MACHINE LEARNING PROJECT

verzeo machine learning main project batch no:ML063B2

In [7]:

```
import pandas as pd
import seaborn as sb
import numpy as np
import matplotlib.pyplot as plt
import scipy.optimize as opt
from sklearn import preprocessing
%matplotlib inline
plt.rcParams['figure.figsize'] = 10,10
```

In [8]:

```
df =pd.read_csv('C:\\Users\\Shreyas Venishetty\\Desktop\\GOAL STREET\\Machine-Learning-master\\DATA\\Information.csv',engine ='python')
```

In [9]:

```
df.head(3)
```

Out[9]:

	_unit_id	_golden	_unit_state	_trusted_judgments	_last_judgment_at	gender	gender:confidence	profile_yn	profile_yn:co
0	**815719226**	**False**	**finalized**	**3**	**10/26/15 23:24**	**male**	**1.0000**	**yes**	
1	**815719227**	**False**	**finalized**	**3**	**10/26/15 23:30**	**male**	**1.0000**	**yes**	
2	**815719228**	**False**	**finalized**	**3**	**10/26/15 23:33**	**male**	**0.6625**	**yes**	

3 rows × 26 columns

In [10]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20050 entries, 0 to 20049
Data columns (total 26 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   _unit_id               20050 non-null  int64
 1   _golden                20050 non-null  bool
 2   _unit_state            20050 non-null  object
```

```
 7   profile_yn              20050 non-null  object
 8   profile_yn:confidence   20050 non-null  float64
 9   created                 20050 non-null  object
 10  description             16306 non-null  object
 11  fav_number              20050 non-null  int64
 12  gender_gold             50 non-null     object
 13  link_color              20050 non-null  object
 14  name                    20050 non-null  object
 15  profile_yn_gold         50 non-null     object
 16  profileimage            20050 non-null  object
 17  retweet_count           20050 non-null  int64
 18  sidebar_color           20050 non-null  object
 19  text                    20050 non-null  object
 20  tweet_coord             159 non-null    object
 21  tweet_count             20050 non-null  int64
 22  tweet_created           20050 non-null  object
 23  tweet_id                20050 non-null  float64
 24  tweet_location          12566 non-null  object
 25  user_timezone           12252 non-null  object
dtypes: bool(1), float64(3), int64(5), object(17)
memory usage: 3.8+ MB
```

Divide the dataset

Here the dataset is divided into three df,male and female dataset for easy handling.

In [11]:

```
female=df.loc[df.gender=='female']
male = df.loc[df.gender=='male']
brand=df.loc[df.gender=='brand']
df = df[df["gender"].isin(['male','female'])]
```

label encoding

Label encoding the gender column

In [12]:

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['gender2']=le.fit_transform(df['gender'])
```

Data Exploration and feature selection

exploring the data and get rid of unwanted columns.

In [13]:

```
df=df[['_trusted_judgments','gender2', 'gender:confidence',
       'profile_yn:confidence', 'fav_number',
       'retweet_count','tweet_count','text']]
```

In [14]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12894 entries, 0 to 20049
Data columns (total 8 columns):
```

```
 2   gender:confidence       12894 non-null  float64
 3   profile_yn:confidence   12894 non-null  float64
 4   fav_number              12894 non-null  int64
 5   retweet_count           12894 non-null  int64
 6   tweet_count             12894 non-null  int64
 7   text                    12894 non-null  object
dtypes: float64(2), int32(1), int64(4), object(1)
memory usage: 856.2+ KB
```

In [15]:

```
df.columns
```

Out[15]:

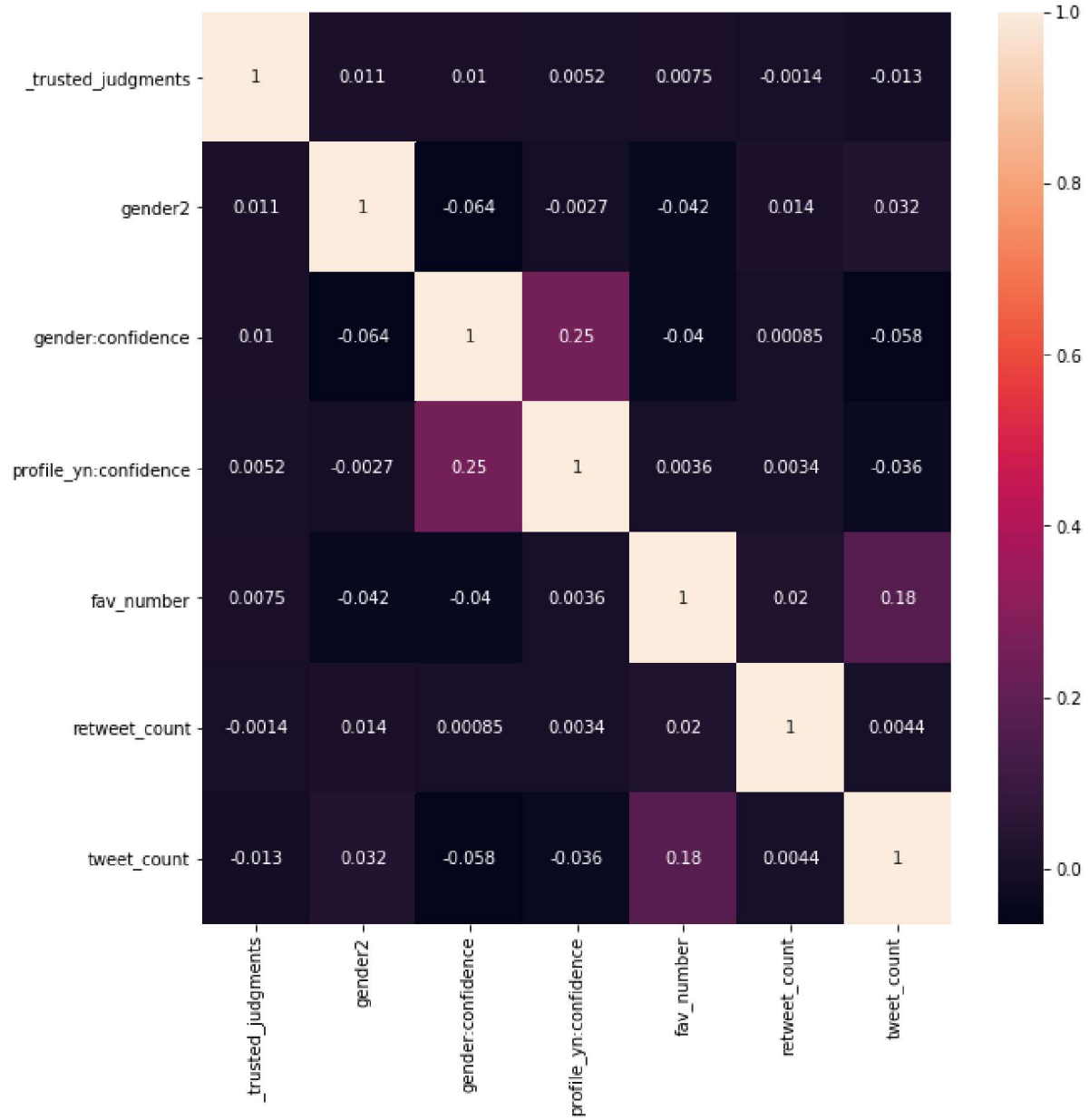
```
Index(['_trusted_judgments', 'gender2', 'gender:confidence',
       'profile_yn:confidence', 'fav_number', 'retweet_count', 'tweet_count',
       'text'],
      dtype='object')
```

In [16]:

```
sb.heatmap(df.corr(), annot =True)
```

Out[16]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x206f7d70cc8>
```



