Lab - 5

PRML
AY 2020-21, Trimester - III

Name : Akshaykumar Kanani(B9EE008)

<u>ANS :</u>

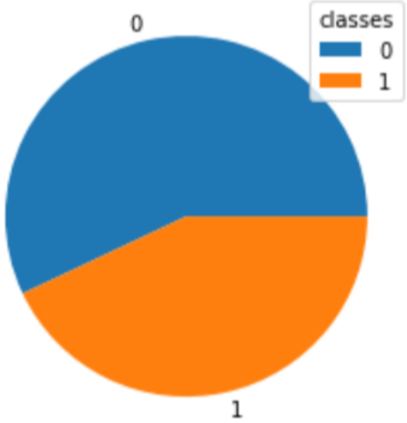
a) Data preparation:

We load data with the help of pandas.

And we get the count of each target as:

0 - 4342
1 - 3271

And from the pie chart, we can see that there is not much difference.

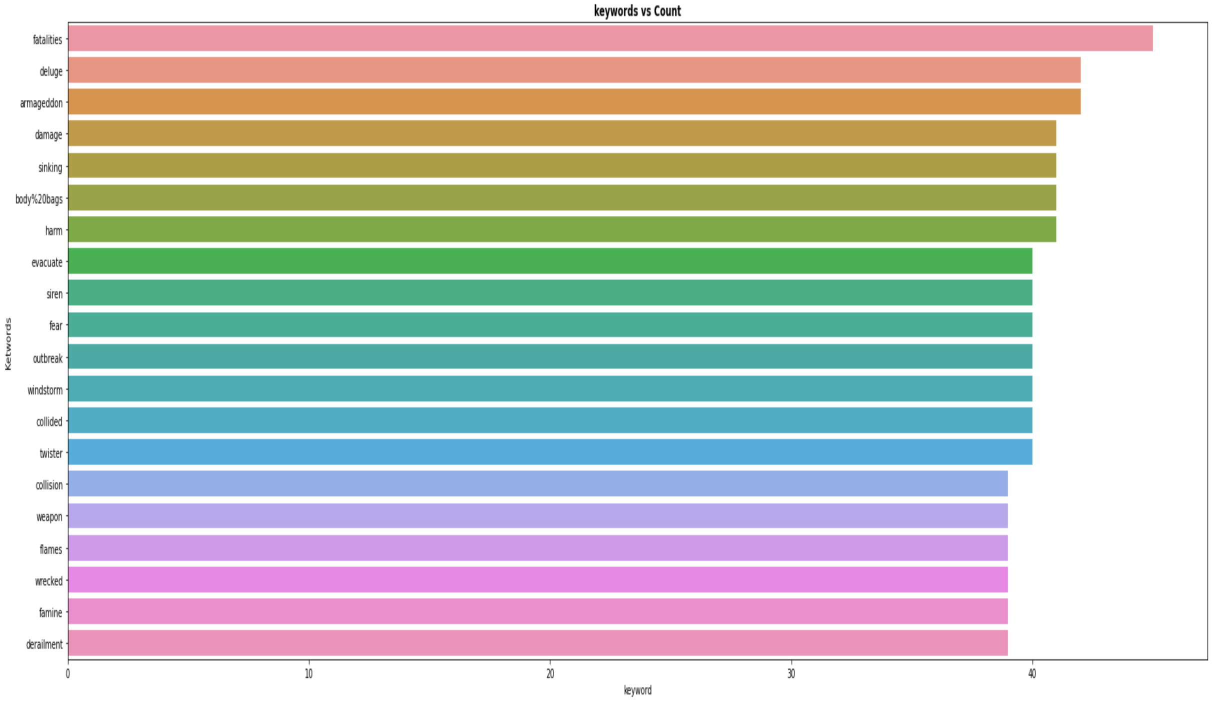


And from the data we get first 20 unique key word with frequency as :

