Approach

Fake News Detection using Python and Machine Learning

Team Name:PSA

Team Members:
1)Mohamed Aslam M
2)Sahana K
3)Prethisha V

Steps taken to achieve the solution for the given problem statement:

1. Data Collection
2. Data Preprocessing
3. Data Cleaning
4. Train and Test Split
5. Text Vectorization
6. Model Training(Logistic Regression)
7. Testing the model
8. Evaluation of model
9. Testing the model with external news

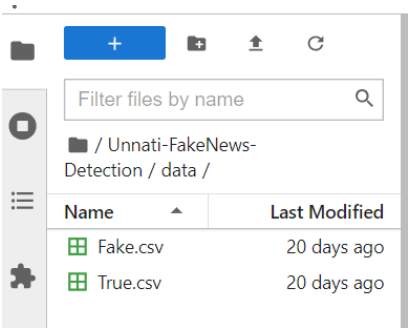
1.Data Collection:

Data Collection involves preparing/collecting the dataset which is going to be utilized in the model. The dataset to be collected varies for each model, based on the needs and constraints.

Reasoning:
The dataset collected here is ISOT Fake News Dataset.The following reasons as to why the above the dataset is preferred is:

1. Easy to acquire
2. Consists of large number of records(44898 records in total)
3. News are obtained from different legitimate sites.

Outcome:
The ISOT Fake news Dataset is downloaded and uploaded into the data directory in the Intel JupyterLab platform



2.Data Preprocessing:

Transforming the collected data for analysis.

Reasoning:
Labels are assigned to indicate the fake and true news of the dataset which are the features required.The fake and true news datasets are merged into a single dataset.

Outcome:
Labels are correctly applied to indicate the fake and true news in the dataset.The fake and true news are then combined into a single dataset named “data”.

Feature Selection:

Assigning Classes to the Dataset

```
[6]: true_data["label"] = 1
     fake_data["label"] = 0
```

Merging Both the Dataset

```
[7]: data = pd.concat([true_data, fake_data], axis=0)
```

3.Data Cleaning:

Process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in the data.

Reasoning:
Data cleaning is performed on the ISOT Fake News Dataset to remove the outliers,useless data such as numbers,non alphanumeric characters,whitespaces,urls,square brackets,HTML tags which are not essential for fake news detection.

Outcome:
The outliers,missing values,useless data are removed from the merged dataset.

Data Cleaning:

```
[13]: def preprocess_text(text):
          text=text.lower()#Text to lowercase
          text=re.sub('\[.*?\]'," ",text)#Content within square brackets are replaced
          text=re.sub('\W'," ",text)#Non alpha numeric characters are replaced
          text=re.sub('https?://\S+|www.\.\S+'," ",text)#URL starting with https/http/www are replaced
          text=re.sub('<.*?>'," ",text)#html tags are replaced
          text=re.sub('[0-9]'," ",text)#Numbers are replaced
          text=re.sub('\n'," ",text)#Newline characters are replaced
          text=re.sub('[%s]'% re.escape(string.punctuation)," ",text)#Punctuation characters are removed
          return text
      data['text'] = data['text'].apply(preprocess_text)
```

4.Train and Test Split:

The dataset is split for training and testing the model. This is done to verify whether the model generalizes well on the unseen data which is observed in later steps,by previously training the model on the training dataset.

Reasoning:
The preprocessed dataset is split into two parts:Training and Testing.Training dataset is utilized for model training and the testing dataset is utilized for testing the model with unseen data

Outcome:
The whole dataset is divided into two sections where 75% of original dataset is allocated for training and 25% is allocated for testing

```
x=data['text']
y=data['label']
```

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25)
```

5.Text Vectorization:

Text vectorization is done to convert text on the datasets to numbers.This is performed with the help of various algorithms such as Count Vectorizer,Hashing Vectorizer and TF-IDF Vectorizer

Reasoning:
Machine learning models understand numbers alone.They can’t comprehend words or string.Therefore,the news in the text column after data preprocessing is vectorized or transformed to numbers using TF-IDF Vectorizer,where it converts the words to numbers based on Total Frequency and Inverse Document Frequency of words.

Outcomes:
The news are transformed into numbers by executing the TF-IDF Vectorizer on them.