COMMON WORDS

Lets find the common words used by male and female. Here we did find the most used words by joining all the ['text'] values using the built in function "join" and count the words and found the answers

common words used by male

In [17]:

```
male.head(2)
```

Out[17]:

	_unit_id	_golden	_unit_state	_trusted_judgments	_last_judgment_at	gender	gender:confidence	profile_yn	profile_yn:co
0	815719226	False	finalized	3	10/26/15 23:24	male	1.0	yes	
1	815719227	False	finalized	3	10/26/15 23:30	male	1.0	yes	

2 rows × 26 columns

In [18]:

```
pd.Series(' '.join(male.text).split()).value_counts()
```

Out[18]:

```
the                4251
and                3685
to                 1906
a                  1645
I                  1570
                   ...
MTV's                 1
@TrulyYours40         1
TRUMP                 1
Bryan,                1
@NerdHeardBen!        1
Length: 27670, dtype: int64
```

common words used by female

In [19]:

```
female.head(2)
```

Out[19]:

	_unit_id	_golden	_unit_state	_trusted_judgments	_last_judgment_at	gender	gender:confidence	profile_yn	profile_yn:co
4	815719230	False	finalized	3	10/27/15 1:15	female	1.0	yes	
5	815719231	False	finalized	3	10/27/15 1:47	female	1.0	yes	

2 rows × 26 columns

In [20]:

```
pd.Series(' '.join(female.text).split()).value_counts()
```

Out[20]:

```
and            4352
the            4049
I              2364
to             2188
a              1454
               ...
"Ugly             1
damages           1
_Ù÷â_ÙÕ¥          1
#equalpay         1
@Valcaakes        1
Length: 26141, dtype: int64
```

therefore the common words used by male and female are "and" , "the" & "to"

AVERAGE WORDS

Lets find the average words used by both gender
here we create a dictionary and iterate through the the "text" column and find how many times the each words occured after that sum of the count was found and divide it with column value.

In [21]:

```
number_of_words={}
for i in range(female.shape[0]):
    words=str(female['text'].values[i]).split(' ')
    for j in words:
        try:
            count=number_of_words[j]
            number_of_words[j]=count+1
        except:
            number_of_words[j]=1

number_of_words
```

