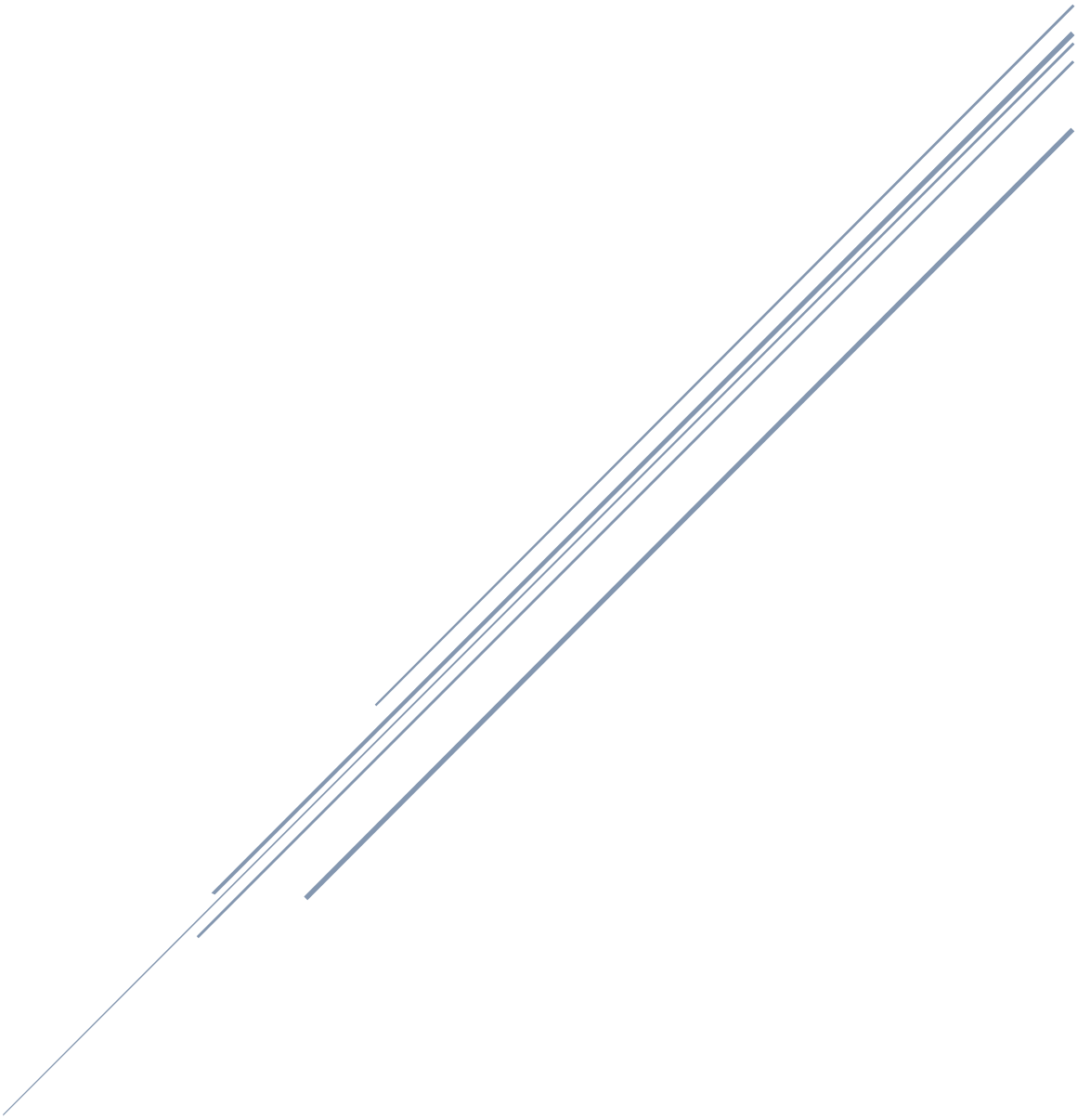
AI WITH PYTHON

PROJECT REPORT



DASARI HARSHA SRI RAM

NEWS CLASSIFICATION USING NLP

ABSTRACT OF THE PROJECT:

This project is about classifying a given news is whether genuine one or else a fake one using Natural Language Processing (NLP) and ML techniques. Programming Language used in this project is Python version 3.9. NLTK and SK Learn tool kits are used in this project.

OBJECTIVE OF THE PROJECT:

Our objective is to build a ML model which will predict whether the given news is fake or real wit the help of training two existing data sets within the model.

INTRODUCTION

This project mainly deals with two things i.e. Natural Language Processing (NLP) and Machine Learning (ML). I use NLTK and SK learn tool kits to work on NLP things and also pandas library to use data sets. I use 2 ML classifiers for knowing about how the model performs with each classifier. By doing this project , I gained some knowledge about building and training ML model, also on how to use NLP features through SK learn.

