Fake News Prediction

Dataset Description

train.csv: A full training dataset with the following attributes:

https://www.kaggle.com/c/fake-news/data?select=train.csv

- id: unique id for a news article
- title: the title of a news article
- author: author of the news article
- text: the text of the article; could be incomplete
- label: a label that marks the article as potentially unreliable
- 1: unreliable
- 0: reliable

test.csv: A testing training dataset with all the same attributes at train.csv without the label.

submit.csv: A sample submission that you can

In [12]:
```
# Importing the libraries
import numpy as np
import pandas as pd
import re # import regular expression
from nltk.corpus import stopwords # import natural language toolkit
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

In [15]:
```
import nltk
nltk.download('stopwords')
```

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

Out[15]:
```
True
```

In [17]:
```
# printing the stopwords in english
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "yo
u'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', 'we
re', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did',
'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'wh
ile', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'thr
ough', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down',
'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'he
re', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few',
'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'sam
e', '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', "had
n't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "might
n't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Data Preprocessing

In [19]:
```
# loading the dataset
news_df = pd.read_csv("train.csv")
news_df.head(40)
```

Out[19]:

	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1
5	5	Jackie Mason: Hollywood Would Love Trump if He...	Daniel Nussbaum	In these trying times, Jackie Mason is the Voi...	0
6	6	Life: Life Of Luxury: Elton John’s 6 Favorite ...	NaN	Ever wonder how Britain’s most iconic pop pian...	1
7	7	Benoît Hamon Wins French Socialist Party’s Pre...	Alissa J. Rubin	PARIS — France chose an idealistic, traditi...	0
8	8	Excerpts From a Draft Script for Donald Trump’...	NaN	Donald J. Trump is scheduled to make a highly ...	0
9	9	A Back-Channel Plan for Ukraine and Russia, Co...	Megan Twohey and Scott Shane	A week before Michael T. Flynn resigned as nat...	0
10	10	Obama’s Organizing for Action Partners with So...	Aaron Klein	Organizing for Action, the activist group that...	0
11	11	BBC Comedy Sketch "Real Housewives of ISIS" Ca...	Chris Tomlinson	The BBC produced spoof on the “Real Housewives...	0
12	12	Russian Researchers Discover Secret Nazi Milit...	Amando Flavio	The mystery surrounding The Third Reich and Na...	1
13	13	US Officials See No Link Between Trump and Russia	Jason Ditz	Clinton Campaign Demands FBI Affirm Trump's Ru...	1
14	14	Re: Yes, There Are Paid Government Trolls On S...	AnotherAnnie	Yes, There Are Paid Government Trolls On Socia...	1
15	15	In Major League Soccer, Argentines Find a Home...	Jack Williams	Guillermo Barros Schelotto was not the first A...	0
16	16	Wells Fargo Chief Abruptly Steps Down - The Ne...	Michael Corkery and Stacy Cowley	The scandal engulfing Wells Fargo toppled its ...	0
17	17	Anonymous Donor Pays $2.5 Million To Release E...	Starkman	A Caddo Nation tribal leader has just been fre...	1
18	18	FBI Closes In On Hillary!	The Doc	FBI Closes In On Hillary! Posted on Home » Hea...	1
19	19	Chuck Todd: ‘BuzzFeed Did Donald Trump a Polit...	Jeff Poor	Wednesday after Donald Trump’s press confere...	0
20	20	News: Hope For The GOP: A Nude Paul Ryan Has J...	NaN	Email \nSince Donald Trump entered the electio...	1