Out[21]:

```
{'Watching': 5,
 'Neighbours': 1,
 'on': 710,
 'Sky+': 1,
 'catching': 4,
 'up': 295,
 'with': 616,
 'the': 4044,
 'Neighbs!!': 1,
 'Xxx': 2,
 '_Ù÷Ä_Ù÷Ä_Ù÷Ä_ÙÔÎ_Ù�È_ÙÔ�_Ù�È': 1,
 'Ive': 6,
 'seen': 30,
 'people': 198,
 'train': 10,
 'lamps,': 1,
 'chairs,': 1,
 'tvs': 1,
```

```
'pic': 6,
'defines': 1,
'all': 353,
'mcd': 2,
'fangirls/fanboys': 1,
'and': 4250,
'shippers': 2,
'xD': 6,
'@Evielady': 1,
'just': 464,
'how': 167,
'lovely': 10,
'is': 977,
'tree': 3,
'this': 403,
'year!': 2,
'Never': 10,
'it': 728,
'as': 156,
'gorgeous': 7,
'#Autumn': 2,
'#colour': 1,
'Just': 118,
'put': 58,
'my': 1065,
'ass': 37,
'line': 10,
'for': 1071,
'you': 1298,
'repay': 1,
'me.': 41,
'will': 180,
'i': 328,
'even': 115,
'need': 144,
'sound': 7,
'effects': 4,
'diviners': 1,
'tonight': 26,
'https://t.co/EROGWTFTYo': 3,
"It's": 90,
'a': 1454,
'glow': 2,
'of': 1208,
'satisfaction': 1,
're': 8,
'The': 541,
'Glow': 1,
'https://t.co/7RdyMCVPKx': 1,
'@giannaaa28': 1,
'lmao': 19,
'_Ù÷â_Ù÷â': 25,
'dude': 9,
"I'm": 481,
'hella': 4,
'scared': 12,
'next': 67,
'episode': 14,
'bc': 43,
'ending': 4,
'to': 2186,
"yesterday's": 2,
'@CraftYear2015': 1,
'@isabelpascual': 1,
'thank': 54,
'retweets': 4,
'All': 38,
'girls': 24,
```

```
'in': 1131,
'floor': 13,
'watched': 26,
'us_Ù÷â_Ù÷â': 1,
'@ChrisAOfficial': 1,
'right': 92,
'side_Ù÷â‰ÏÎ_Ù_•ü xxx': 1,
'@SydnieJR': 1,
'except': 4,
'once': 25,
'Hallmark': 2,
'movies': 9,
'start': 30,
'I': 2342,
"won't": 26,
'get': 312,
'anything': 36,
'done!!': 1,
'_Ù÷__Ù__Ù_ã': 1,
'You': 107,
'leave': 34,
'group': 24,
'chat': 10,
'more': 136,
'than': 70,
'2': 119,
'mins': 6,
'miss': 55,
'made': 63,
'shit': 83,
'Me': 31,
'week': 53,
"Brandon's": 1,
'birthday:': 1,
"there's": 34,
'no': 130,
'such': 30,
'thing': 87,
'birthday': 44,
'u': 110,
'weirdo\nMe': 1,
'November': 8,
'1st:': 1,
"it's": 214,
'month,': 1,
'bow': 3,
'down': 65,
'me': 781,
'This': 93,
'boy': 16,
'was': 413,
'El': 1,
'wit': 2,
'his': 114,
'3': 56,
'daughters': 1,
'they': 212,
'under': 21,
'5': 49,
'@MarkHicks1204': 1,
'wrong': 31,
'Nandos': 1,
'but': 337,
'found': 39,
'eventually': 1,
'#10chilliesequalsfreenandos': 1,
'Those': 5,
'who': 123,
```

```
'those': 39,
'abandon': 1,
'their': 92,
'friends': 56,
'worse': 7,
'scum': 1,
'VIDEO:': 2,
'James': 8,
'Bond': 6,
'Spectre': 4,
'world': 66,
'premiere:': 7,
'After': 7,
'months': 18,
'build': 8,
'Spectre,': 2,
'latest%Û_': 1,
'https://t.co/uV38Wlg5bE': 1,
'#UK': 2,
'Once': 2,
'complete,': 1,
'lift': 3,
'off': 104,
'attempt': 5,
'connect': 5,
'alpha': 1,
'point': 23,
'rift': 2,
'we': 219,
'observing.': 1,
'(1-2)': 1,
'camerons': 1,
'side': 21,
'bed': 35,
'smells': 5,
'SO': 31,
'bad': 39,
'@wishbonecon': 1,
'going': 115,
'1.30': 1,
'one': 269,
'??': 6,
'https://t.co/nRN2mGLd2E\nAm': 1,
'only': 146,
'loves': 15,
'part': 33,
'Merlin': 1,
"Regina's": 1,
'face?': 1,
':D': 11,
'#OUAT': 1,
'Please': 17,
'God,': 3,
'let': 71,
'house': 34,
'that': 607,
'has': 121,
'fireplace': 2,
'bedroom!!': 1,
'have': 476,
'great': 67,
'ideas': 6,
'room..': 2,
"that's": 71,
'8%': 1,
'hidden..': 1,
'UMPH.': 1,
'https://t.co/OnebBSAG93': 1,
```

```
'tomorrow': 30,
'say': 82,
'should': 76,
'be': 447,
'around': 46,
'20': 12,
'day': 179,
'office': 13,
'-': 197,
'catch': 13,
'@thedjbrisk': 1,
'so': 486,
'wish': 38,
'had': 124,
'x': 20,
'Walter': 1,
'invented': 3,
'bolt-action': 1,
'rifle,': 1,
'liquor,': 2,
'sexual': 2,
'intercourse,': 1,
'football--': 1,
'order.': 2,
'Amazing:': 1,
'See': 14,
'Igbo': 1,
'Village': 1,
'In': 42,
'United': 3,
'States': 2,
'OfåÊAmerica': 1,
'https://t.co/Z8A95hAQpE': 1,
'https://t.co/qdQ6HnE735': 1,
'@R_M_Appleyard': 1,
'alarm': 2,
'wont': 3,
'work': 95,
"can't": 130,
'stops': 3,
'Leeds.': 1,
'Too': 9,
'much': 96,
'traffic': 5,
'ect.': 1,
'A': 92,
'20min': 2,
'bus': 16,
'ride': 9,
'can': 226,
'take': 89,
'50': 1,
'or': 133,
'12': 19,
'mins,': 1,
'ya': 16,
'know?': 2,
'#Akinator,': 1,
'Genie': 1,
'App,': 1,
'guessed': 2,
'thinking': 26,
'Katniss': 1,
"Everdeen's": 1,
'Daughter': 2,
'#what?': 1,
'how?': 1,
'Move': 1,
```

```
'Stretch': 1,
'‰Û_': 11,
':': 14,
'https://t.co/kxkCEFUxQB': 1,
'...': 26,
'https://t.co/n7PtOHGaPQ': 1,
'@RepComstock': 1,
'supporting': 7,
'Father': 2,
'Sequester': 1,
'Speaker.': 1,
'Cut': 1,
'vet': 1,
'pensions,': 1,
'military,': 1,
'medicare.': 1,
'Shame': 1,
'see': 128,
'now': 145,
'way': 82,
'resist': 1,
'Shattered': 1,
'Empire': 4,
'business.': 2,
'Noto': 1,
'cover': 11,
'alone!': 3,
'https://t.co/nLmIC5xDmp': 1,
'accuracy': 1,
'though': 15,
'lol': 74,
'https://t.co/frFforEeLC': 1,
'Thanks': 37,
'#ACA': 2,
'adults': 2,
'eligible': 1,
'#Medicaid': 1,
'#Illinois': 1,
'&amp;': 166,
'able': 14,
'apply': 5,
'online.': 3,
'@YoungInvincible\n\nhttps://t.co/cfwlLqv4K9': 1,
'IBMSocialBiz:': 1,
'future,': 1,
'there': 111,
'narrowing': 1,
'gap': 2,
"people's": 4,
'abilities': 1,
'use': 42,
'tools': 3,
'proficiently.': 1,
'JenniferMcClure': 1,
'#H2HChat‰Û_': 1,
'@iampoojabalaji': 1,
'after': 66,
'wedding': 9,
'anniversary...': 1,
'Sucks': 1,
'suck': 3,
'When': 108,
"you're": 104,
'early': 8,
'an': 133,
'advising': 1,
'appointment': 6,
'bitch': 23,
```

```
'10': 37,
'mins_Ù÷Ô': 1,
'https://t.co/EROGWToizO': 1,
'Sissy‰Ûªs': 2,
'wife': 5,
'loved': 16,
'l': 2,
'Choosing': 1,
'Lingerie': 1,
'‰ÛÒ': 7,
'Very': 8,
'Important': 1,
'Role': 2,
'For': 20,
'Sissy': 3,
'https://t.co/sRf506OsHy': 1,
'If': 80,
'ion': 1,
'job': 21,
',': 50,
'other': 77,
'.': 56,
'Lrt': 2,
'at': 432,
'allllllllllllllll': 1,
'time!!!!!': 1,
'Flexible': 2,
'Diet': 1,
'Plan:': 2,
'Revealed': 1,
'first': 79,
'time,': 5,
'secret': 7,
'methods': 3,
'losing': 3,
'weight,': 1,
'without': 33,
'having': 55,
'give': 46,
'up‰Û_': 1,
'https://t.co/yGQYYVCYKi': 1,
'Would': 12,
'‰Û÷The': 1,
'Affair‰Ûª': 1,
'Be': 18,
'Easier': 1,
'To': 32,
'Watch': 9,
'Kids': 3,
'Weren‰Ûªt': 1,
'Such': 2,
'Brats?': 1,
'[Spoilers]:': 1,
'Showtime': 1,
'series': 6,
'Affa...': 1,
'https://t.co/o6jpVaUFdz': 1,
'NBD,': 1,
'Comet': 1,
'Is': 35,
'Spewing': 1,
'Gallons': 1,
'Wine': 2,
'Space:': 1,
'Scientists': 1,
'discovered': 2,
'comet': 1,
'Lovejoy': 1,
```

```
'With': 23,
'Wind': 1,
'Clark': 1,
'Gable': 1,
'Vivien': 1,
'Leigh': 1,
'#Photo': 1,
'https://t.co/yVO2B34jrw': 1,
'#Film': 1,
'#Actress': 1,
'good': 118,
'seeing': 21,
'@Caspar_Lee': 6,
'_Ù÷÷': 9,
'enjoy': 20,
'film': 13,
'xx': 9,
'Photo:': 15,
'Fellowship': 1,
'Ring:': 1,
'Book': 5,
'One': 29,
'Lord': 8,
'Rings‰Û_:': 1,
'https://t.co/4A1v5WdDd4': 1,
'https://t.co/Hyosq3fr5e': 1,
'kept': 8,
'telling': 19,
'myself': 41,
'temptations': 1,
'would': 111,
'in,\nAnd': 1,
'did.': 6,
'Fool': 2,
'three': 10,
'times,': 1,
'fuck': 52,
'peace': 6,
'signs\nLoad': 1,
'chopper,': 1,
'rain': 7,
'world(esp': 1,
'EqAnon': 1,
'island)needs': 1,
'like': 400,
'@EQcompliments,': 1,
'@HorseShowTutor,': 1,
'@StudioOnWhitney,': 1,
'@DejavuEq': 1,
'@AthleticEq': 1,
'#AmazingLadies': 1,
'No': 23,
'matter': 7,
'results,': 1,
'words': 15,
'express': 2,
'proud': 17,