METHODOLOGY

I have clearly mentioned about the methodology in the code itself using Comments feature in Python language. So, I believe that one can understand the methodology of my project by going through the code.

NEWS CLASSIFICATION USING NATURAL LANGUAGE PROCESSIING (NLP)

In [20]:
```
# First I am  downloading  and importing  libraries that I use in this project
```

In [21]:
```
# I will now import pandas and  download nltk
```

In [22]:
```
import nltk
import pandas as pd
```

In [23]:
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to C:\Users\Harsha Sri
[nltk_data]     Ram\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[23]: True

In [24]:
```
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to C:\Users\Harsha Sri
[nltk_data]     Ram\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[24]: True

In [ ]:
```
# so this says that successfully installed nltk.
```

In [ ]:
```
#now I name the 2 different data sets as 'real' and 'fake'
```

In [25]:
```
real = pd.read_csv(r'C:\Users\Harsha Sri Ram\Downloads\FAKE AND TRUE DATA SETS OF NEWS\True.csv')
fake = pd.read_csv(r'C:\Users\Harsha Sri Ram\Downloads\FAKE AND TRUE DATA SETS OF NEWS\Fake.csv')
```

In [26]:
```
# I also imported datasets successfully, to check them lets try:
```

In [27]:
```
fake
```

Out[27]:

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017
...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016

23481 rows × 4 columns

In [28]:
```
real
```

Out[28]:

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017
...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017

21417 rows × 4 columns

In [ ]:
```
# both above shows that data imported successfully.

# now I will add a label to differentiate both fake and real news.
```

In [29]:
```
fake["correctness"] = 0
real["correctness"] = 1

#lets again see both the data sets with new label added
```

```
fake
```

Out[29]:

	title	text	subject	date	correctness
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
...	...	...	...	...	...
23476	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
23479	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

23481 rows × 5 columns

In [30]:
```
real
```

Out[30]:

	title	text	subject	date	correctness
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1
...	...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

21417 rows × 5 columns

In [31]:
```
# above datasets are now also showing newly added label correctly.

# lets concatenate both the data sets fake and real.

#lets name that new set as data.

data = pd.concat([fake,real],axis=0)
data
```

Out[31]:

	title	text	subject	date	correctness
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0
...	...	...	...	...	...
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

44898 rows × 5 columns

In [32]:
```
# now I remove the unwanted features from the data using pandas.

data = data.reset_index(drop=True)

data = data.drop(['title','subject','date'], axis = 1)

data
```

Out[32]:

	text	correctness
0	Donald Trump just couldn t wish all Americans ...	0
1	House Intelligence Committee Chairman Devin Nu...	0
2	On Friday, it was revealed that former Milwauk...	0
3	On Christmas day, Donald Trump announced that ...	0
4	Pope Francis used his annual Christmas Day mes...	0
...	...	...
44893	BRUSSELS (Reuters) - NATO allies on Tuesday we...	1
44894	LONDON (Reuters) - LexisNexis, a provider of l...	1
44895	MINSK (Reuters) - In the shadow of disused Sov...	1
44896	MOSCOW (Reuters) - Vatican Secretary of State ...	1
44897	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	1

44898 rows × 2 columns

In [ ]:
```
# I successfully imported the data we have with us.
# we now proceed onto data preprocessing.
```

Data Preprocessing

In [33]:
```
# In word preprocessing we normally do 3 steps as follows

# 1.Tokenization
from nltk.tokenize import word_tokenize

data['text'] = data['text'].apply(word_tokenize)
```

In [34]:
```
# 2.Stemming
from nltk.stem.snowball  import SnowballStemmer

sb = SnowballStemmer("english",ignore_stopwords=False)
```

In [35]:
```
def stem_it(text):
    return [sb.stem(word) for word in text]
```

In [36]:
```
data['text'] = data['text'].apply(stem_it)
```

In [37]:
```
# 3.Stop Word Removal
def stopword_remover(text):
    return [word for word in text if len(word)>>2]
```

In [39]:
```
data['text'] = data['text'].apply(' '.join)
```

In [40]:
```
data
```

Out[40]:

	text	correctness
0	donald trump just couldn t wish all american a...	0
1	hous intellig committe chairman devin nune is ...	0
2	on friday , it was reveal that former milwauke...	0
3	on christma day , donald trump announc that he...	0
4	pope franci use his annual christma day messag...	0
...	...	...
44893	brussel (reuter) - nato alli on tuesday welc...	1
44894	london (reuter) - lexisnexi , a provid of le...	1
44895	minsk (reuter) - in the shadow of disus sovi...	1
44896	moscow (reuter) - vatican secretari of state...	1
44897	jakarta (reuter) - indonesia will buy 11 suk...	1

44898 rows × 2 columns

In [ ]:
```
# I also completed Data Preprocessing. Now next I split the data set into 2 parts
# i) one part is for testing
# ii) other is for training the model
```

Splitting the Data set

In [ ]:
```
# To perform data splitting and remaining things I need sklearn(sci-kit learn).
# so I will import each tool from sklearn  when they are required in project.
```

In [43]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(data['text'],data['correctness'],test_size = 0.25)
```

