Data Mining Project
"FAKE News Detection"

Group Details :-

1. Muhammad Saqib 19I-0494
2. Haseeb Ramzan 19I-0475
3. Amjid Arshad 19I-504

Brief Description of Project :-

In order to accurately classify a collection of news as real or fake we have to build a machine learning model.

To deal with the detection of fake or real news, we will develop the project in python with the help of 'sklearn', we will use 'TfidfVectorizer' in our news data which we will gather from online media.

After the first step is done, we will initialize the classifier, transform and fit the model. In the end, we will calculate the performance of the model using the appropriate performance matrix/matrices. Once we calculate the performance matrices we will be able to see how well our model performs.

1. Dataset for FAKE-News :-

A full training dataset with the following attributes :-

1. **Id** :- unique id for a news article
2. **Title** :- the title of a news article
3. **Author** :- author of the news article
4. **Text** :- the text of the article; could be incomplete
5. **Label** :- label that marks article as potentially unreliable (1 : FAKE & 0 : REAL)

Source of DataSet :- Fake-News.csv

id	title	author	text	label
0	House Dem Aide	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Letter Until Jason Chaffetz Tweeted It By Darrell Lucus on October 30, 2016 Subscribe With apologies to Keith Olbermann, there is no doubt who the Worst Person in The World is this week–FBI Director James Comey. But As we now know, Comey notified the Republican chairmen and Democratic ranking members of the House Intelligence, Judiciary, and — Jason Chaffetz (@jasoninthehouse) October 28, 2016	1
1	FLYNN: Hillary C	Daniel J. Flynn	Ever get the feeling your life circles the roundabout rather than heads in a straight line toward the intended destination? [Hillary Clinton	0
2	Why the Truth M	Consortiumnews	Why the Truth Might Get You Fired October 29, 2016 The tension between intelligence analysts and political policymakers has always been between honest assessments and desired results By Lawrence Davidson For those who might wonder why foreign policy makers repeatedly make bad choices, some insight might be drawn from the following a Back in the early spring of 2003, George W. Bush initiated the invasion of Iraq. One of his key public reasons for doing so was the claim For our purposes, we will concentrate on the belief that Iraq was about to become a hostile nuclear power. Why did President Bush and The short answer is Bush wanted, indeed needed, to believe it as a rationale for invading Iraq. At first he had tried to connect Saddam H But the nuclear weapons gambit proved more fruitful, not because there was any hard evidence for the charge, but because supposedly What we had was a U.S. leadership cadre whose worldview literally demanded a mortally dangerous Iraq, and informants who, in order So the U.S. and its allies insisted that the United Nations send in weapons inspectors to scour Iraq for evidence of a nuclear weapons p On March 19, 2003, Bush launched the invasion of Iraq with the expectation was that, once in occupation of the country, U.S. inspectors Social and Behavioral Sciences to the Rescue? The various U.S. intelligence agencies were thoroughly shaken by this affair, and today, 13 years later, their directors and managers are A "partnership" is being forged between the Office of the Director of National Intelligence (ODNI), which serves as the coordinating cente Despite this effort, it is almost certain that the "social and behavioral sciences" cannot give the spy agencies what they want – a way of The Believers It is simply not true, as the ODNI leaders seem to assert, that U.S. intelligence agency personnel cannot tell, more often than not, that th Therefore, if someone feeds them "snake oil," they usually know it. However, having an accurate grasp of things is often to no avail beca Listen to Charles Gaukel, of the National Intelligence Council – yet another organization that acts as a meeting ground for the 16 intellig I can certainly tell you what it means historically. It means that for the power brokers, "truth" must match up, fit with, their worldview – the On the other hand, as long as what you're selling the leadership matches up with what they want to believe, you can peddle them anyth What does this sad tale tell us? If you want to spend millions of dollars on social and behavioral science research to improve the assess It has happened this way so often, and in so many places, that it is the source of Shakespeare's determination that "what is past, is prel	1

2. Dataset Analysis :-

Dataset selected for the training and testing model is analyzed so that we could decide which preprocessing techniques are sufficient to be used .

Shape of Dataset :- **(3000, 6)** out of which 1521 are FAKE and 1479 are REAL

In [316]:
```
#Data Set Loaded as Pandas Dataset
data = pd.read_csv("dataset.csv",nrows=3000)
data.head()
```

Out[316]:

	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

In [317]:
```
data['text corpus'] = data['title']+' '+data['author']+' '+data['text']
data = data[["id", "title","author","text","text corpus", "label"]]
data.head()
```

Out[317]:

	id	title	author	text	text corpus	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...	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...	FLYNN: Hillary Clinton, Big Woman on Campus - ...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	Why the Truth Might Get You Fired Consortiumne...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	15 Civilians Killed In Single US Airstrike Hav...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	Iranian woman jailed for fictional unpublished...	1

3. DATA Preprocessing :-

For Data Preprocessing , these are the four steps which have been applied .

1. Cleaning missing values
2. Replacing Numbers and Punctuations with whitespaces
3. Remove all the words which have been detected as the stopwords
4. Reduce the word to its origin or root word
5. Converting all the textual data into numerical vector
6. Selecting desired features

4. Train & Test Split :-

Preprocessed data is then splitted into two parts with 0.18 equality using the **train_test_split** function from sklearn library .

3. Splitting :- TRAIN & TEST Data

```
Data from csv file is being spiltted into two parts with 0.18 as Training data and 0.82 as Test data . Also , this split is
random . Each time , TEST and TRAIN data changes .
```

In [335]:
```
def train_test_split(data_train,data_test,folder_train,foler_test) :

    os.mkdir(folder_train)
    train_ind=list(data_train.index)

    # Train folder
    for i in tqdm(range(len(train_ind))):
        os.system('cp '+data_train[train_ind[i]]+' ./'+ folder_train + '/'  +data_train[train_ind[i]].split('/')[2])

    # Test folder
    for j in tqdm(range(len(test_ind))):
        os.system('cp '+data_test[test_ind[j]]+' ./'+ folder_test + '/'  +data_test[test_ind[j]].split('/')[2])
```

In [282]:
```
#Splitting DataSet in Train && Test
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.18, stratify=Y, random_state=124)
```

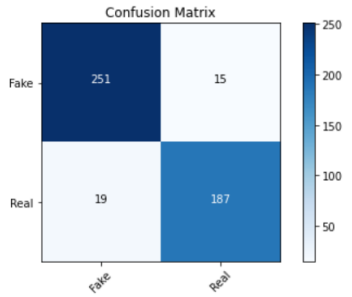
4. LOGISTIC Regression :-

Logistic Regression Model is trained in order to classify the data vector as FAKE or REAL . The accuracy of the model comes out to be 94% on the training set and 90% on the test set. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). ... Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables..

5. Evaluation of Accuracy :-

Confusion matrix is drawn in order to best understand the accuracy of trained machine learning models as the dataset is a bit **Unbalanced** so predicate accuracy calculated in the step above mighn't work well .

```
In [336]:
cm = metrics.confusion_matrix(Y_test, X_test_prediction)
plot_confusion_matrix(cm, classes=['Fake', 'Real'])
```



5. Prediction of Trained Model :-

In order to test our trained model , we used some strings , inputted them to our model and analyzed the output against actual and expected values . This acted as our Validation Set through which we could know how well our model performs on unknown inputs .

5. Prediction using Trained MODEL

```
In [331]:
X_new = X_test[100]

prediction = model1.predict(X_new)

if (prediction[0] == 0):
  print('The news is Real')
else:
  print('The news is Fake & Unreliable')
```

```
The news is Real
```