'am': 95,
'@csunfasa!': 1,
'dedication': 2,
'these': 57,
'new': 151,
'old': 41,
'members': 6,
'gave': 17,
'left': 38,
'speechless!': 1,
'Hello': 10,
```

```
'yep': 2,
'wait': 54,
'until': 27,
'‰ÛÏ@thacelebritea': 1,
'Seen': 3,
'Scene:': 1,
'Spotted': 1,
'#RavenSymone': 1,
'out': 288,
'about': 271,
'earlier': 6,
'[r.p': 1,
'@theybfdaily': 1,
']': 3,
'https://t.co/sjNH93a0L3‰Û�': 1,
'‰÷¼•ü‰÷¼•ü‰÷¼•ü�': 2,
'@justinbieber': 19,
'swag': 1,
'Acoustic': 1,
'version.': 1,
'too': 68,
'cool,': 1,
'silent': 2,
'makes': 131,
'laughter': 1,
'end': 28,
'vibratos': 1,
'_Ù÷�_Ù÷É': 2,
'Listen': 12,
'@johnmdonnelly': 1,
'discussing': 3,
'exclusive': 7,
'reporting': 1,
'potentially': 2,
'fatal': 2,
'flaws': 2,
'#F35': 1,
'ejector': 1,
'seat': 8,
'https://t.co/rZ1BEKbY6P': 1,
'@CQnow': 1,
'@BarCampPhilly': 1,
'@FlipperPA': 1,
'Yes!': 1,
'love': 249,
'throw': 12,
'hat': 1,
'couple': 18,
'local': 3,
'folks': 2,
'@iQmediacorp.': 1,
'Happy': 26,
'arrange': 1,
'meet.': 1,
'Finally': 9,
'realized': 7,
"I'll": 65,
'never': 94,
'answers': 3,
'want': 183,
'Starbucks': 6,
'Mobile': 2,
'Ordering': 2,
'Comes': 2,
'UK': 5,
'Marketing': 6,
'https://t.co/YeCugANJLy\n#emarketingmkt412': 1,
'#Earth': 1,
```

```
'still': 146,
'light.': 1,
'@starsandglitta': 1,
"_ÙÐ¥_Ù_I'm": 1,
'not': 328,
'fake': 10,
'friends,': 3,
'Take': 9,
'back': 135,
'when': 247,
'your': 459,
'clothes': 11,
'took': 27,
'drawers': 1,
'@ell_morton': 1,
'face': 33,
'timed': 1,
'him': 86,
'ago': 13,
"they're": 28,
'@': 37,
'airport': 2,
'Dallas,': 1,
'getting': 65,
'plane': 2,
'hours.': 3,
"They're": 6,
'okay': 16,
'_Ù÷É_Ù÷É': 1,
'luv': 2,
'them': 145,
'Photo': 6,
'https://t.co/TeMkNK17aO': 1,
'best': 148,
'busty': 4,
'boobs': 5,
'click': 8,
'here': 84,
'image#bustyfriends': 4,
'bignaturals,': 3,
'b‰Û_': 3,
'https://t.co/RSbkiea90U': 1,
'woes_Ù÷__Ù_÷_ÙÕâ': 1,
'https://t.co/CXWHx6DTgW': 1,
'ISIS': 1,
'Launches': 1,
'BIZARRE': 1,
'New': 32,
'Weapon': 1,
'War': 6,
'Against': 2,
'Infidel': 1,
'‰_Ê': 1,
'FLOATING': 1,
'CONDOM': 1,
'BOMBS': 1,
'(Video)': 1,
'|': 26,
'BB4SP': 1,
'https://t.co/c4nA4Cc4dK': 1,
'excited': 30,
'spend': 9,
'Halloween': 36,
'night': 46,
'PIT': 1,
'@sigmonmakenzie': 1,
'@ChrisYoungMusic': 1,
'_Ù�È_Ù__': 1,
```

```
'tomorrow.': 8,
'Growing': 5,
'https://t.co/MIjN2Phyly': 1,
'#VoteOneDirection': 1,
'@onedirection': 62,
'#AMAs': 68,
'Artist': 32,
'Year': 19,
'https://t.co/pnDV9MyCVY': 1,
'Alysha': 1,
'half': 39,
'mom': 37,
'worlds': 2,
'greatest': 10,
'friend': 30,
'creeps': 2,
'again': 45,
'@CocoM2Z': 1,
'https://t.co/8ZTKJXOn8v': 1,
'#fate': 1,
'Europe‰Ûªs': 1,
'net': 2,
'neutrality': 2,
'decided': 4,
'https://t.co/zgqa8d49qX': 1,
'https://t.co/5Qk8zDNPQq': 1,
'got': 172,
'_Ù÷Ä': 5,
'bought': 8,
'tickets': 21,
'Post': 3,
'Glamour': 1,
'Girls': 8,
'Cookie,': 1,
'Leicester': 1,
'https://t.co/BP7THGHF37!': 1,
'https://t.co/efU07V6cN3': 1,
'via': 88,
'@ents24': 1,
'@crowmatriarchy': 1,
'@suzuyin': 1,
'@breadfuck': 1,
'likeness': 1,
'purely': 1,
'disgusting.': 1,
'mean,': 4,
'honestly.': 1,
'Look': 9,
'yourself': 20,
'mirror': 5,
'mor': 1,
'Hippocampus': 1,
'Long-term': 1,
'Memory': 1,
'https://t.co/SRCXtbYrMC': 1,
'#ArtistOfTheYear': 58,
'\n\nFrom': 2,
'four': 10,
'albums': 2,
'heard': 19,
'MM': 1,
'fav': 9,
'could': 66,
'changed': 8,
'rly': 1,
'fast': 8,
'hear': 23,
'fift': 1,
```

```
'anyone': 21,
'_Ù÷â': 71,
'@LindsayLkin': 1,
'know': 177,
'squirrel': 1,
'emoji.': 1,
'Oh': 14,
'iPhone': 4,
'family': 47,
'_Ù�À_Ù÷â': 1,
'@tombertram91': 1,
'NO': 25,
'WAY!': 1,
"You're": 28,
'opening': 4,
'act,': 1,
"aren't": 18,
'you?!': 1,
'chronicles': 1,
'Emerland.': 1,
'Solitaire.': 1,
'Games': 3,
'Adventure': 2,
'Mac': 4,
'App': 2,
'*****': 2,
'$4.': 1,
'‰Û_:': 1,
'https://t.co/xCJnfYaPvw': 1,
'https://t.co/jRlgh3fNSg': 1,
'Coming': 4,
'worst': 28,
'attitude': 5,
'while': 40,
'trying': 45,
'help': 49,
'_Ù÷Õ': 20,
'also': 46,
'where': 69,
"steven's": 1,
'interviews?': 1,
'emily': 2,
'did': 73,
'7': 20,
'million': 3,
'beth': 1,
'died': 13,
'actors': 4,
'afterwards': 1,
'@bethneilson12': 1,
'happy': 63,
'Beth!!': 1,
'Hope': 10,
'_ÙÕÏ_ÙÕÏ_ÙÕÏ_ÙÕÏ': 1,
'sweetest': 4,
'ever': 93,
'And': 260,
'two': 60,
'night.': 5,
'https://t.co/Nms52Uverj': 1,
'@SavageJaspy': 1,
'purple': 3,
'seem': 13,
'green': 10,
'person?': 1,
'saying': 30,
'things': 63,
'game': 27,
```

```
'doing': 46,
'https://t.co/9ZrbUY19FY': 1,
"kid's": 1,
'party': 24,
'were': 84,
'ones': 25,
'fun': 32,
'balloons.': 1,
'any‰Û_': 1,
'https://t.co/0mtdf9CmGr': 1,
'Now': 20,
'already': 39,
'Appreciate': 1,
'well': 44,
'Congrats!!!': 1,
'Enjoy': 4,
'GC': 1,
'phone!!!': 1,
'@thedomesticexec': 1,
'@dmthoma': 1,
'#TreatYourFamily': 5,
'@RIGHTWERK': 1,
'@musicnews_facts': 1,
"isn't": 27,
'she': 161,
'hated': 3,
'wearing': 17,
'suit': 1,
'Remember': 5,
'esteban': 1,
'calling': 6,
'dead': 19,
'hold': 12,
'https://t.co/XJ1Tjd8d5z': 1,
'Do': 20,
'americans': 1,
'find': 44,
'weird': 13,
'eat': 33,
'fries': 5,
'mayo': 1,
'cos': 4,
'do': 226,
'ur': 39,
'advisor': 2,
'classes': 9,
'semester': 3,
'year': 55,
'behind': 13,
'_ÙªÄ': 19,
'loving': 3,
'harrypotter': 1,
'Sky': 2,
'channel': 3,
'‰Ï¬_ÙÓ¨': 1,
'Last': 6,
'teacher': 14,
'said': 66,
'HBCU': 1,
"doesn't": 45,
'accurate': 3,
'perception': 2,
'what': 275,
'ican': 1,
'turn': 20,
'damn': 29,
'w/o': 3,
'havin': 1,
```

```
'KNs': 1,
'#PushAwardsKathNiels': 2,
'@mdeedoubleyou': 1,
'cry.': 2,
'dropping': 2,
'move': 23,
'Cali.': 1,
'‰Ïö•ü': 1,
'@Bekkimarshall3': 1,
'Haaa': 1,
'lipstick': 5,
"you'll": 19,
'perfect': 27,
'pout': 1,
'Holiday': 4,
'Candy': 2,
'Course:': 1,
'Learn': 9,
'art': 16,
'gourmet': 2,
'candy': 6,
'making': 32,
'make': 132,
'chocolates': 1,
'own': 38,
'kitchen!': 1,
'https://t.co/j1BKKeIqBA': 1,
'Starting': 2,
'Conner': 1,
'little': 77,
'#CoronationStreet': 1,
'‰ÛÏThat‰Ûªs': 1,
'drop': 8,
'em‰Û': 1,
'#BigTits': 1,
'#Boobs': 1,
'#Tits': 1,
'https://t.co/Wv3mjvXxd2': 1,
'https://t.co/H8p2ExMWqX': 1,
'working': 25,
"out&amp;amp;I'm": 1,
'exhausted': 4,
'breaking': 9,
'any': 42,
'sec': 2,
'listening': 18,
'Hall': 1,
'Fame': 1,
'always': 100,
'gives': 11,
'kick': 7,
'go': 184,
'on_ÙÕ»_Ù_': 1,
'#Mens': 1,
'#Shirt': 1,
'#BreakingBad': 1,
'WALT': 1,
'"I': 19,
'AM': 8,
'THE': 129,
'DANGER': 1,
'WHITE': 3,
'#Tshirt': 1,
'SIZE:': 1,
'MEDIUM': 1,
'**NEW**': 1,
'https://t.co/Ds5Zl3iODk': 1,
'#Bestseller': 1,
```

```
 'rest': 26,
 '@KAtheIntellect': 1,
 'easy': 9,
 'af': 11,
 'pencil': 1,
 'huh': 1,
 'Preparing': 1,
 'praying': 5,
 'best!': 3,
 '@scotthoying': 1,
 '"WE': 1,
 'COULD': 1,
 'BE': 17,
 'KINGS': 1,
 'OF': 25,
 'WORLD"': 1,
 '_Ù÷__Ù÷__Ù÷_': 12,
 'hi': 16,
 '@Harry_Styles': 29,
 ',thank': 4,
 'being\nso': 4,
 'nice.': 5,
 'fave': 10,
 'even\nthe': 4,
 'artist': 20,
 "couldn't": 24,
 'draw': 6,
 '‰ªÁ': 6,
 '‰ÑÎ.\nfollow': 4,
 'me,': 29,
 'please?': 6,
 '‰ÛÓ79,605': 1,
 'thinks': 10,
 'n': 12,
 'word,': 2,
 '@dinahfatty': 1,
 "what's": 27,
 'good?': 2,
 'old...I': 1,
 'F5...i': 1,
 'rehaul...I': 1,
 'Vanneka': 1,
 ...}
```

