

ML Project

Fake News Prediction

Name – Gautam Kumar Mahar

Branch – CSE

Overview

News Data --> Data pre-processing (Text to meaning full number) --> Train Test split (Train Our ML Model with training dataset) --> Logistic Regression model (In this project we define whether News is Fake or Not Fake so we are dealing with Binary Classification so Logistic Regression model is useful for this) --> Trained Logistic Regression model --> New data + Trained Loge... --> Gives Result (News is Fake Or Real).

jupyter Fake_News_Prediction Last Checkpoint: 17 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

In [7]: `# print the first 5 rows of the dataframe
news_dataset.head()
we see in OUPUT Level --> 1 is for Fake News And 0 for Real news`

Out[7]:

	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 Aistr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

In [8]: `# Here i counting the number missing values in the dataset
news_dataset.isnull().sum()`

Out[8]:

```
id          0
title       558
author     1957
text        39
label       0
dtype: int64
```

In [9]: `# Now i replacing the null values with empty string
news_dataset = news_dataset.fillna('')`

jupyter Fake_News_Prediction Last Checkpoint: 17 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

In [10]: `# Now merging the author name and news title
news_dataset['content'] = news_dataset['author']+' '+news_dataset['title']`

In [11]: `# So, Now print the out dataset
print(news_dataset['content'])
Now we seen in the output first author name then attached title`

```
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 [12]: `# separating the data & Label
X = news_dataset.drop(columns='label', axis=1)
Y = news_dataset['label']`

In [13]: `print(X)`

jupyter Fake_News_Prediction Last Checkpoint: 17 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

In [13]: `print(X)
print(Y)`

```
id          title \
0      House Dem Aide: We Didn't Even See Comey's Let...
1      FLYNN: Hillary Clinton, Big Woman on Campus - ...
2              Why the Truth Might Get You Fired
3      15 Civilians Killed In Single US Airstrike Hav...
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
```

jupyter Fake_News_Prediction Last Checkpoint: 17 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3 (ipykernel)

20/39

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 Ainstr...
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 Lucas 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
```

jupyter Fake_News_Prediction Last Checkpoint: 17 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3 (ipykernel)

Stemming :

Stemming is the process of reducing a word to its Root word.

Like Example - actor, actress, acting - so for all these root word is "act"

In [14]: port_stem = PorterStemmer() # So this is Loader to this variable

In [15]: def stemming(content): # create a function using def <- because this is not a inbuilt function in python# content represent the i
stemmed_content = re.sub('[^a-zA-Z]', ' ', content) # sub = substitute certain values and content basically the combined fram th
stemmed_content = stemmed_content.lower() # after we need to covert all those to Lower Letter , because upper case Letter mea
stemmed_content = stemmed_content.split() # so once we done with above step then it will be splitted and converted to list
stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words('english')] # so now we ta
stemmed_content = ' '.join(stemmed_content) # in this step we will join all the words
return stemmed_content

In [21]: news_dataset['content'] = news_dataset['content'].apply(stemming) # Now We need to above this function to out content column # Nc

In [22]: print(news_dataset['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 ainstr...
4 david swanson what keeps the f-35 alive

jupyter Fake_News_Prediction Last Checkpoint: 17 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3 (ipykernel)

In [62]: #seprating the data and Label
X = news_dataset['content'].values
Y = news_dataset['label'].values

In [63]: 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 new york time'
'alex ansari nato russia hold parallel exercis balkan'
'david swanson keep f aliv']

In [64]: print(Y)

[1 0 1 ... 0 1 1]

In [65]: Y.shape

Out[65]: (20800,)

In []: # converting the textual data to numerical data
vectorizer = TfidfVectorizer()
vectorizer.fit(X)

X = vectorizer.transform(X)

jupyter Fake_News_Prediction Last Checkpoint: 17 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
In [80]: print(X) # Now in below we can see that text to number changed (Now computer understand these value )
```

```
(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
:            :
```

jupyter Fake_News_Prediction Last Checkpoint: 17 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
:
:
(20797, 13122)    0.2482526352197606
(20797, 12344)    0.27263457663336677
(20797, 12138)    0.24778257724396507
(20797, 10306)    0.08038079000566466
(20797, 9588)     0.174553480255222
(20797, 9518)     0.2954204005420313
(20797, 8988)     0.36160868928090795
(20797, 8364)     0.22322585870464118
(20797, 7042)     0.21799848897828688
(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)    0.3192406370187028
(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
```

jupyter Fake_News_Prediction Last Checkpoint: 17 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

