FLIP ROBO

Fake News Prediction

Submitted by:

PAVANKUMAR BOLLI

ACKNOWLEDGMENT

I would like to express my special thanks to flip robo technlogies for given me such a wonderfull of apportunity to explore and insite work on a data science projects.

And and also thanks to Datatraiend institute's mentors to motivate and understand the concepts of Data science. I have got lot of help from their resourses and concept to complete these project.

Link that helps me to complete this project

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

INTRODUCTION

- ## Business Problem Framing

  The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate, or false information acquires a tremendous potential to cause real-world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed.

- ## Motivation for the Problem Undertaken

  Motivation of this project is to predict the news which is getting viral is real or fake. Now Days each and every minutes some news are getting and happening in the world. We will get lots of news each day but it becomes hard to verify which news are real or fake. So many people are to promote there product or to visit there sire the make fake news. To solve this i have made a model which can predict that news is real or fake on the basis of text present in the news.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Dataset contains 20800 observations with 6 columns. In this data there are some null values in written by and headlines columns. I have handled it by replacing null values with no author and no headlines. Rather dropping observation not losing the data i have imputed with some other text. Here two columns have to predict news is real are fake that is Headlines or News. But i have worked on News column. So model will create with large text data.

- Data Sources and their formats

  All features are in text and target column is in binary format i.e. 1/0

  There are 6 columns in the dataset. The description of each of the column is given below:

  "id": Unique id of each news article

  "headline": It is the title of the news.

  "news": It contains the full text of the news article

  "Unnamed:0": It is a serial number

  "written_by": It represents the author of the news article

  "label": It tells whether the news is fake (1) or not fake (0).

- ## Data Preprocessing Done

  In Data pre-processing I have worked on News column which is in text. While cleaning the data i have removed all the punctuations, numbers, email address, emogies and stop words. Stop words are those which does not affect much to predict. These are small words which will comes consistently in each sentence which makes data redundancy. After cleaning the data i've created new columns which contains the length of each news before cleaning and after cleaning. And also i have represented in graphical manner how much data is cleaned.

- ## Hardware and Software Requirements and Tools Used

  ListinTools and libraries that i have used to solve project:

  - Software: Anaconda, Jupyter Notebook, Python3
  - Libraries: Numpy, Pandas, Matplotlib, Seaborn, Sklearn.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Describe the approaches you followed, both statistical and analytical, for solving of this problem.

- Testing of Identified Approaches (Algorithms)

  Listing down all the algorithms used for the training and testing.

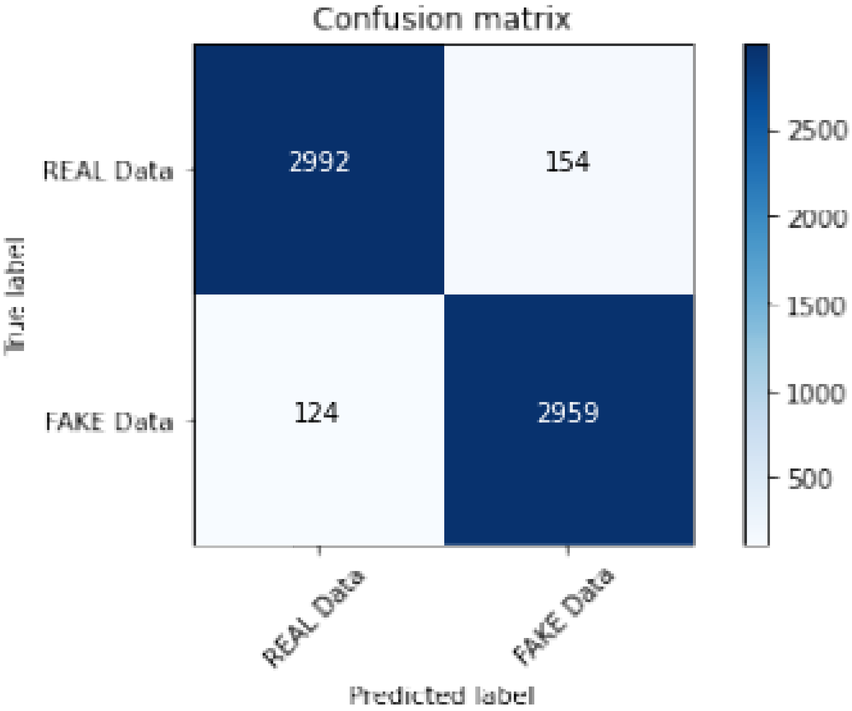
  The data set contain lable data and target data lable has cleaned by some EDA steps. Target data is in the form of classified i.e (Label). By seeing the our target column we can conclude that this problem comes under supervised and Classification Problem. Target column having in factor. For more testing Multinomial, Passive Aggressive Classifier Algorithm have used. I have tested both model by using Count Vectorizer and TFIDF