In [22]:

```
total_no_words=sum(number_of_words.values())
```

In [23]:

```
female.shape
```

Out[23]:

```
(6700, 26)
```

In [24]:

```
average_number_words_female=total_no_words/female.shape[0]
```

In [ ]:

```

```

In [25]:

```
number_of_words={}
for i in range(male.shape[0]):
```

```
            count=number_of_words[j]
            number_of_words[j]=count+1
        except:
            number_of_words[j]=1

number_of_words
```

Out[25]:


```
{'Robbie': 1,
 'E': 2,
 'Responds': 1,
 'To': 37,
 'Critics': 1,
 'After': 12,
 'Win': 3,
 'Against': 3,
 'Eddie': 2,
 'Edwards': 4,
 'In': 66,
 'The': 616,
 '#WorldTitleSeries': 1,
 'https://t.co/NSybBmVjKZ': 1,
 '‰ÛÏIt': 2,
 'felt': 5,
 'like': 319,
 'they': 251,
 'were': 80,
 'my': 536,
 'friends': 31,
 'and': 3651,
 'I': 1550,
 'was': 359,
 'living': 11,
 'the': 4248,
 'story': 22,
 'with': 498,
 'them‰Û': 1,
 'https://t.co/arngE0YHNO': 1,
 '#retired': 1,
 '#IAN1': 2,
 'https://t.co/CIzCANPQFz': 1,
 'i': 251,
 'absolutely': 7,
 'adore': 2,
 'when': 199,
 'louis': 6,
 'starts': 15,
 'songs': 16,
 'it': 611,
 'hits': 4,
 'me': 418,
 'hard': 21,
 'but': 265,
 'feels': 11,
 'good': 134,
 'Hi': 15,
 '@JordanSpieth': 1,
 '-': 314,
 'Looking': 13,
 'at': 386,
 'url': 3,
 'do': 226,
 'you': 1125,
 'use': 44,
 '@IFTTT?!': 1,
 '': 603,
 "Don't": 31,
 'typically': 1,
```

```
 'user': 4,
 'on': 758,
 '@PGATOUR!': 1,
 'https://t.co/H68ou5PE9L': 1,
 'Gala': 5,
 'Bingo': 4,
 'clubs': 7,
 'bought': 11,
 'for': 982,
 'å£241m:': 4,
 "UK's": 5,
 'largest': 20,
 'High': 8,
 'Street': 7,
 'bingo': 4,
 'operator,': 4,
 'Gala,': 4,
 'is': 940,
 'being': 107,
 'taken': 10,
 'over': 98,
 'by‰Û_': 5,
 'https://t.co/HzeeykJUd3': 1,
 '@coolyazzy94': 1,
 'Ditto': 2,
 "I'm": 272,
 'still': 123,
 'learning': 8,
 'favourites': 1,
 'retweet': 4,
 'stuff': 19,
 'least': 16,
 'sucks': 4,
 'less': 20,
 'than': 88,
 'Facebook': 10,
 'haha': 8,
 ':P': 1,
 '@CaribBros': 1,
 '@JstSaleem': 1,
 "don't": 185,
 'understand': 10,
 'how': 154,
 'to': 1902,
 'get': 296,
 'server': 1,
 'YALL': 1,
 'LMFAOO': 1,
 'RIGHT': 2,
 'WHEN': 3,
 'THE': 94,
 'CHORUS': 1,
 'CAME': 1,
 'ON,': 1,
 'A': 93,
 'TEAR': 1,
 'ROLLED': 1,
 'DOWN': 3,
 'HIS': 2,
 'FACE': 3,
 'https://t.co/aYuQDPtvsE': 1,
 'James': 17,
 'Bond': 17,
 'premier': 1,
 'night': 31,
 '@Everymancinema': 1,
 'in': 1062,
 'Oxted': 1,
 '@SidiEdey_': 1,
```

```
 'up': 271,
 'expectation!': 1,
 '#SPECTRE': 7,
 'As': 19,
 'opposed': 3,
 'Pump': 2,
 'where': 54,
 "it's": 179,
 'HI': 1,
 'HOPE': 2,
 'YOU': 25,
 'LIKE': 4,
 'DOING': 1,
 'JUMPS': 1,
 'WHERE': 2,
 'SPREAD': 1,
 'YOUR': 12,
 'FEET': 1,
 'ACROSS': 1,
 'ENTIRE': 1,
 'STAGE': 1,
 'All': 45,
 '#magic': 1,
 'Hath': 1,
 'No': 42,
 'FURY': 1,
 'based': 13,
 'REAL': 3,
 '#Magick!': 1,
 'https://t.co/jwpsVhAU1E': 1,
 'And': 227,
 'got': 136,
 'more': 145,
 'yards': 2,
 'AND': 67,
 'points': 8,
 'Jets': 4,
 'gave': 13,
 'all': 314,
 'season.': 8,
 'https://t.co/gdfkaOxcDD': 1,
 'Did': 15,
 'Alot': 2,
 'Up': 5,
 'Past': 2,
 'Ion': 1,
 'Wont': 1,
 'Back': 6,
 '@TheRiddler109': 1,
 '@CNN': 3,
 'mean': 32,
 'not': 315,
 'Mainstream': 1,
 'new': 179,
 'media': 15,
 'supposed': 6,
 'feed': 4,
 'fact': 17,
 'nowadays...': 1,
 'How': 61,
 'many': 44,
 'followers': 30,
 'everyday?': 3,
 '1': 63,
 'last': 102,
 'day.': 30,
 'Growing': 9,
 'daily': 10,
 'https://t.co/IzckR3ub8H': 1,
```

```
'real': 19,
'#AMAs\n\nHO': 1,
'FAME': 1,
'@kbonimtetezi': 1,
'mheshimiwa': 1,
'travellers': 1,
'along': 17,
'that': 556,
'stretch': 1,
'road': 14,
'(lubao)': 1,
'r': 8,
'hurting': 1,
'nobody': 8,
'seems': 12,
'be': 485,
'raising': 3,
'this': 364,
'issue!': 1,
'Greenville': 1,
'Thursday.': 1,
'Yall': 2,
'holla': 1,
'need': 118,
'anything.': 4,
'Know': 6,
'plug': 4,
'@AndyRobsonTips': 1,
'Cardiff': 1,
'or': 136,
'drew': 2,
'1.5': 1,
'match': 9,
'goals': 10,
'sounds': 11,
'https://t.co/XprInke1m2': 1,
'Best': 18,
'thing': 68,
'about': 228,
'having': 29,
'audition': 1,
'west': 1,
'side': 22,
'able': 14,
'eat': 33,
'lunch': 3,
'Komodo': 1,
'Discipline': 1,
'bridge': 3,
'between': 52,
'accomplishment.': 1,
'Jim': 5,
'Rohn': 1,
'20': 16,
'minutes': 26,
'Wednesday': 7,
'fan': 18,
'has': 138,
'tweeted': 2,
'us': 43,
'3': 49,
'times,': 2,
'no,': 5,
'obsessed,': 1,
'even': 79,
'a': 1645,
'little': 33,
'bit': 21,
'ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ ÙÑâ': 1,
```

```
'conversion': 5,
'App': 4,
'https://t.co/obCkhxkOm7': 1,
'@Hakeem_NLT': 1,
'Boii': 1,
'make': 117,
'wafflw': 1,
'link': 12,
'reference': 1,
'then': 106,
'word': 23,
'count': 7,
'should': 71,
'clear': 6,
'lol': 82,
'@Harry_Styles': 35,
'always': 82,
'remember,': 4,
'never': 67,
'forget': 17,
"you're": 68,
'best': 122,
'greatest': 17,
'world': 63,
'just': 361,
'till': 17,
'end': 39,
'time': 165,
'POLITICS': 1,
'evidence': 6,
'+': 12,
'results': 5,
'international': 4,
'development.': 1,