	id	title	author	text	label
21	21	Monica Lewinsky, Clinton Sex Scandal Set for '...	Jerome Hudson	Screenwriter Ryan Murphy, who has produced the...	0
22	22	Rob Reiner: Trump Is 'Mentally Unstable' - Bre...	Pam Key	Sunday on MSNBC's "AM Joy," actor and director...	0
23	23	Massachusetts Cop's Wife Busted for Pinning Fa...	NaN	Massachusetts Cop's Wife Busted for Pinning Fa...	1
24	24	Abortion Pill Orders Rise in 7 Latin American ...	Donald G. McNeil Jr. and Pam Belluck	Orders for abortion pills by women in seven La...	0
25	25	Nukes and the UN: a Historic Treaty to Ban Nuc...	Ira Helfand	Email \nIn an historic move the United Nations...	1
26	26	EXCLUSIVE: Islamic State Supporters Vow to 'Sh...	Aaron Klein and Ali Waked	JERUSALEM — Islamic State sympathizers and ...	0
27	27	Humiliated Hillary Tries To Hide What Camera C...	Amanda Shea	Humiliated Hillary Tries To Hide What Camera C...	1
28	28	Andrea Tantaros of Fox News Claims Retaliation...	Jim Dwyer	Andrea Tantaros, a former Fox News host, charg...	0
29	29	How Hillary Clinton Became a Hawk - The New Yo...	Mark Landler	Hillary Clinton sat in the hideaway study off ...	0
30	30	Chuck Todd to BuzzFeed EIC: 'You Just Publishe...	Ian Hanchett	During a discussion of BuzzFeed's story on a d...	0
31	31	Israel is Becoming Pivotal to China's Mid-East...	NaN	Country: Israel While China is silently playin...	1
32	32	Having Won, Boris Johnson and 'Brexit' Leaders...	Steven Erlanger	LONDON — With their giddy celebrations of "...	0
33	33	Texas Oil Fields Rebound From Price Lull, but ...	Clifford Krauss	MIDLAND, Tex. — In the land where oil jobs ...	0
34	34	Bayer Deal for Monsanto Follows Agribusiness T...	Leslie Picker, Danny Hakim and Michael J. de l...	Don Halcomb, a farmer in Adairville, Ky. is...	0
35	35	Russia Moves to Ban Jehovah's Witnesses as 'Ex...	Andrew Higgins	VOROKHOBINO, Russia — A dedicated pacifist ...	0
36	36	Re: Why We Are Still In 'The Danger Zone' Unti...	greanfinisher .	Why We Are Still In 'The Danger Zone' Until Ja...	1
37	37	Open Thread (NOT U.S. Election) 2016-39	b	Open Thread (NOT U.S. Election) 2016-39 \nNews...	1
38	38	Democrat Gutierrez Blames Chicago's Gun Violen...	AWR Hawkins	Rep. Luis Gutierrez ( ) made the rounds on MS...	0
39	39	Avoiding Peanuts to Avoid an Allergy Is a Bad ...	Aaron E. Carroll	This article originally ran in April. We are r...	0

In [20]:
```
news_df.shape
```

Out[20]:
```
(20800, 5)
```

In [21]:
```
# checking the missing values
news_df.isnull().sum()
```

Out[21]:
```
id           0
title      558
author    1957
text        39
label        0
dtype: int64
```

In [22]:
```
# Replacing the null values with empty string
news_df = news_df.fillna('')
```

In [23]:
```
# checking again the missing values
news_df.isnull().sum()
```

Out[23]:
```
id        0
title     0
author    0
text      0
label     0
dtype: int64
```

In [24]:
```
# merging the author name and news title
news_df['content'] = news_df['author']+' '+news_df['title']
```

In [25]:
```
print(news_df['content'])
```

```
0        Darrell Lucus House Dem Aide: We Didn’t Even S...
1        Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2        Consortiumnews.com Why the Truth Might Get You...
3        Jessica Purkiss 15 Civilians Killed In Single ...
4        Howard Portnoy Iranian woman jailed for fictio...
                               ...
20795    Jerome Hudson Rapper T.I.: Trump a ’Poster Chi...
20796    Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797    Michael J. de la Merced and Rachel Abrams Macy...
20798    Alex Ansary NATO, Russia To Hold Parallel Exer...
20799              David Swanson What Keeps the F-35 Alive
Name: content, Length: 20800, dtype: object
```

In [29]:
```
# Separating the content and label column
X = news_df.drop(columns='label',axis=1)
Y = news_df['label']
```

