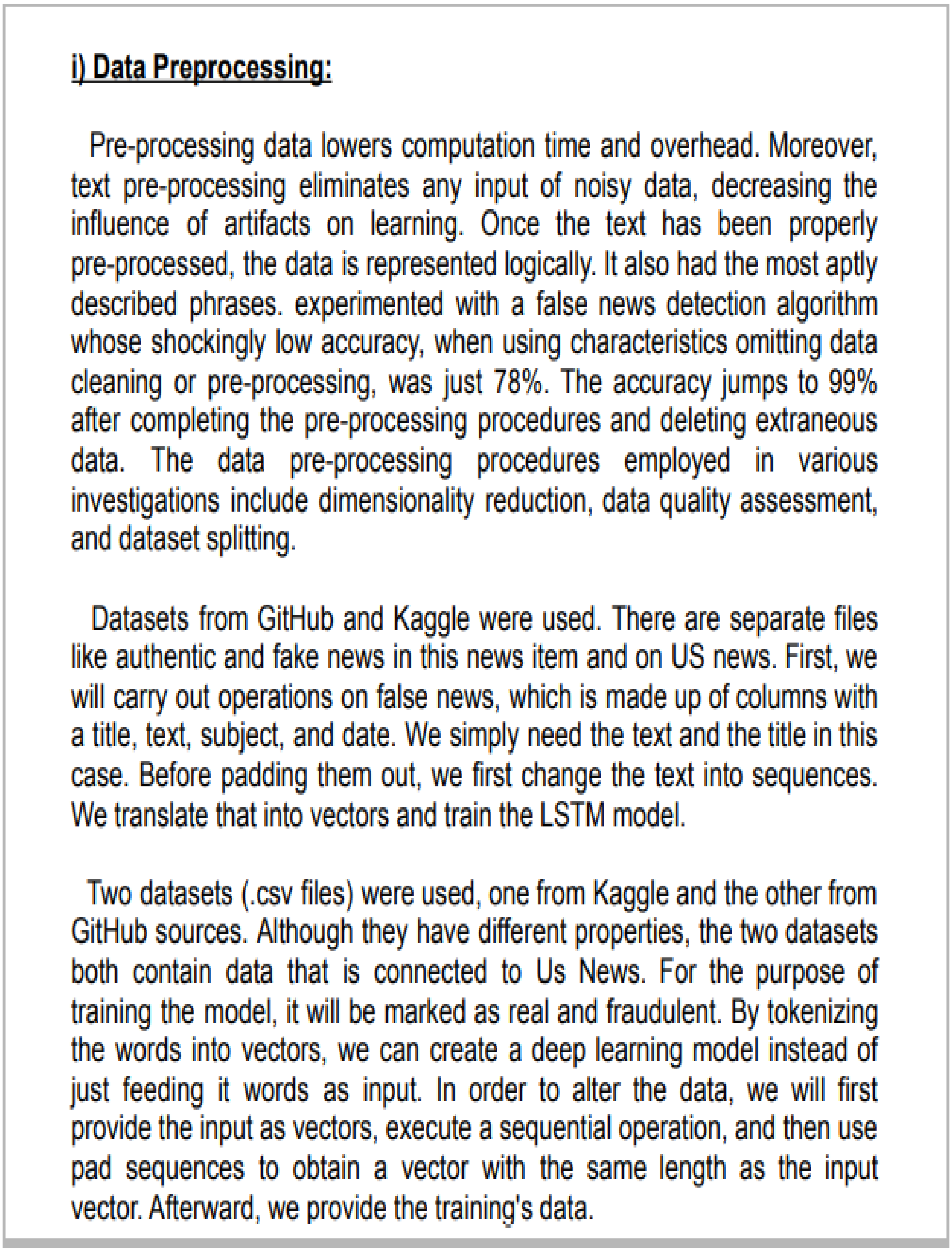
Fake News Classification using LSTM



i) Data Preprocessing:

Pre-processing data lowers computation time and overhead. Moreover, text pre-processing eliminates any input of noisy data, decreasing the influence of artifacts on learning. Once the text has been properly pre-processed, the data is represented logically. It also had the most aptly described phrases. experimented with a false news detection algorithm whose shockingly low accuracy, when using characteristics omitting data cleaning or pre-processing, was just 78%. The accuracy jumps to 99% after completing the pre-processing procedures and deleting extraneous data. The data pre-processing procedures employed in various investigations include dimensionality reduction, data quality assessment, and dataset splitting.

Datasets from GitHub and Kaggle were used. There are separate files like authentic and fake news in this news item and on US news. First, we will carry out operations on false news, which is made up of columns with a title, text, subject, and date. We simply need the text and the title in this case. Before padding them out, we first change the text into sequences. We translate that into vectors and train the LSTM model.

Two datasets (.csv files) were used, one from Kaggle and the other from GitHub sources. Although they have different properties, the two datasets both contain data that is connected to Us News. For the purpose of training the model, it will be marked as real and fraudulent. By tokenizing the words into vectors, we can create a deep learning model instead of just feeding it words as input. In order to alter the data, we will first provide the input as vectors, execute a sequential operation, and then use pad sequences to obtain a vector with the same length as the input vector. Afterward, we provide the training's data.

ii) Workflow:

1. Imported the necessary modules, libraries, and packages.

2.Datasets for the test and train data have been acquired from GitHub and Kaggle.

3. Visualizing the dataset's content using wordcloud and countplot

4.Dividing the data into test and training sets, assigning the labels to y and the total text to x.