```
#TODO: explore different vectorization available with sklearn.feature_extraction.text
from sklearn.feature_extraction.text import TfidfVectorizer
Tfidf=TfidfVectorizer()
xv_train=Tfidf.fit_transform(x_train)
xv_test=Tfidf.transform(x_test)
```

6.Model Selection and Training:

This phase involves selecting the type of model that is required for us to achieve the solution.The required model varies based on the problem type and needs,for instance,for classification problems,using linear SVC, Naive Bayes,Logistic regression,Random forest are well suited.

Reasoning:
The machine learning is trained using **Logistic Regression** algorithm,a supervised machine learning algorithm which is utilized for binary classification.Here we classify the news into fake and true news,and detect the fake news.Due to news being classified into two categories(fake and true),which comes under binary classification we utilize **Logistic Regression** algorithm.

We train the model so the model can understand the relationship between the independent and dependent variables and have the ability to make future predictions.

Outcome:
The model is trained successfully using the vectorized training dataset.

```
[17]: #TODO: Model training and print the accuracy score
      LR=LogisticRegression()
      LR.fit(xv_train,y_train)
```

7.Testing the model:

Testing the model is essential as we can determine the model's predictions on unseen and can later be used for comparing its results with the actual results in the dataset.

Reasoning:
Once the model has been trained successfully,we test the model using testing dataset(unseen data) to compare the model's results with actual results.In addition,we test the model to learn the model's understanding upon the relationship between the independent and dependent variables.

Outcome:
The results of the model are stored in a variable which is later used for evaluation of the model

```
LR_prediction=LR.predict(xv_test)
```

8.Evaluation of Model:

Evaluation of model involves using metrics such as accuracy,f1 score,recall,precision etc to determine the trained model's performance.

Reasoning:
The model is evaluated to assess its performance and the accuracy of its predictions using both training and testing datasets. This evaluation helps identify the model's strengths and weaknesses. Evaluation metrics like accuracy score, classification report, and confusion matrix are utilized. The accuracy score compares the model's results with the actual results. The classification report provides information on precision, F1 score, and other performance measures. The confusion matrix displays the number of correct and incorrect predictions made by the model.

Outcome:
The model achieved an accuracy of 98.83% and the corresponding classification report and confusion matrix are displayed

```
[19]: #TODO: Model training and print the accuracy score
LR=LogisticRegression()
LR.fit(xv_train,y_train)
LR_prediction=LR.predict(xv_test)
accuracy=accuracy_score(y_test,LR_prediction)
print(accuracy)
```

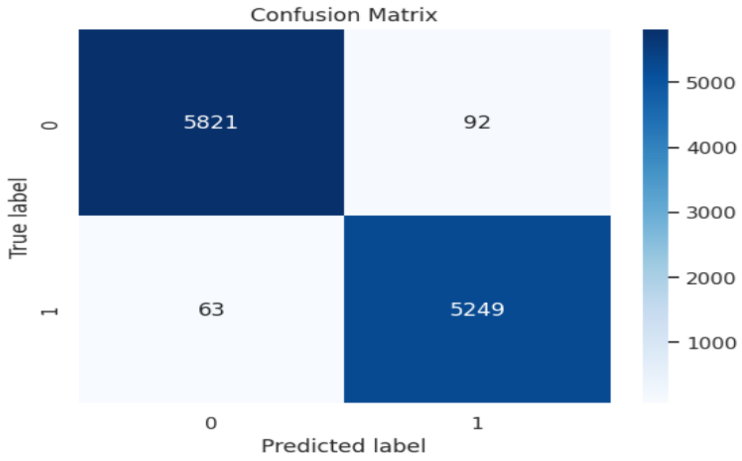
```
0.9883296213808463
```

```
[20]: # Display the Confusion matrix of Results from your classification algorithm
from sklearn.metrics import classification_report,confusion_matrix
print("Classification Report\n",classification_report(y_test,LR_prediction))
cm=confusion_matrix(y_test,LR_prediction)
sns.heatmap(cm, annot=True, cmap='Blues', fmt='d')
plt.title('Confusion Matrix')
plt.xlabel('Predicted label')
plt.ylabel('True label')
plt.show()
```