In [31]:
```
print(X)
print(Y)
```

```
          id                                              title  \
0          0  House Dem Aide: We Didn’t Even See Comey’s Let...
1          1  FLYNN: Hillary Clinton, Big Woman on Campus - ...
2          2                  Why the Truth Might Get You Fired
3          3  15 Civilians Killed In Single US Airstrike Hav...
4          4  Iranian woman jailed for fictional unpublished...
...      ...                                                ...
20795  20795  Rapper T.I.: Trump a ’Poster Child For White S...
20796  20796  N.F.L. Playoffs: Schedule, Matchups and Odds -...
20797  20797  Macy’s Is Said to Receive Takeover Approach by...
20798  20798  NATO, Russia To Hold Parallel Exercises In Bal...
20799  20799                          What Keeps the F-35 Alive

                                      author  \
0                               Darrell Lucus
1                             Daniel J. Flynn
2                           Consortiumnews.com
3                              Jessica Purkiss
4                               Howard Portnoy
...                                        ...
20795                           Jerome Hudson
20796                         Benjamin Hoffman
20797  Michael J. de la Merced and Rachel Abrams
20798                             Alex Ansary
20799                           David Swanson

                                                    text  \
0      House Dem Aide: We Didn’t Even See Comey’s Let...
1      Ever get the feeling your life circles the rou...
2      Why the Truth Might Get You Fired October 29, ...
3      Videos 15 Civilians Killed In Single US Airstr...
4      Print \nAn Iranian woman has been sentenced to...
...                                                  ...
20795  Rapper T. I. unloaded on black celebrities who...
20796  When the Green Bay Packers lost to the Washing...
20797  The Macy’s of today grew from the union of sev...
20798  NATO, Russia To Hold Parallel Exercises In Bal...
20799    David Swanson is an author, activist, journa...

                                                 content
0      Darrell Lucus House Dem Aide: We Didn’t Even S...
1      Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2      Consortiumnews.com Why the Truth Might Get You...
3      Jessica Purkiss 15 Civilians Killed In Single ...
4      Howard Portnoy Iranian woman jailed for fictio...
...                                                  ...
20795  Jerome Hudson Rapper T.I.: Trump a ’Poster Chi...
20796  Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797  Michael J. de la Merced and Rachel Abrams Macy...
20798  Alex Ansary NATO, Russia To Hold Parallel Exer...
20799            David Swanson What Keeps the F-35 Alive

[20800 rows x 5 columns]
0        1
1        0
2        1
3        1
4        1
        ..
20795    0
20796    0
20797    0
20798    1
```

```
20799    1
Name: label, Length: 20800, dtype: int64
```

Stemming:

Stemming is the process of reducing the word to its root word.

In [32]:
```
port_stem = PorterStemmer()
```

In [33]:
```
def stemming(content):
    stemmed_content = re.sub('[^a-zA-Z]',' ',content)
    stemmed_content = stemmed_content.lower()
    stemmed_content = stemmed_content.split()
    stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word
    stemmed_content = ' '.join(stemmed_content)
    return stemmed_content
```

In [34]:
```
news_df['content'] = news_df['content'].apply(stemming)
```

In [35]:
```
print(news_df['content'])
```

```
0        darrel lucu hous dem aid even see comey letter...
1        daniel j flynn flynn hillari clinton big woman...
2                  consortiumnew com truth might get fire
3        jessica purkiss civilian kill singl us airstri...
4        howard portnoy iranian woman jail fiction unpu...
                              ...
20795    jerom hudson rapper trump poster child white s...
20796    benjamin hoffman n f l playoff schedul matchup...
20797    michael j de la merc rachel abram maci said re...
20798    alex ansari nato russia hold parallel exercis ...
20799                            david swanson keep f aliv
Name: content, Length: 20800, dtype: object
```

In [37]:
```
# Separating the data and label
X = news_df['content'].values
Y = news_df['label'].values
```

In [38]:
```
print(X)
```

```
['darrel lucu hous dem aid even see comey letter jason chaffetz tweet'
 'daniel j flynn flynn hillari clinton big woman campu breitbart'
 'consortiumnew com truth might get fire' ...
 'michael j de la merc rachel abram maci said receiv takeov approach hudson bay ne
w york time'
 'alex ansari nato russia hold parallel exercis balkan'
 'david swanson keep f aliv']
```