- Run and Evaluate selected models

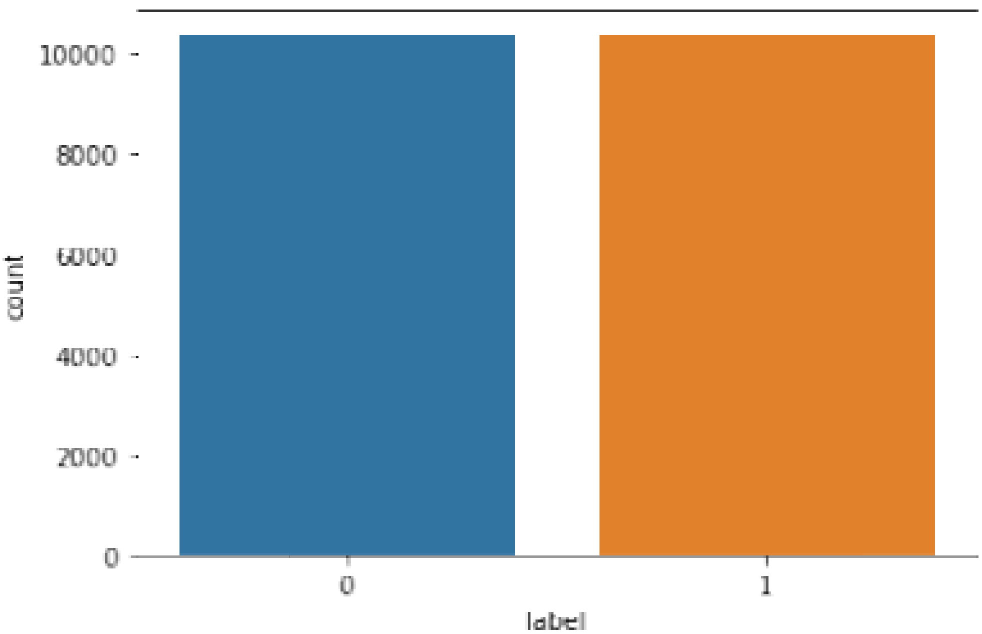
  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.
  After working on multinomial naive bayes algorithms with both tfidf and Count Vectorizer with hyper parameters using grid search c vi have got accuracy with 91% but large number of errors are coming in the section of Type 2. To solve this i have use another algorithm that is passive aggressive classifier i have got much better output with the 95% accuracy and also with a default parameters. And also i tried with bigram on this algorithm but accuracy has decreased.