```
Classification Report
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5794
           1       0.99      0.99      0.99      5431

    accuracy                           0.99     11225
   macro avg       0.99      0.99      0.99     11225
weighted avg       0.99      0.99      0.99     11225
```



9.Testing the model with external news:

Finally,we test the model's consistency by testing the model with news of real world.

Reasoning:
The above steps explains that the model is effective in detecting fake news within the ISOT Fake News Dataset. To test the model's performance on external sources like social platforms or online channels, the input news from these sources needs to be vectorized and cleaned. To achieve this, two functions are defined: one function assigns a label indicating fake or genuine news, and the other function vectorizes and cleans the input news by removing irrelevant data and outliers. These functions enable the model to accurately process and evaluate the input news.

```
def testing(news):
    testnews={"text":[news]}
    df_test=pd.DataFrame(testnews)
    df_xtest=df_test["text"].apply(preprocess_text)
    df_xvtest=Tfidf.transform(df_xtest)
    Pred=LR.predict(df_xvtest)

    print("\n\nLogistic Regression Prediction",format(output_label(Pred[0])))
```

Outcome:

The external news are correctly detected as fake and not fake.The model performs well in detecting fake news even in real life applications.

Testing with not Fake News:

Ukraine's Zelensky says the most intense fighting is happening on the southern front From CNN staff Ukrainian President Volodymyr Zelensky says the toughest fighting is taking place on Ukraine's southern front, and he praised Kyiv's forces for holding off Russian assaults in the east. Zelensky made the comments in his daily address Sunday, saying Ukrainian troops are "advancing, position by position, step by step" and "are moving forward." The opening stages of Kyiv's counteroffensive have been marked by probing attacks – seemingly testing the Russian lines of defense – and modest gains, but no apparent major breakthroughs. Russian troops "continue to focus their main efforts on the Lyman, Bakhmut, Avdiivka and Marinka directions," the Ukrainian military's General Staff said Sunday, referring to a series of eastern Ukrainian frontline cities stretching from north to south. "Heavy fighting continues." Zelensky said that not a single US Patriot air defense system has been destroyed in Ukraine, and that nearly three dozen missiles and about 50 attack drones were destroyed over the past week. He also said the Ukrainian Air Force carried out more than 100 strikes on enemy positions over the past week. The latest from Moscow: The Russian defense ministry said in its daily report Sunday that "the Armed Forces of Ukraine are most actively advancing in the (southern) Zaporizhzhia direction, with forces of up to 3 battalion groups, reinforced with tanks and armored combat vehicles." A Russia-backed official said earlier Sunday that Ukraine has retaken a village near Zaporizhzhia city. Moscow denies the report, saying troops repelled attacks there. Russian forces also repelled eight Ukrainian army attacks in various settlements east and northeast of Donetsk city, the defense ministry claimed in its report

Logistic Regression Prediction Not a Fake news

Testing with Fake News:

News outlets around the world are reporting on the news that Pope Francis has made the unprecedented decision to endorse a US presidential candidate. His statement in support of Donald Trump was released from the Vatican this evening: "I have been hesitant to offer any kind of support for either candidate in the US presidential election but I now feel that to not voice my concern would be a dereliction of my duty as the Holy See. A strong and free America is vitally important in maintaining a strong and free world and in that sense what happens in American elections affects us all. The Rule of Law is the backbone of the American government as it is in any nation that strives for freedom and I now fear that the Rule of Law in America has been dealt a dangerous blow. The FBI, in refusing to recommend prosecution after admitting that the law had been broken on multiple occasions by Secretary Clinton, has exposed itself as corrupted by political forces that have become far too powerful. Though I don't agree with Mr. Trump on some issues, I feel that voting against the powerful political forces that have corrupted the entire American federal government is the only option for a nation that desires a government that is truly for the people and by the people. For this primary reason I ask, not as the Holy Father, but as a concerned citizen of the world that Americans vote for Donald Trump for President of the United States." Sources within the Vatican reportedly were aware that the Pope had been discussing the possibility of voicing his concern in the US presidential election but apparently were completely unaware that he had made a decision on going forward with voicing this concern until his statement was released this evening from the Vatican. Stay tuned to WTOE 5 News for more on this breaking news.

Logistic Regression Prediction Fake News