fatalities	45
deluge	42
armageddon	42
damage	41
sinking	41
body%20bags	41
harm	41
evacuate	40
siren	40
fear	40
outbreak	40
windstorm	40
collided	40

twister	40
collision	39
weapon	39
flames	39
wrecked	39
famine	39
derailment	39

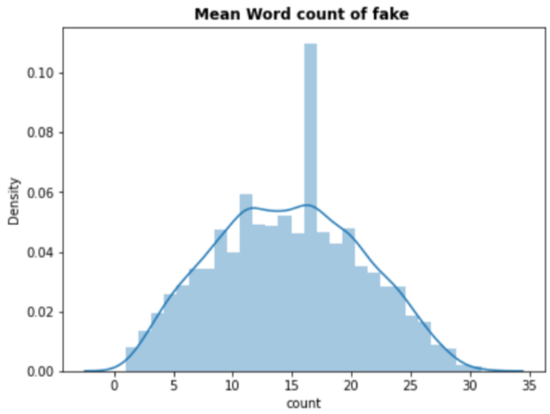
And we get the plot of count for this key words as:

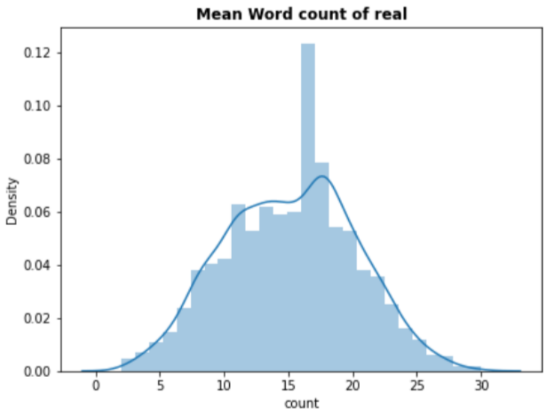


Visualize the correlation of the length of a tweet with its target- we get:

Mean length of text column = 101.03743596479706
mean of the count of words of fake class = {14.704744357438969}
mean of the count of words of real class = {15.167532864567411}

And graph as





From the graph and the means, we can clearly see that it is not specifically possible to be at any conclusion only on the basis of the length of the text. As the mean of the count of words are also of the same level.

Print the null values in a column

We got the number of null values in each columns as:
location   2533
keyword   61

Removing null values

from the above result its clear that there is no null value in text and target row so we don't need to remove this because if we remove this data we left with small traning data set without any good so i think we not need to detele null data

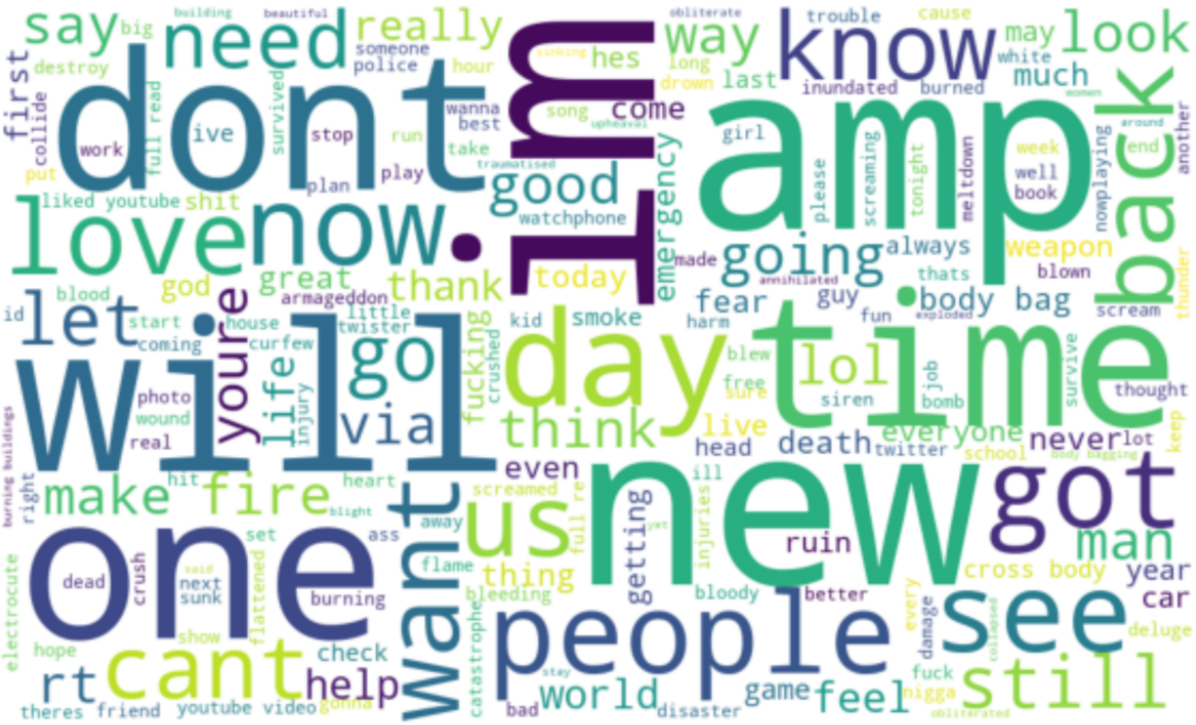
Removing Double Spaces, Hyphens and arrows, Emojis, URL, another   Non-English or special symbol