In [39]:
```
print(Y)
```

```
[1 0 1 ... 0 1 1]
```

In [40]:
```
# Converting the text data into numerical
vectorizer = TfidfVectorizer()
vectorizer.fit(X)
X = vectorizer.transform(X)
```

In [41]:
```
print(X)
```

```
  (0, 15686)    0.28485063562728646
  (0, 13473)    0.2565896679337957
  (0, 8909)     0.3635963806326075
  (0, 8630)     0.29212514087043684
  (0, 7692)     0.24785219520671603
  (0, 7005)     0.21874169089359144
  (0, 4973)     0.233316966909351
  (0, 3792)     0.2705332480845492
  (0, 3600)     0.3598939188262559
  (0, 2959)     0.2468450128533713
  (0, 2483)     0.3676519686797209
  (0, 267)      0.27010124977708766
  (1, 16799)    0.30071745655510157
  (1, 6816)     0.1904660198296849
  (1, 5503)     0.7143299355715573
  (1, 3568)     0.26373768806048464
  (1, 2813)     0.19094574062359204
  (1, 2223)     0.3827320386859759
  (1, 1894)     0.15521974226349364
  (1, 1497)     0.2939891562094648
  (2, 15611)    0.41544962664721613
  (2, 9620)     0.49351492943649944
  (2, 5968)     0.3474613386728292
  (2, 5389)     0.3866530551182615
  (2, 3103)     0.46097489583229645
  :     :
  (20797, 13122)        0.2482526352197606
  (20797, 12344)        0.27263457663336677
  (20797, 12138)        0.24778257724396507
  (20797, 10306)        0.08038079000566466
  (20797, 9588) 0.174553480255222
  (20797, 9518) 0.2954204003420313
  (20797, 8988) 0.36160868928090795
  (20797, 8364) 0.22322585870464118
  (20797, 7042) 0.21799048897828688
  (20797, 3643) 0.21155500613623743
  (20797, 1287) 0.33538056804139865
  (20797, 699)  0.30685846079762347
  (20797, 43)   0.29710241860700626
  (20798, 13046)        0.22363267488270608
  (20798, 11052)        0.4460515589182236
  (20798, 10177)        0.3192496370187028
  (20798, 6889) 0.32496285694299426
  (20798, 5032) 0.4083701450239529
  (20798, 1125) 0.4460515589182236
  (20798, 588)  0.3112141524638974
  (20798, 350)  0.28446937819072576
  (20799, 14852)        0.5677577267055112
  (20799, 8036) 0.45983893273780013
  (20799, 3623) 0.37927626273066584
  (20799, 377)  0.5677577267055112
```

Train Test Split

In [43]:
```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.2, stratify=Y,
```

Training the Model

Logistic Regression

In [44]:
```
lr_model = LogisticRegression()
```

In [45]:
```
lr_model.fit(X_train,Y_train)
```

Out[45]:
```
▾ LogisticRegression

LogisticRegression()
```

Evaluation

In [49]:
```
# Accuracy score on training data
X_train_predict = lr_model.predict(X_train)
train_accuracy = accuracy_score(Y_train, X_train_predict)
print('Accuracy score of the training data : ', train_accuracy)
```

```
Accuracy score of the training data :  0.9874399038461539
```

In [50]:
```
# Accuracy score on test data
X_test_predict = lr_model.predict(X_test)
test_accuracy = accuracy_score(Y_test, X_test_predict)
print('Accuracy score of the training data : ', test_accuracy)
```

```
Accuracy score of the training data :  0.9752403846153846
```

Making a Predictive System

In [65]:
```
X_new = X_test[10]
prediction = lr_model.predict(X_new)
print(prediction)


if (prediction[0]==0):
    print("The News is Real")
else:
    print("The News is Fake")
```

```
[1]
The News is Fake
```

In [66]:
```
# checking the actual label for given X_test[10]
print(Y_test[10])
```

```
1
```

In [ ]:
```

```