PROJECT REPORT
ON
FAKE NEWS

Submitted by

MANISHA

Introduction

Business Problem Framing:

Fake news has become one of the biggest problems of our age. It has serious impact on our online as well as offline discourse. One can even go as far as saying that, to date, fake news poses a clear and present danger to western democracy and stability of the society.

Conceptual Background of the Domain Problem:

Fake news's simple meaning is to incorporate information that leads people to the wrong path. Nowadays fake news spreading like water and people share this information without verifying it. This is often done to further or impose certain ideas and is often achieved with political agendas.

Review of Literature:

There are two datasets one for fake news and one for true news. We are combined both datasets using pandas built-in function. Machine learning data only works with numerical features so we have to convert text data into numerical columns. So we have to preprocess the text by steaming, lemmatization, remove stopwords, remove special symbols and numbers, etc.

Motivation for the Problem Undertaken:

We have to detect that the news are published on websites these are fake news or not. For this we analyse our data and then apply model to get better prediction regarding the news.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem:

We use Statistical techniques and analytics modeling in our projects, such as:

- describe() : use to calculate the statistical values that are mean, standard deviation, quantile deviation, minimum and maximum values.
- corr(): use to calculate the relation between feature variable with the target variable.

Data Source and their formats:

There are two datasets one for fake news and one for true news. In true news, there is 21417 news, and in fake news, there is 23502 news.

```
fake_data=pd.read_excel('C:\\Users\\Dell\\Desktop\\Fake.xlsx')
fake_data.head()
```

	title	text	subject	date	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	...	Unnamed: 162	Unnamed: 163	Unnamed: 164	Unn
0	Donald Trump Sends Out Embarrassing New Yeara...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	
3	Trump Is So Obsessed He Even Has Obamaâ€™s Na...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	

```
true_data=pd.read_excel('C:\\Users\\Dell\\Desktop\\True.xlsx')
true_data.head()
```

	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

We insert one label column zero for fake news and one for true news then we combined both datasets using pandas built-in function. The data set of the Fake News Project as show in the fig:

```
data=data.append(true_data,ignore_index=True)

data.head()
```

	title	text	subject	date	label
0	Donald Trump Sends Out Embarrassing New Yeará...	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 Obamaâ€™s Na...	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

Data Pre-processing

In fake dataset there are 168 columns having NaN, so we drop all of them. In-text preprocess we are cleaning our text by steaming, remove stopwords, remove special symbols and numbers, etc. After cleaning the data we have to convert this text data into numerical features using encoding technique.

Data inputs-Logic-Output Relationships

The ‘label’ column is our target variable. In this problem label column is weak correlated with the title column. The text, date and subject columns have good correlation with the target variable.

Hardware and Software Requirements and Tools used

Hardware:

- Memory 16GB minimum
- Hard Drive SSD is preferred 500GB
- Processor intel i5 minimum
- Operating system Windows 10

Software:

- Jupyter notebook (Python)

Libraries:

pandas (used to create the data and read the data)

numpy (used with the mathematical function)

seaborn (used to create a different types of graphs)

matplotlib (used to plot the graph)

regexp_tokenize (used to remove numbers and symbols)

stopwords (used to remove the unnecessary words)

train_test_split (used to split the data into train and test data)

accuracy_score (used to calculate accuracy score for train and test)

classification_report (to display precision, f1 score)

confusion_matrix (form the matrix)

roc_curve (used to plot the area under curve)

Model/s Development and Evaluation

Identification of possible problem:

We approach to both statistical and analytical problem

- ❖ Plot a bar graph for nominal data and distribution graph for continuous data.
- ❖ describe () use to calculate mean, standard deviation, minimum, maximum and quantile deviation.
- ❖ corr() used to calculate the correlation of input variable with the output variables.
- ❖ Scatter plot between target variable to the feature variables.

Testing of Identified Approaches:

Here we work on the classification problem so the machine learning models are:

- Logistic Regression
- K Neighbors Classifier
- Random Forest Classifier
- Decision Tree Classifier

Run and Evaluate selected models:

- ➢ Logistic Regression

```
x_train,x_test,y_train,y_test = train_test_split(x,y, test_size=0.30,random_state=46)
```

```
lr.fit(x_train,y_train)
y_pred1 = lr.predict(x_test)
accuracy = accuracy_score(y_test,y_pred1)*100
print("accuracy score:",accuracy)
```

```
accuracy score: 56.58044523597507
```

```
cm= confusion_matrix(y_test,y_pred1)
print(cm)
```

```
[[3600 3339]
 [2523 4014]]
```

```
clr=classification_report(y_test,y_pred1)
print(clr)
```

```
              precision    recall  f1-score   support

           0       0.59      0.52      0.55      6939
           1       0.55      0.61      0.58      6537

    accuracy                           0.57     13476
   macro avg       0.57      0.57      0.56     13476
weighted avg       0.57      0.57      0.56     13476
```

The accuracy score of logistic regression is 56.5%.And the precision is 59, recall is 52 and f1-score is 55. The sum of true negative and false negative is 5862 and the area under the curve is 56.64.

