```



The result :

```
Male_Words.plot(kind='bar',stacked=True, colormap='Paired')
```

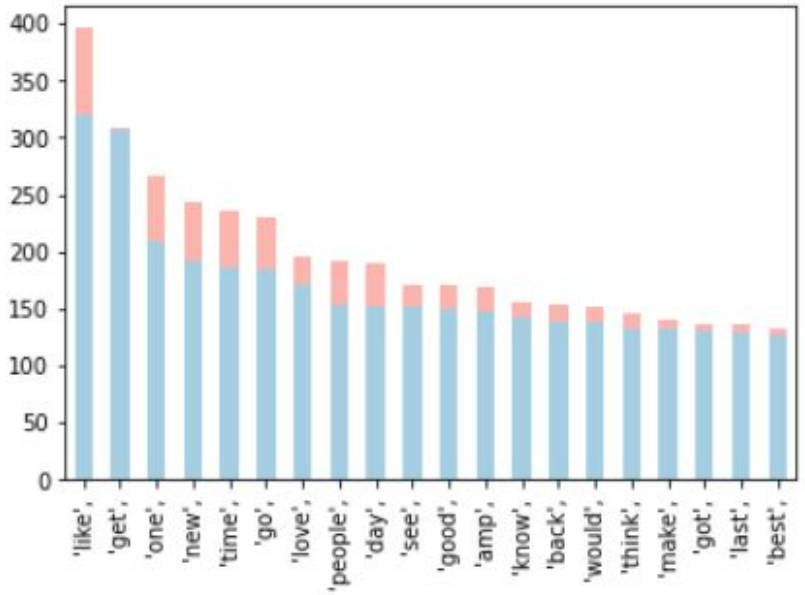
```
<AxesSubplot:>
```



The result :

```
#A common word between female , male
Female_Words.plot(kind='bar',stacked=True, colormap='Pastel1')
Male_Words.plot(kind='bar',stacked=True, colormap='Paired')
```

```
<AxesSubplot:>
```



Thanks a lot