'https://t.co/T7N2c2TTBf': 1,
'Looks': 7,
'great.': 6,
'h/t': 2,
'@duncan_ids': 1,
'@rosalindeyben': 1,
'(Surgical': 1,
'enhancements': 2,
'are': 397,
'different': 15,
'part': 26,
'body,': 1,
'and,': 6,
'kidding...)': 1,
'swear': 9,
'if': 171,
'she': 65,
'touches': 1,
'one': 189,
'time.': 27,
'_Ù÷__Ù÷_': 4,
'pisses': 2,
'off.': 8,
"It's": 84,
'stupid': 16,
'cause': 26,
'fall': 12,
'trick': 2,
'as': 241,
'well': 62,
'_Ù÷Ô': 4,
'She': 14,
'beautiful': 30,
'm,': 2,
'talking': 20,
```

```
 '@LOVSIZWorldwide': 1,
 'Say': 2,
 'He': 55,
 'one.': 10,
 'eternal.': 1,
"doesn't": 40,
 'have': 462,
 'kids': 29,
 'nor': 2,
 'did': 79,
 'anyone': 18,
 'Him.': 1,
 'There': 20,
 'nothing': 40,
 'One.': 1,
 'Now': 24,
 'Changed!': 1,
 'same': 89,
 'Man': 16,
 'once': 18,
 'Yesterday': 2,
 'better': 54,
 'watch': 43,
 '&amp;': 133,
 'take': 63,
 'notes': 7,
 '@TekifyUK': 1,
 'am': 65,
 'trying': 35,
 'unbrick': 2,
 'Kindle': 1,
 'fire,': 2,
 'red': 14,
 'screen.': 1,
 'followed': 16,
 'your': 372,
 'youtube': 3,
 'video,': 1,
 'using': 25,
 'tool.': 1,
 'If': 120,
 'say': 75,
 'someone': 51,
 'Wale': 2,
 'Meek': 2,
 'Wiz': 1,
 'gotten': 5,
 'big': 42,
 'past': 39,
 'five': 11,
 'years': 35,
 'moron.': 1,
 'They': 42,
 'mixtape': 4,
 'game.': 11,
 'Reimagining': 2,
 '#webdesign': 2,
 'process': 7,
 'by': 254,
 '@InVisionApp': 2,
 'https://t.co/Vmb0OZU67e': 2,
 'https://t.co/hFlWR8tfOl': 2,
 '@deptulahasrage': 1,
 'Clemson': 1,
 'only': 135,
 'stack': 2,
 'line': 20,
 'against': 27,
 'FSU': 1,
```

```
'Chris': 1,
'chan': 1,
'fame': 4,
'Im': 23,
'weakkkkk_Ù÷â_Ù÷â_Ù÷â_Ù÷â_Ù÷â\nTbh': 1,
'thats': 17,
'way': 92,
'shut': 11,
'down': 77,
'girls': 20,
'who': 135,
'flex': 2,
'Chris': 4,
'Got': 19,
'On': 27,
'Black': 12,
'Toe': 1,
'1s': 3,
'_Ù÷_': 22,
'@paulmgardner': 1,
'@cwellssun': 1,
'knew': 14,
'guy': 34,
'military': 6,
'used': 23,
'because': 73,
'show': 44,
'drug': 7,
'tests.': 1,
'My': 86,
'mom': 11,
'listening': 11,
'"Hello"': 3,
"what's": 11,
'happening': 4,
'Total': 1,
'loss': 6,
'respect': 14,
'twat': 3,
'https://t.co/eQstpYJEck': 1,
'he': 223,
'needs': 24,
'his': 174,
'own': 28,
'intro': 3,
'pick': 15,
'phone': 23,
"'CLIFFORD": 1,
'ELLIS,': 1,
'GREATEST': 1,
'BROCKER': 1,
'IN': 27,
"WORLD'": 1,
'lmfao': 7,
'Fortune': 1,
'favors': 1,
'funny': 19,
'Guatemala': 1,
'election:': 1,
'https://t.co/E1Qa16AweK': 1,
'https://t.co/c8nvJn3VaN': 1,
'@megwacha': 1,
'Actually': 5,
'UC': 1,
'announcement': 2,
'could': 80,
'related': 2,
'that.': 15,
'Stats': 24,
```

```
'via': 98,
'https://t.co/VuMDKp70A7.': 1,
"Kaito's": 1,
'song': 34,
'chasing': 3,
'after': 59,
'idealized': 1,
'version': 5,
'himself,': 2,
'real': 44,
'world,': 5,
"he's": 48,
'alone.': 3,
'#starmyu': 1,
'https://t.co/TvLbxrYMRK': 1,
'Photo:': 8,
'Snow': 5,
'Queen': 5,
'Vlad': 1,
'Barbe': 1,
'|': 30,
'Kids': 3,
'Family': 4,
'|716822518:': 1,
'fairy': 2,
'tale': 1,
'Hans': 1,
'Christian...': 1,
'https://t.co/CC4qqNHvjY': 1,
'@bentorkington': 1,
'@NZInlandRevenue': 1,
"...don't": 1,
'bother': 3,
'KiwiSaver': 1,
'until': 26,
'kickstart': 1,
'fund': 4,
'restored': 1,
'"Came': 1,
'into': 76,
'game': 79,
'white': 17,
'rappers': 4,
"weren't": 3,
'cliche"\n\n@classified': 1,
"spittin'": 1,
'truth': 12,
'Coffee': 1,
"Taster's": 1,
'Flavor': 1,
'Wheel': 1,
'from': 218,
'SCAA\n#coffee\nhttps://t.co/GOLukszOLU': 1,
'https://t.co/HJHczkaiQy': 1,
'We': 65,
'chance': 16,
'first': 81,
'decades': 1,
'elect': 1,
'will': 229,
'owned': 5,
'lobby': 1,
'groups': 1,
'corporations.': 1,
'Or': 13,
'Saudi': 6,
'Arabia': 2,
'saw': 25,
'met': 14,
```

```
'today.': 17,
'forever': 6,
'grateful': 1,
'opportunity': 5,
'given.': 1,
'_ÙÕ¥_ÙÕ_': 1,
'https://t.co/vLrCSnqN2I': 1,
'things': 37,
'heated': 2,
'#Pennsylvania': 1,
'Believing': 1,
'yourself': 15,
'secret': 11,
'success!': 1,
'https://t.co/9sIPL3VcxE': 1,
'Corporate': 3,
'welfare': 3,
'worst.': 2,
'Dark': 10,
'funds': 1,
'Military': 3,
'industrial': 1,
'complex': 1,
'another.': 1,
'\n\nAll': 1,
'these': 35,
'posts': 4,
'with...': 2,
'https://t.co/awmdynzFxZ': 1,
'Rapper': 2,
'Tyga': 1,
'appear': 2,
'next': 69,
'season': 35,
'KUWTK': 1,
'Will': 20,
'Get': 34,
'$25k': 1,
'Per': 2,
'Appearance\n-': 1,
'See': 21,
'&gt;&gt;...': 1,
'https://t.co/b1Rw1LkjHG': 1,
'Nothing': 12,
'midwest.': 1,
'Had': 6,
'great': 68,
'weekend': 16,
"Ruffin's": 1,
'wedding.': 2,
'glad': 10,
'we': 225,
'came': 24,
'together‰Û_': 1,
'https://t.co/DrK2UkEKxV': 1,
'Mentally': 1,
'ill': 6,
'woman': 9,
'her': 90,
'two': 41,
'pink-dyed': 1,
'poodles': 1,
'threatened': 1,
'shoot': 6,
'someone,': 1,
'throwing': 3,
'1/2': 7,
'At': 21,
'Steidl': 1,
```

```
'#IRS': 1,
'state': 13,
'scary.': 1,
'can': 215,
'help:': 1,
'https://t.co/riC1oHcD1A': 1,
'Drugs,': 1,
'greed,': 1,
'kidnapping': 1,
'plague': 1,
'Dani': 2,
'Liz': 1,
'Bluewater': 2,
'Ganja,': 1,
'9th': 1,
'Thrillers': 1,
'#sailing': 1,