- K Neighbors Classifier

```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.30,random_state=68)
```

```
knn.fit(x_train,y_train)
y_pred2 = knn.predict(x_test)
accuracy = accuracy_score(y_test,y_pred2)*100
print("accuracy score:",accuracy)
```

```
accuracy score: 87.76343128524785
```

```
cm= confusion_matrix(y_test,y_pred2)
print(cm)
```

```
[[6360  670]
 [ 979 5467]]
```

```
clr=classification_report(y_test,y_pred2)
print(clr)
```

```
              precision    recall  f1-score   support

           0       0.87      0.90      0.89      7030
           1       0.89      0.85      0.87      6446

    accuracy                           0.88     13476
   macro avg       0.88      0.88      0.88     13476
weighted avg       0.88      0.88      0.88     13476
```

The accuracy score of K Neighbors classification is 87.76%. And the precision is 87, recall is 90 and f1-score is 89. The sum of true negative and false negative is 1649 and the area under the curve is 87.64.

- Random Forest Classifier

```
x_train,x_test,y_train,y_test = train_test_split(x,y, test_size=0.30,random_state=93)
```

```
rfc.fit(x_train,y_train)
y_pred3 = rfc.predict(x_test)
accuracy = accuracy_score(y_test,y_pred3)*100
print("accuracy score:",accuracy)
```

```
accuracy score: 99.53250222617987
```

```
cm= confusion_matrix(y_test,y_pred3)
print(cm)
```

```
[[7058   25]
 [  38 6355]]
```

```
clr=classification_report(y_test,y_pred3)
print(clr)
```

```
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      7083
           1       1.00      0.99      1.00      6393

    accuracy                           1.00     13476
   macro avg       1.00      1.00      1.00     13476
weighted avg       1.00      1.00      1.00     13476
```

The accuracy score of Random Forest classification is 99.53%. And the precision is 99, recall is 100 and f1-score is 100. The sum of true negative and false negative is 63 and the area under the curve is 99.52.

- Decision Tree Classifier

```
x_train,x_test,y_train,y_test = train_test_split(x,y, test_size=0.30,random_state=40)
```

```
clf.fit(x_train,y_train)
y_pred4 = clf.predict(x_test)
accuracy = accuracy_score(y_test,y_pred4)*100
print("accuracy score:",accuracy)
```

```
accuracy score: 99.88127040664885
```

```
cm= confusion_matrix(y_test,y_pred4)
print(cm)
```

```
[[7006   12]
 [   4 6454]]
```

```
clr=classification_report(y_test,y_pred4)
print(clr)
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      7018
           1       1.00      1.00      1.00      6458

    accuracy                           1.00     13476
   macro avg       1.00      1.00      1.00     13476
weighted avg       1.00      1.00      1.00     13476
```

The accuracy score of Decision Tree classification is 99.88%. And the precision is 100, recall is 100 and f1-score is 100. The sum of true negative and false negative is 16 and the area under the curve is 99.88.

The **Decision Tree Classifier** gives **better** accuracy score**,** precision score, recall and f1-score. The total of True Negative and False Negative in the

confusion matrix is less in the same model and area under the curve is also higher for the testing data.

Visualisation:

On visualising the continuous data we see that our target variable is balance. In subject column there are two subject that are politics, politicsnews and News, worldnews both are same so we replace politics by politicsnews and News by worldnews.

Interpretation of the Results:

On our analysis basis we go through various models and then we conclude better model on the basis of various classifications. Our data is balance so we do consider accuracy score for model testing. Then we go with the precision, f1-score, confusion matrix and area under the curve. After that we will predict the test data on the basis of train data.

Conclusion

Key Findings and Conclusions of the Study:

On study the fake data we see that there are 168 columns having missing values so we drop them and add fake data with the true data. We see that target variable is balanced. The relation of feature variables are good with the target variable but not good with each other.

Learning Outcomes of the Study in respect of Data Science:

Here we first clean the data by dropping the columns from dataset whom having huge null values. Removing the unnecessary word, symbols from the title and text. In analysis we do describe the statistical values and correlation. Fit some classification models and find the better one i.e. Decision Tree Classifier Model. Calculate accuracy score, confusion metrics, classification report and ROC curve and these are better in the same model.

Advantage in Future:

- We will not take that news correct.
- Not effect on original informations.