NLP FAKE NEWS CLASSIFICATION

NAME - KSHITIJ BHARDWAJ

SECTION - DS

CLASS ROLL NO. - 31

UNIV. ROLL NO. - 2013652

UNIVERSITY - Graphic Era University,Dehradun

In [4]:

```
import nltk
nltk.download("punkt")
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\kshit\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

Out[4]:

```
True
```

In [2]:

```
import pandas as pd
```

In [3]:

```
fake = pd.read_csv("D:\\True-210604-161650.csv")
true = pd.read_csv("D:\\Fake-210604-161841.csv")
```

In [6]:

```
display(fake.info)
display(true.info())
```

```
<bound method DataFrame.info of                                                    title  \
0      As U.S. budget fight looms, Republicans flip t...
1      U.S. military to accept transgender recruits o...
2      Senior U.S. Republican senator: 'Let Mr. Muell...
3      FBI Russia probe helped by Australian diplomat...
4      Trump wants Postal Service to charge 'much mor...
...                                                  ...
21412  'Fully committed' NATO backs new U.S. approach...
21413  LexisNexis withdrew two products from Chinese ...
21414  Minsk cultural hub becomes haven from authorities
21415  Vatican upbeat on possibility of Pope Francis ...
21416  Indonesia to buy $1.14 billion worth of Russia...

                                                    text       subject  \
0      WASHINGTON (Reuters) - The head of a conservat...  politicsNews
1      WASHINGTON (Reuters) - Transgender people will...  politicsNews
2      WASHINGTON (Reuters) - The special counsel inv...  politicsNews
3      WASHINGTON (Reuters) - Trump campaign adviser ...  politicsNews
4      SEATTLE/WASHINGTON (Reuters) - President Donal...  politicsNews
...                                                  ...           ...
21412  BRUSSELS (Reuters) - NATO allies on Tuesday we...     worldnews
21413  LONDON (Reuters) - LexisNexis, a provider of l...     worldnews
21414  MINSK (Reuters) - In the shadow of disused Sov...     worldnews
21415  MOSCOW (Reuters) - Vatican Secretary of State ...     worldnews
21416  JAKARTA (Reuters) - Indonesia will buy 11 Sukh...     worldnews

                     date
0      December 31, 2017
1      December 29, 2017
2      December 31, 2017
3      December 30, 2017
4      December 29, 2017
...                   ...
21412    August 22, 2017
21413    August 22, 2017
21414    August 22, 2017
21415    August 22, 2017
21416    August 22, 2017

[21417 rows x 4 columns]>
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23481 entries, 0 to 23480
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
 ---  ------   --------------  -----
 0   title    23481 non-null  object
 1   text     23481 non-null  object
 2   subject  23481 non-null  object
 3   date     23481 non-null  object
dtypes: object(4)
memory usage: 733.9+ KB
None
```

In [14]:

```
fake.subject.value_counts()
```

Out[14]:

```
politicsNews    11272
worldnews       10145
Name: subject, dtype: int64
```

In [13]:

```
print('\n')
true.subject.value_counts()
```

Out[13]:

```
News               9050
politics           6841
left-news          4459
Government News    1570
US_News             783
Middle-east         778
Name: subject, dtype: int64
```

In [9]:

```
fake['target'] = 0
true['target'] = 1
```

In [15]:

```
fake.head()
```

Out[15]:

	title	text	subject	date	target
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	0
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	0
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	0
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	0
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	0

In [16]:

```
true.head()
```

Out[16]:

	title	text	subject	date	target
0	Donald Trump Sends Out Embarrassing New Year’...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	1
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	1
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	1
3	Trump Is So Obsessed He Even Has Obama’s Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	1
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	1

In [18]:

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

In [19]:

```
data = data.reset_index(drop=True)
```

In [20]:

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

In [21]:

```
data.columns
```

Out[21]:

```
Index(['text', 'target'], dtype='object')
```

In [22]:

```
data
```

Out[22]:

	text	target
0	WASHINGTON (Reuters) - The head of a conservat...	0
1	WASHINGTON (Reuters) - Transgender people will...	0
2	WASHINGTON (Reuters) - The special counsel inv...	0
3	WASHINGTON (Reuters) - Trump campaign adviser ...	0
4	SEATTLE/WASHINGTON (Reuters) - President Donal...	0
...	...	...
44893	21st Century Wire says As 21WIRE reported earl...	1
44894	21st Century Wire says It s a familiar theme. ...	1
44895	Patrick Henningsen 21st Century WireRemember ...	1
44896	21st Century Wire says Al Jazeera America will...	1
44897	21st Century Wire says As 21WIRE predicted in ...	1

44898 rows × 2 columns

TOKENIZATION

In [25]:

```
from nltk.tokenize import word_tokenize
data['text']=data['text'].apply(word_tokenize)
```

In [26]:

```
data.head(10)
```

Out[26]:

	text	target
0	[WASHINGTON, (, Reuters,), -, The, head, of, ...	0
1	[WASHINGTON, (, Reuters,), -, Transgender, pe...	0
2	[WASHINGTON, (, Reuters,), -, The, special, c...	0
3	[WASHINGTON, (, Reuters,), -, Trump, campaign...	0
4	[SEATTLE/WASHINGTON, (, Reuters,), -, Preside...	0
5	[WEST, PALM, BEACH, ,, Fla./WASHINGTON, (, Reu...	0
6	[WEST, PALM, BEACH, ,, Fla, (, Reuters,), -, ...	0
7	[The, following, statements, were, posted, to,...	0
8	[The, following, statements, were, posted, to,...	0
9	[WASHINGTON, (, Reuters,), -, Alabama, Secret...	0

STEMMING

In [28]:

```
from nltk.stem.snowball import SnowballStemmer
porter = SnowballStemmer("english", ignore_stopwords=False)
```

In [29]:

```
def stem_it(text):
    return [porter.stem(word) for word in text]
```

In [30]:

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

In [31]:

```
data.head(10)
```

Out[31]:

	text	target

0	[washington, (, reuter,), -, the, head, of, a...	0
1	[washington, (, reuter,), -, transgend, peopl...	0
2	[washington, (, reuter,), -, the, special, co...	0
3	[washington, (, reuter,), -, trump, campaign,...	0
4	[seattle/washington, (, reuter,), -, presid, ...	0
5	[west, palm, beach, ,, fla./washington, (, reu...	0
6	[west, palm, beach, ,, fla, (, reuter,), -, p...	0
7	[the, follow, statement, were, post, to, the, ...	0
8	[the, follow, statement, were, post, to, the, ...	0
9	[washington, (, reuter,), -, alabama, secreta...	0

STOPWORD REMOVAL

In [32]:

```
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
stopwords.words('english')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\kshit\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
```

Out[32]:

```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his',
 'himself',
 'she',
 "she's",
 'her',
 'hers',
 'herself',
 'it',
 "it's",
 'its',
 'itself',
 'they',
 'them',
 'their',
 'theirs',
 'themselves',
 'what',
 'which',
 'who',
 'whom',
 'this',
 'that',
 "that'll",
 'these',
 'those',
 'am',
 'is',
 'are',
 'was',
 'were',
 'be',
 'been',
 'being',
 'have',
 'has',
 'had',
 'having',
 'do',
 'does',
 'did',
 'doing',
 'a',
 'an',
 'the',
 'and',
 'but',
 'if',
 'or',
 'because',
 'as',
 'until',
 'while',
 'of',
 'at',
 'by',
 'for',
 'with',
 'about',
 'against',
 'between',
 'into',
 'through',
 'during',
 'before',
 'after',
 'above',
 'below',
 'to',
 'from',
 'up',
 'down',
 'in',
 'out',
 'on',
 'off',
 'over',
 'under',
 'again',
 'further',
 'then',
 'once',
 'here',
 'there',
 'when',
 'where',
 'why',
 'how',
 'all',
 'any',
 'both',
 'each',
 'few',
 'more',
 'most',
 'other',
 'some',
 'such',
 'no',
 'nor',
 'not',
 'only',
 'own',
 'same',
 'so',
 'than',
 'too',
 'very',
 's',
 't',
 'can',
 'will',
 'just',
 'don',
 "don't",
 'should',
 "should've",
 'now',
 'd',
 'll',
 'm',
 'o',
 're',
 've',
 'y',
 'ain',
 'aren',
 "aren't",
 'couldn',
 "couldn't",
 'didn',
 "didn't",
 'doesn',
 "doesn't",
 'hadn',
 "hadn't",
 'hasn',
 "hasn't",
 'haven',
 "haven't",
 'isn',
 "isn't",
 'ma',
 'mightn',
 "mightn't",
 'mustn',
 "mustn't",
 'needn',
 "needn't",
 'shan',
 "shan't",
 'shouldn',
 "shouldn't",
 'wasn',
 "wasn't",
 'weren',
 "weren't",
 'won',
 "won't",
 'wouldn',
 "wouldn't"]
```

In [33]:

```python
def stop_it(t):
    dt = [word for word in t if len(word)>2]
    return dt
```

In [34]:

```python
data['text']=data['text'].apply(stop_it)
```

In [35]:

```python
data.head(10)
```

Out[35]:

	text	target
0	[washington, reuter, the, head, conserv, repub...	0
1	[washington, reuter, transgend, peopl, will, a...	0
2	[washington, reuter, the, special, counsel, in...	0
3	[washington, reuter, trump, campaign, advis, g...	0
4	[seattle/washington, reuter, presid, donald, t...	0
5	[west, palm, beach, fla./washington, reuter, t...	0
6	[west, palm, beach, fla, reuter, presid, donal...	0
7	[the, follow, statement, were, post, the, veri...	0
8	[the, follow, statement, were, post, the, veri...	0
9	[washington, reuter, alabama, secretari, state...	0

In [36]:

```python
data['text']=data['text'].apply(' '.join)
```

Splitting up of data

Splitting up of data

In [38]:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(data['text'], data['target'], test_size=0.25)
display(X_train.head())
print('\\n')
display(y_train.head())
```

```
12506    phnom penh reuter the european union has suspe...
19587    riyadh reuter larg saudi public univers announ...
263      washington reuter presid donald trump former c...
7959     washington reuter u.s. lawmak express doubt th...
21334    vatican citi reuter perhap onli matter time be...
Name: text, dtype: object
\n
12506    0
19587    0
263      0
7959     0
21334    0
Name: target, dtype: int64
```

Vectorization

In [39]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
my_tfidf = TfidfVectorizer( max_df=0.7)
tfidf_train = my_tfidf.fit_transform(X_train)
tfidf_test = my_tfidf.transform(X_test)
```

In [40]:

```
tfidf_train
```

Out[40]:

```
<33673x91729 sparse matrix of type '<class 'numpy.float64'>'
	with 5992341 stored elements in Compressed Sparse Row format>
```

LogisticRegression

In [41]:

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

In [42]:

```
model_1 = LogisticRegression(max_iter=900)
model_1.fit(tfidf_train, y_train)
pred_1 = model_1.predict(tfidf_test)
cr1    = accuracy_score(y_test,pred_1)
print(cr1*100)
```

```
98.79732739420936
```

PassiveAggressiveClassifier

In [43]:

```
from sklearn.linear_model import PassiveAggressiveClassifier
model = PassiveAggressiveClassifier(max_iter=50)
model.fit(tfidf_train, y_train)
```

Out[43]:

```
PassiveAggressiveClassifier(max_iter=50)
```

In [44]:

```
y_pred = model.predict(tfidf_test)
accscore = accuracy_score(y_test, y_pred)
print('The accuracy of prediction is ',accscore*100)
```

```
The accuracy of prediction is  99.52783964365256
```

```
In [ ]:
```

Loading [MathJax]/extensions/Safe.js