'https://t.co/60BrKidlAP': 1,
'ISIS': 3,
'Launches': 2,
'BIZARRE': 1,
'New': 27,
'Weapon': 1,
'War': 1,
'Infidel': 1,
'‰_Ê': 1,
'FLOATING': 1,
'CONDOM': 1,
'BOMBS': 1,
'(Video)': 1,
'BB4SP': 1,
'https://t.co/ffjxmiCEt7': 1,
'@Nataliaa_45': 1,
'@Caspar_Lee': 1,
'MY': 16,
'PHONE': 4,
'IPAD': 1,
'ARE': 11,
'GOING': 3,
'OUT': 10,
'WINDOW': 1,
'MINUTE': 1,
'OMG': 3,
'CANT': 2,
'BREATHE': 1,
'JOE': 2,
'FAVED': 1,
'CASPAR': 1,
'SAID': 3,
'THAT': 7,
'MET': 2,
'sky': 5,
'making': 32,
'weird': 8,
'ass': 29,
'noises': 2,
'MT': 1,
'@nytimesarts:': 1,
'longest-running': 1,
'play': 57,
'America': 11,
'set': 19,
'open': 22,
'NYC': 4,
'1st': 10,
'https://t.co/65oIQVi9M8': 1,
'https://t.co/6WIFY90anM': 1,
'apparently': 7,
'contract': 3,
```

```
'no': 139,
'contracts': 2,
'phone.': 2,
'.@verizon': 1,
'One': 43,
'You': 131,
'Fall': 3,
'Love': 22,
'With': 24,
'Should': 11,
'Feel': 5,
'Like': 15,
'Nothin': 1,
'Else': 1,
'Matters..': 1,
'#Fr': 1,
'Selling': 2,
'#Recruiting': 1,
'Process': 2,
'Isn‰Ûªt': 1,
'Gamble': 1,
'https://t.co/Qbsl8ix3dy': 1,
'#HR': 3,
'#TheCandEs': 1,
'#TChat': 1,
'#HRTechConf': 3,
'https://t.co/Dzu9HX59Yb.': 1,
'October': 12,
'Apostrophe': 1,
'https://t.co/Mhng687f9w': 1,
'#publishing': 1,
'#authors': 1,
'720p': 1,
'option': 6,
'Fantastic': 7,
'launch': 6,
'@Snapdeal': 3,
'apps': 7,
'now': 159,
'run': 24,
'smoothly': 3,
'OnePlus': 3,
'thanks': 35,
'quad': 4,
'core': 4,
'processor!': 3,
'https://t.co/2yMa7R3PcU': 3,
'4742': 1,
'7': 21,
'unfollowers': 4,
'5': 46,
'(hello!': 1,
'hello!)': 1,
'week.': 14,
'Via': 4,
'https://t.co/WL2CRq09dV': 1,
'Except': 3,
'couple': 15,
'broken': 4,
'plays,': 5,
'really': 102,
'job': 25,
'countering': 1,
'what': 199,
'England': 4,
'does': 37,
'O': 3,
'exploiting': 1,
'D': 4,
```

```
'Barrow': 3,
'putting': 8,
'Audrey': 4,
'Roberts': 2,
'old': 40,
'rascal!': 1,
'Be': 14,
'Lookout!!!': 1,
'_ÙÔà_ÙÔà_ÙÔà': 1,
'https://t.co/dfZqv6Kjbh': 1,
'@NBA': 13,
'hyped': 4,
'bc': 14,
'im': 52,
'so': 316,
'excited': 16,
'offs!': 1,
"who's": 6,
'team': 36,
'win': 43,
"who'll": 1,
'star!_Ù�Û': 1,
'#ThisIsWhyWePlay': 12,
'#Sweepstakes': 8,
'Hey': 14,
'guys': 29,
"I've": 64,
'decided': 8,
"I'll": 71,
'maybe': 18,
'most': 73,
'each': 22,
'!': 22,
'@axeslasher': 1,
'@MetalShayne2000': 1,
'super': 11,
'hooked': 4,
'Woodland': 1,
'Tortuary': 1,
'talk.': 3,
'Hot': 7,
'presses!': 1,
'https://t.co/fZK5fjSWEJ': 1,
'@stloto': 1,
'any': 52,
'update': 5,
'site': 9,
'again?': 3,
'Thanks': 29,
'@WinkMartindale': 1,
'shout': 4,
'out!': 4,
'I‰Ûªm': 15,
'looking': 33,
'pretty': 31,
'75!': 1,
'#FamousName': 1,
'https://t.co/UMrFosWY0u': 1,
'fairly': 2,
'normal': 4,
"could've": 2,
'died': 9,
'other': 74,
'tho': 9,
'Jewish': 3,
'Voice': 4,
'Peace': 1,
'"': 16,
'Analysis': 1,
```

```
'which': 57,
'"dehumanize': 1,
'Palestinians': 2,
'and...': 9,
'https://t.co/fTKu2RMQN1': 1,
'tonight': 17,
'hit': 37,
'39': 1,
'total': 3,
'more,': 4,
'11': 7,
'$5': 1,
'scratch': 3,
'offs': 2,
'coming': 31,
'https://t.co/T7UOjvQqjT': 1,
'@JoseFranco_': 1,
'caller': 3,
'wins': 7,
'contest': 4,
'quickly': 3,
'joking': 2,
'tweets': 8,
'entirely': 1,
'too': 69,
'seriously': 4,
'#AMAs': 44,
'https://t.co/DsjN39Gzs1': 1,
'@CineBroughton': 1,
'hour': 19,
'bull': 2,
'ready': 20,
'#007': 3,
'DJ': 2,
'Polish': 1,
'Sausage': 1,
'follow!': 5,
'sample': 5,
'colors': 1,
'encountered': 2,
'during': 18,
'our': 89,
'hike': 1,
'Mount': 1,
'LeConte.': 1,
'This': 66,
'picture': 17,
'was‰Û_': 1,
'https://t.co/qlz7fookM4': 1,
'Tell': 7,
'@SenateDems': 1,
'stand': 9,
'Monsanto': 1,
'protect': 6,
'#GMO': 1,
'labeling.': 1,
'Sign': 4,
'petition:': 1,
'https://t.co/SL82I7iqh8': 1,
'#p2': 2,
'#RightToKnow': 1,
'#DARKAct': 1,
'@FitStar': 1,
'sent': 12,
'DM': 3,
'describing': 1,
'issue': 6,
'email': 4,
'address': 5,
```

```
'Amazon': 2,
'celebrates': 1,
'vinyl': 1,
'store': 9,
'unveiling': 1,
'‰Ûª13': 1,
'Day': 13,
'Vinyl‰Û_': 1,
'https://t.co/OlKFmaMmtC': 1,
'#EDM': 3,
'#Followback': 1,
'https://t.co/Z0cq8di0sp': 1,
'Rose': 1,
'without': 20,
'restriction': 1,
'opener:': 1,
'Chicago': 4,
'Bulls': 2,
'era': 3,
'under': 26,
'coach': 12,
'F...': 1,
'https://t.co/6ec5nDJw9W': 1,
'Colt': 2,
'here!': 4,
'sure': 31,
'Nicole': 3,
'group!': 1,
'10/27': 3,
'start': 38,
'fun': 22,
'games': 25,
'Release': 2,
'RECKLESS!...': 1,
'unfollower': 2,
'https://t.co/K1SASMDLEB': 1,
'@PrestigeDiesels': 1,
'Yeah...': 1,
'had': 114,
'nail': 3,
'mat': 1,
'floor!': 2,
'@LeahRebeccaUK': 1,
'ADA': 1,
'you.': 32,
'regulation': 1,
'affect': 2,
'office?': 1,
'#askHCSI': 1,
'https://t.co/i3CLomKeMd': 1,
'Would': 14,
'‰Û÷The': 2,
'Affair‰Ûª': 1,
'Easier': 1,
'Watch': 15,
'Weren‰Ûªt': 1,
'Such': 2,
'Brats?': 1,
'[Spoilers]:': 1,
'Showtime': 1,
'series': 6,
'Affa...': 1,
'https://t.co/PZU4G93AdO': 1,
'@maxschrems': 1,
'check': 19,
'out': 276,
'work': 66,
'‰ÛÏconsent': 1,
'receipt‰Û': 1,
```

```
 analysis' : 2,
 ...}
```