Replace wrong spellings with correct ones

This all part are done in code file with the help of demo file provided in the class during the lab.

Plot a word cloud of the real and fake target

For fake tweet- we got:



For Real tweet:

new
fire
amp
us
will
via
storm
flood
one
death
suicide bomber
police
near
area
two
today
now
man
flooding
burning buildings
emergency
people
world
city
rt
building
attack
look
oil spill
year
car
damage
im
back
dont
day
say
another
homes razed
going
hiroshima
time
northern california
terrorist
california wildfire
pkk suicide
suicide bombing
families affected
detonated bomb
confirmed mh
evacuate
mass murder
families sue
bomber detonated
severe thunderstorm
latest homes
declares disaster
old pkk
obama declares
please
report
rain
call
debris found
mh malaysia
atomic bomb
hit
road
video
fatal outbreak
life
yr old
first
razed northern
weapon
got
update
help
see
dead
go
legionnaires
sue legionnaires
st
way
affected fatal
evacuation
wild fire
match airport
refugio oil
police officer
officer wounded
may costlier
heat wave
drought
india
minute watch
crash
spill may
bombing
rioting

Remove all columns except text and target
Split data into train and validation
Done in the code file.

B part done on the code file.
C) Find the frequency of words in class 0 and 1.

For class 0 top 4 words with frequency:
'the': 1931, 'i': 1474, 'a': 1265, 'to': 1221……………..

For class 1 top 4 words with frequency:
'the': 1412, 'in': 1169, 'a': 944, 'of': 935,...................

D) Does the sum of the unique words in target 0 and 1 sum to the total number of unique words in the whole document? Why or why not? Explain in the report.

Yes the sum of the unique words in the target 0 and 1 sum to the total number of unique words in the whole text column data because we count all the words from both the class 0 and 1 and we know that there are only two class 0 and 1 without any NULL value so we can say that C0+C1=Ct. Thats why we got this result.

E and F part) Most of the work is done on then code file. Some result we got as:
likelihood_prob_0
0.9999999999997653

likelihood_prob_1
1.0000000000001097

Prior_0
0.5694581280788177

Prior_1
0.43054187192118226

G) precision, recall and f1 score and confusion matrix

tp= 275 ,fp= 411 , tn= 463 , fn= 374

Actual/predicted	Predicted true	Predicted false
Actual true	275	374

Actual false	411	463

classwise accuracy= 0.4767385486560912
Total accuracy= 0.48456992777413
Precision : 0.4008746355685131
Recall : 0.423728813559322
f1-score : 0.41198501872659177

Thank you