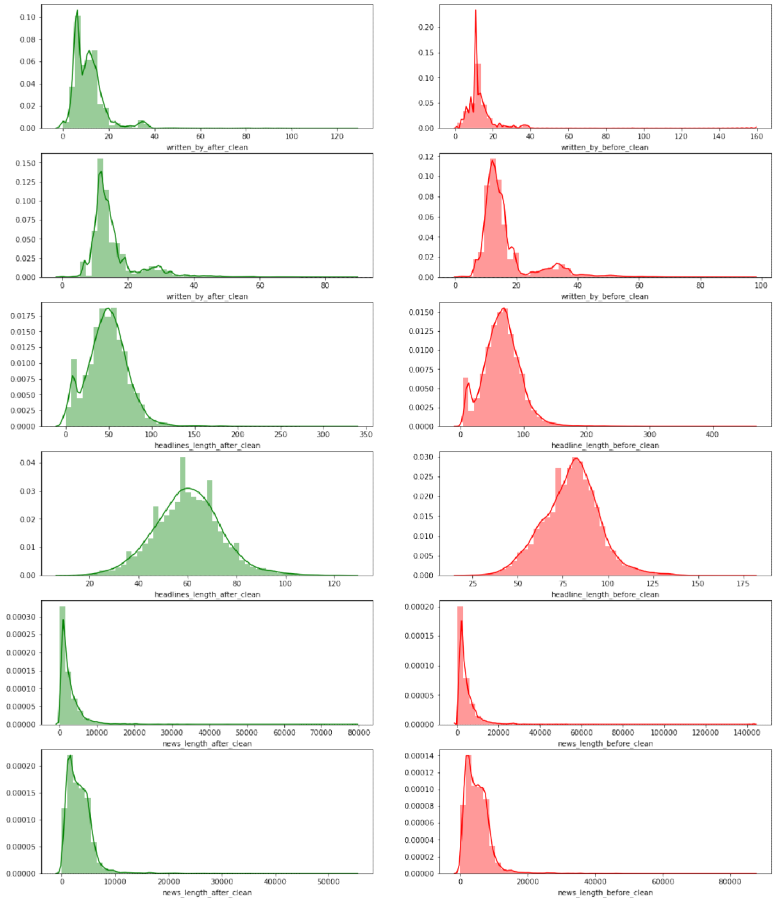
- Key Metrics for success in solving problem under consideration

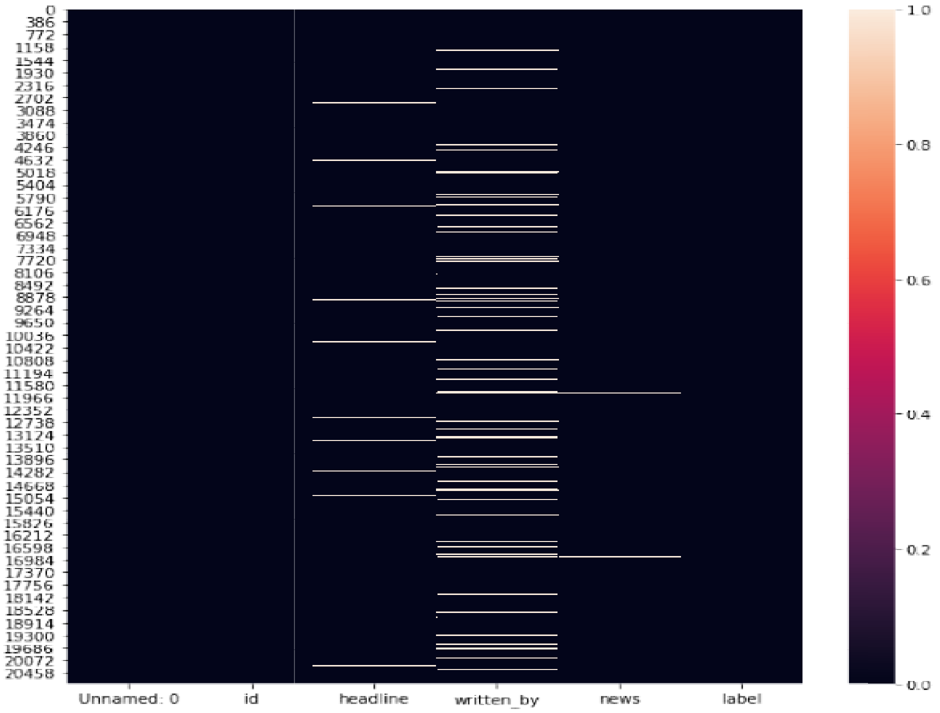
  For testing how good model is generated. I have splitted data set into two parts (70% 30%) and each part is splited into further two part that are lable data   and target data. First 70% splited data is used for training the model and another rest of 30% used for Prediction testing the model how good model is predicted in terms of percentage. For evaluation i have used confusion matrix, accuracy and classification report.