In [26]:

```
total_no_words=sum(number_of_words.values())
```

In [27]:

```
male.shape
```

Out[27]:

```
(6194, 26)
```

In [28]:

```
average_number_words_male=total_no_words/male.shape[0]
```

In [29]:

```
print("The average number of words used by FEMALE in thier tweet is:",average_number_word
s_female)
print("The average number of words used by MALE in thier tweet is:",average_number_words_
male)
```

```
The average number of words used by FEMALE in thier tweet is: 15.850298507462687
The average number of words used by MALE in thier tweet is: 16.237003551824348
```

Ensemble Machine learning Modelling

assign independent and dependent variables. Here gender is the dependent variable

In [30]:

```
X = df[['_trusted_judgments','gender:confidence',
       'profile_yn:confidence', 'fav_number',
       'retweet_count','tweet_count']].values
Y =df[['gender2']].values
```

TESTING AND SPLITTING THE DATASET

In [31]:

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y)
```

KNN ALGORITHM

finding the knn algorithm accuracy

In [32]:

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
```

```
C:\Users\Shreyas Venishetty\anaconda3\lib\site-packages\ipykernel_launcher.py:3: DataConv
ersionWarning: A column-vector y was passed when a 1d array was expected. Please change t
he shape of y to (n_samples, ), for example using ravel().
  This is separate from the ipykernel package so we can avoid doing imports until
```

Out[32]:

In [33]:

```
from sklearn import metrics
y_pred = knn.predict(X_test)
print("Test set Accuracy: ",metrics.accuracy_score(Y_test, y_pred))
```

```
Test set Accuracy:  0.538151364764268
```

SUPPORT VECTOR MACHINE ALGORITHM

svm algorithm accuracy

In [34]:

```
from sklearn.svm import SVC
svc = SVC(kernel='rbf')
# training Linear Regression model on training data
svc.fit(X_train, Y_train)# The coefficients
```

```
C:\Users\Shreyas Venishetty\anaconda3\lib\site-packages\sklearn\utils\validation.py:760:
DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please
change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

Out[34]:

```
SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
    max_iter=-1, probability=False, random_state=None, shrinking=True,
    tol=0.001, verbose=False)
```

In [35]:

```
y_pred = svc.predict(X_test)

from sklearn.metrics import classification_report, confusion_matrix
print ("TEST ACCURACY:",metrics.accuracy_score(Y_test, y_pred))
```

```
TEST ACCURACY: 0.5186104218362283
```

RANDOM FOREST ALGORITHM

finding accuracy of random forest algorithm

In [36]:

```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier()
# training Linear Regression model on training data
rfc.fit(X_train, Y_train)
```

```
C:\Users\Shreyas Venishetty\anaconda3\lib\site-packages\ipykernel_launcher.py:4: DataConv
ersionWarning: A column-vector y was passed when a 1d array was expected. Please change t
he shape of y to (n_samples,), for example using ravel().
  after removing the cwd from sys.path.
```

Out[36]:

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
```

In [37]:

```
y_pred = rfc.predict(X_test)
from sklearn.metrics import classification_report, confusion_matrix
print ("TEST ACCURACY:",metrics.accuracy_score(Y_test, y_pred))
```

```
TEST ACCURACY: 0.5384615384615384
```

ACCURACY OF THREE ALGORITHM

from the above we find each algorithm accuracy and listed down

accuracy of each algorithm

1.KNN :53%

2.SVM :52%

3.RANDOM FOREST :54%

Comparing the Accuracy of all three, the ML algorithms suits best for the given problem is RANDOM FOREST ALGORITHM

In [ ]:

```

```