In [44]:
```
X_train
```

Out[44]:
```
31379    san francisco/washington ( reuter ) - a settle...
8892     that dread night in benghazi , libya in septem...
12462    a true stori of how american reject social , c...
25795    washington ( reuter ) - presid donald trump pr...
39015    tokyo ( reuter ) - u.s. presid donald trump sa...
                               ...
28841    munich ( reuter ) - the unit state on saturday...
17338    i heard mr. mclellan on the radio yesterday an...
39451    tripoli ( reuter ) - libya s coastguard interc...
76       donald trump just got his ass hand to him by n...
27429    los angel ( reuter ) - a much-delay u.s. rule ...
Name: text, Length: 33673, dtype: object
```

In [45]:
```
X_test
```

Out[45]:
```
19902    mayb comedi is a better career choic for gari ...
44543    beirut ( reuter ) - u.s.-back syrian militia h...
31333    miami , fla. ( reuter ) - hillari clinton recr...
35991    denpasar , indonesia ( reuter ) - indonesian p...
39960    berlin ( reuter ) - the three parti explor a p...
                               ...
7398     look back at all of the offens , ridicul thing...
41790    strasbourg ( reuter ) - the eu execut call aga...
38759    moscow ( reuter ) - the kremlin said on thursd...
13129    john kerri decid to reveal how he truli feel a...
28454    ( reuter ) - share of hospit and health insur ...
Name: text, Length: 11225, dtype: object
```

In [46]:
```
y_train
```

Out[46]:
```
31379    1
8892     0
12462    0
25795    1
39015    1
        ..
28841    1
17338    0
39451    1
76       0
27429    1
Name: correctness, Length: 33673, dtype: int64
```

In [47]:
```
y_test
```

Out[47]:
```
19902    0
44543    1
31333    1
35991    1
39960    1
        ..
7398     0
41790    1
38759    1
13129    0
28454    1
Name: correctness, Length: 11225, dtype: int64
```

In [ ]:
```
# with all these above tools, I successfully splitted the dataset into 2 parts.
```

VECTORIZATION (TFIDF)

In [ ]:
```
# during vectorization we treat words as numbers as finally to get their predictions whether those words
# exist or not in given data.
```

In [48]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf = TfidfVectorizer(max_df = 0.7)


tfidf_train = tfidf.fit_transform(X_train)
tfidf_test  = tfidf.transform(X_test)
```

BUILDING OF ML MODEL

In [ ]:
```
# using two differnet ML classifiers I will check the prediction of the news
# i)Logistic Regression  ii) Passive Aggressive Classifier
```

In [50]:
```
from sklearn.linear_model import LogisticRegression

LR = LogisticRegression(max_iter = 900)

LR.fit(tfidf_train,y_train)
```

Out[50]:
```
LogisticRegression(max_iter=900)
```

In [51]:
```
pred1 = LR.predict(tfidf_test)
pred1
```

Out[51]:
```
array([0, 1, 0, ..., 1, 0, 1], dtype=int64)
```

In [52]:
```
y_test
```

Out[52]:
```
19902    0
44543    1
31333    1
35991    1
39960    1
        ..
7398     0
41790    1
38759    1
13129    0
28454    1
Name: correctness, Length: 11225, dtype: int64
```

In [ ]:
```
# to get accuracy of our perdiction, I use accuracy_score tool in sklearn
```

In [61]:
```
from sklearn.metrics import accuracy_score

score1 = accuracy_score(y_test,pred1)

score1
```

Out[61]:
```
0.9889532293986637
```

So I got accuracy 98.89% using Regression Classifier

In [ ]:
```
# now  I try using Passive Aggressive Classifier
```

In [58]:
```
from sklearn.linear_model  import PassiveAggressiveClassifier

PAC = PassiveAggressiveClassifier(max_iter = 100)

PAC.fit(tfidf_train,y_train)
```

Out[58]:
```
PassiveAggressiveClassifier(max_iter=100)
```

In [59]:
```
pred2 = PAC.predict(tfidf_test)
pred2
```

Out[59]:
```
array([0, 1, 1, ..., 1, 0, 1], dtype=int64)
```

In [60]:
```
y_test
```

Out[60]:
```
19902    0
44543    1
31333    1
35991    1
39960    1
        ..
7398     0
41790    1
38759    1
13129    0
28454    1
Name: correctness, Length: 11225, dtype: int64
```

In [62]:
```
from sklearn.metrics import accuracy_score
score2 = accuracy_score(y_test,pred2)
score2
```

Out[62]:
```
0.996792873051225
```

So I got accuracy 99.67% using Passive Aggerssive Classifier

Finally we Classified news using Natural Language Processing