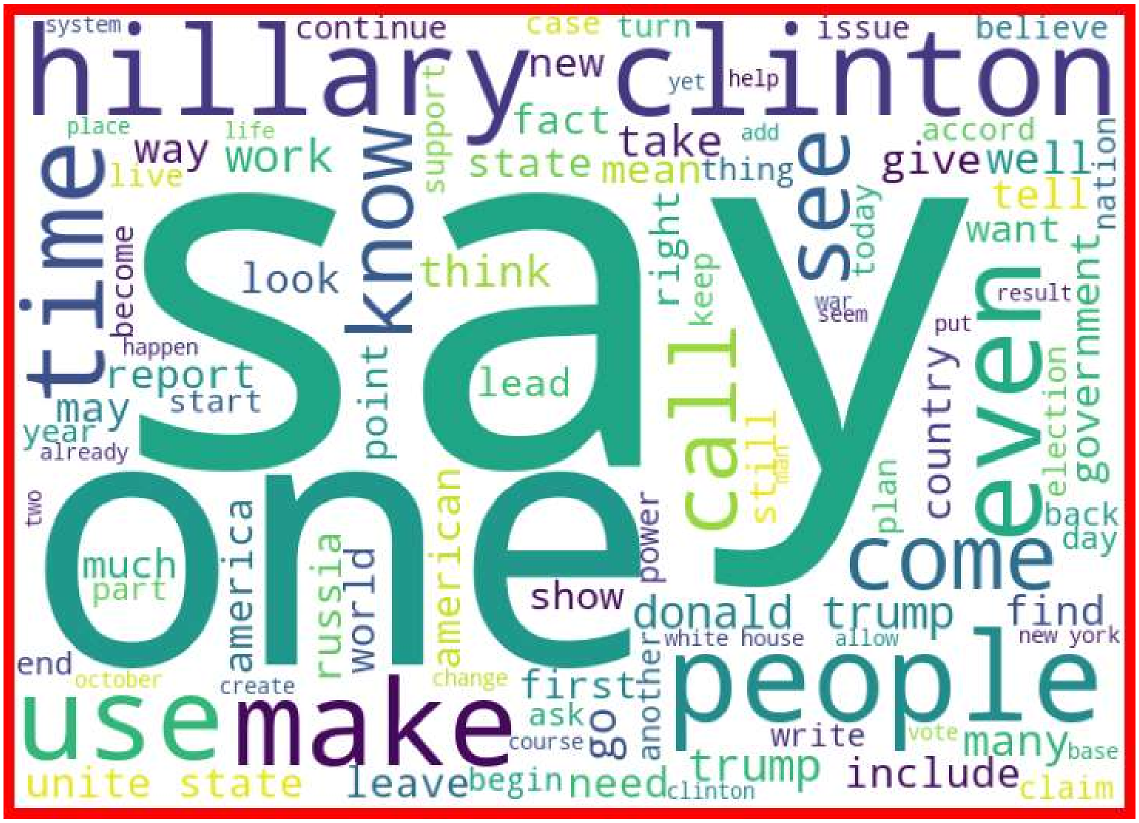


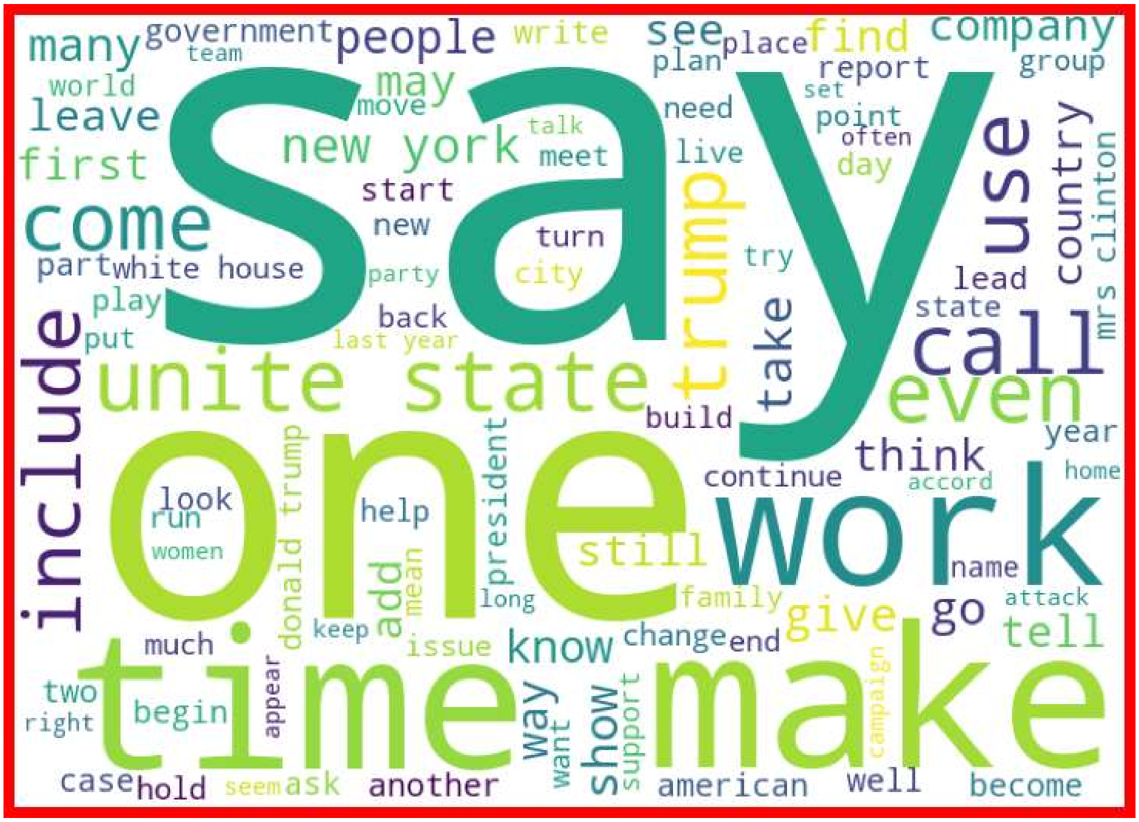
- Visualizations






0
386
772
1158
1544
1930
2316
2702
3088
3474
3860
4246
4632
5018
5404
5790
6176
6562
6948
7334
7720
8106
8492
8878
9264
9650
10036
10422
10808
11194
11580
11966
12352
12738
13124
13510
13896
14282
14668
15054
15440
15826
16212
16598
16984
17370
17756
18142
18528
18914
19300
19686
20072
20458
Unnamed: 0
id
headline
written_by
news
label
1.0
0.8
0.6
0.4
0.2
0.0


say
one
hillary clinton
time
know
see
even
call
come
people
make
use
system
continue
case
turn
issue
believe
new
yet
help
place
life
fact
add
accord
way
work
take
state
mean
thing
give
well
live
support
tell
want
nation
today
right
become
look
think
keep
government
war
seem
result
put
happen
report
lead
start
may
point
country
election
year
already
still
plan
man
two
back
day
power
much
part
america
russia
world
american
show
donald trump
find
end
white house
allow
new york
october
create
change
first
go
another
ask
write
vote
many
base
course
unite state
leave
begin
need
clinton
trump
include
claim


say
one
time
make
work
unite state
come
include
call
use
even
trump
new york
people
many
government
team
world
leave
first
may
move
write
see
place
plan
find
report
company
group
set
point
often
need
talk
meet
live
day
country
mrs clinton
start
new
turn
try
part
white house
party
city
lead
state
play
put
back
last year
take
build
year
think
continue
accord
home
look
run
women
help
donald trump
president
add
mean
still
name
family
long
attack
keep
much
give
go
tell
know
issue
change
end
two
way
show
begin
right
appear
want
support
campaign
case
hold
seem
ask
another
american
well
become

CONCLUSION

- Learning Outcomes of the Study in respect of Data Science

  I have learn some inbuild funtion of python at the time of data cleaning. Object columns to Integer are done by using python map function over on lambda function. The main challenging is that to run all model once and get the best model name with best parameters also with a AUC ROC curve plot. This is done by using pipeline. This will helps me to get out from overfitting and underfitting.