Splitting the dataset to training & test data

```
In [81]: X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, stratify=Y, random_state=2)
```

Training the Model: Logistic Regression


```
In [82]: model = LogisticRegression()
In [83]: model.fit(X_train, Y_train)
Out[83]: LogisticRegression()
```

Evaluation

Accuracy score


```
In [88]: # accuracy score on the training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

In [89]: print('Accuracy score of the training data: ', training_data_accuracy)
# So we see in output accuracy is very good because of logistic model - this model is best for binary classification
```

 jupyter












Fake_News_Prediction

Last Checkpoint: 17 minutes ago (unsaved changes)

 Logout

FileEditViewInsertCellKernelWidgetsHelp

Not TrustedPython 3 (ipykernel)



Code

Evaluation

Accuracy score

```
In [88]: # accuracy score on the training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

In [89]: print('Accuracy score of the training data: ', training_data_accuracy)
# So we see in output accuracy is very good because of Logistic model - this model is best for binary classification

Accuracy score of the training data:  0.9865985576923076

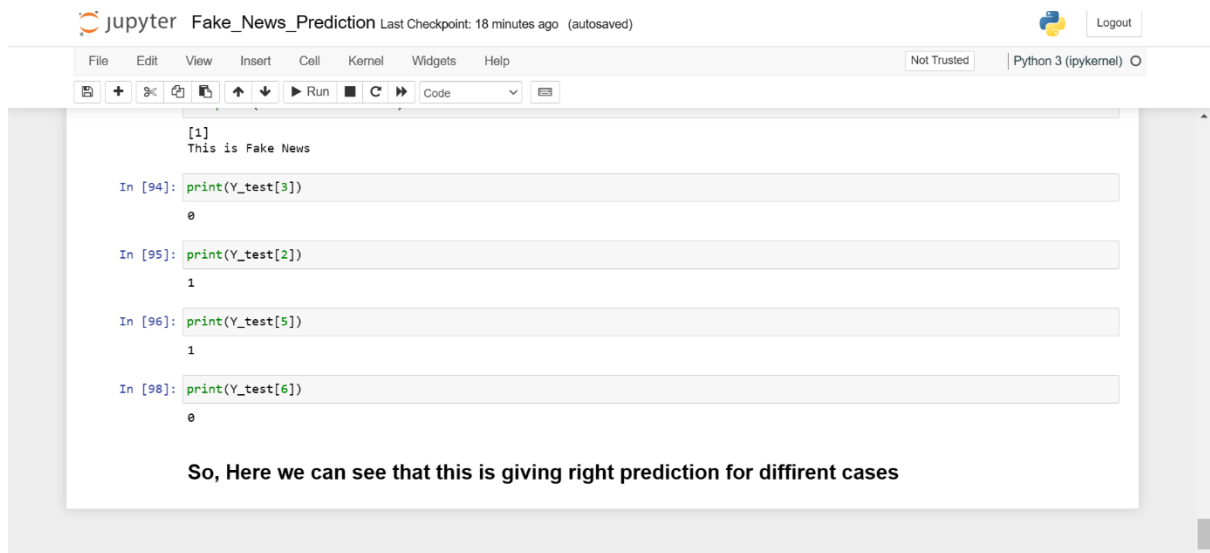
In [90]: # accuracy score on the training data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

In [91]: print('Accuracy score of the test data: ', test_data_accuracy)

Accuracy score of the test data:  0.9790865384615385
```

Making a Predictive System

```
In [93]: X_new = X_test[0]
```



The image shows a Jupyter Notebook interface with the title "Fake_News_Prediction". The top bar indicates the last checkpoint was 18 minutes ago and the notebook is autosaved. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running, and code execution. The notebook content shows a code cell with the following output:

```
[1]  
This is Fake News
```

Below the code cell, there are four input cells, each containing a print statement and its corresponding output:

```
In [94]: print(Y_test[3])  
0
```

```
In [95]: print(Y_test[2])  
1
```

```
In [96]: print(Y_test[5])  
1
```

```
In [98]: print(Y_test[6])  
0
```

At the bottom of the notebook, there is a text cell stating: "So, Here we can see that this is giving right prediction for diffirent cases".

Here we got the right predictions.

- THANK YOU -