5. Turning the text into tokens by breaking it up into words and sentences.

6. To make the array the same length, we pad the sequences.

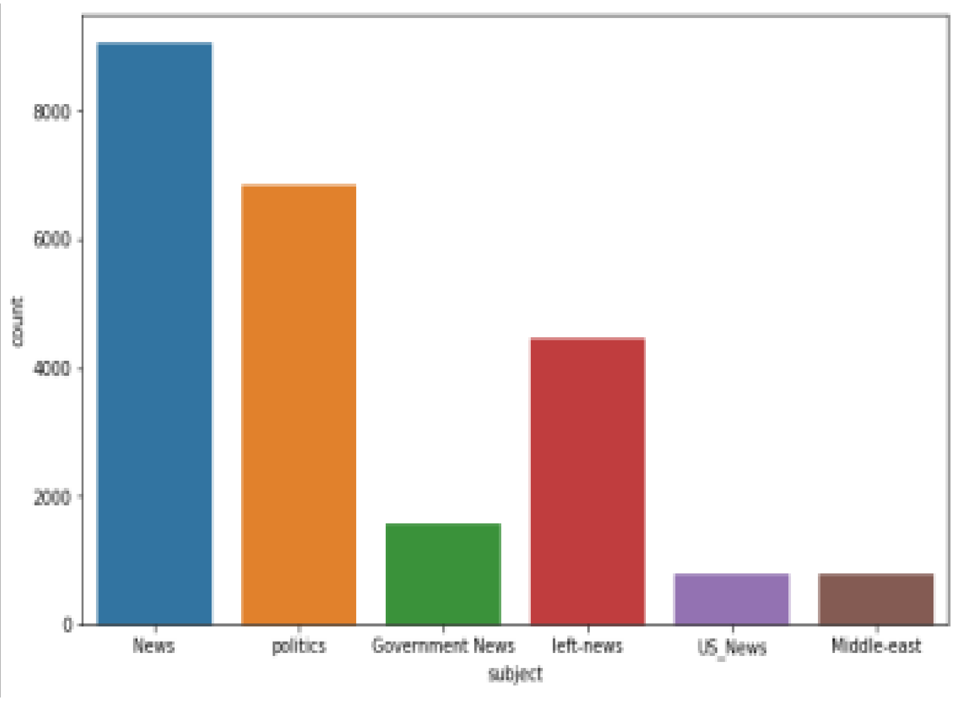
7. For the sequences following padding, a sequential model was utilized.

8. For our model, we added an embedding and an LSTM layer to the sequential model.

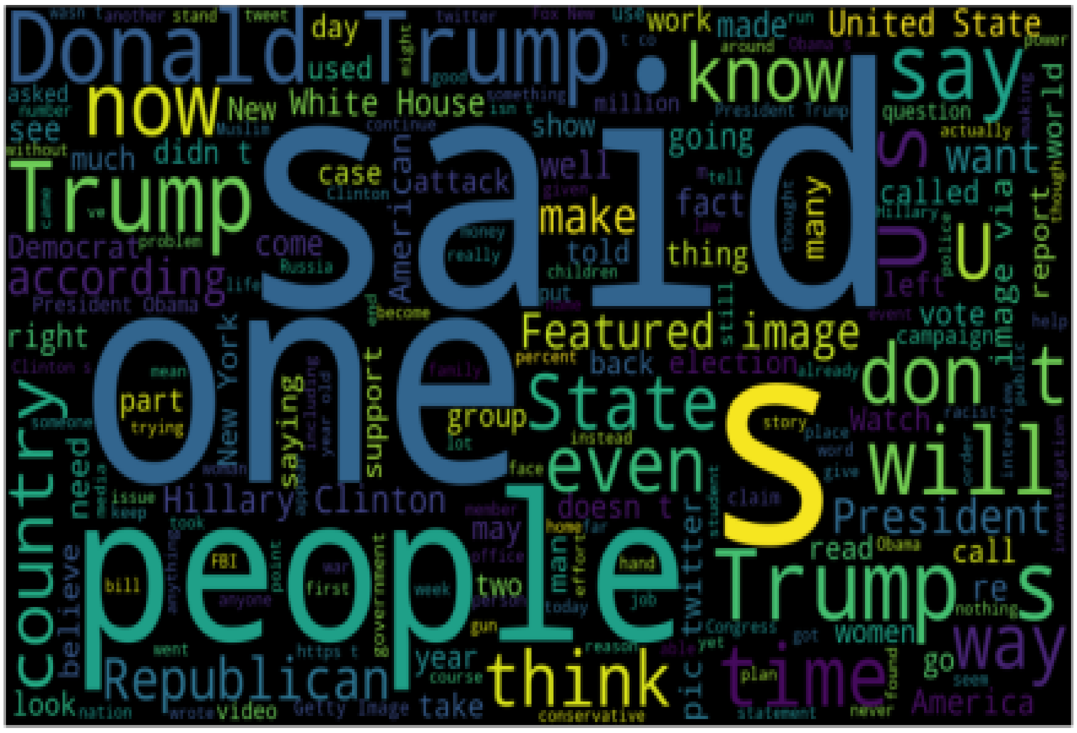
9. For our model, we will forecast the accuracy and confusion matrices.

iii) Exploratory Data Analysis (EDA):
Before performing any data preparation, we employed a number of data visualization tools, including word clouds and bar plots, to determine the word frequency and the relationship between both numerical and categorical variables.

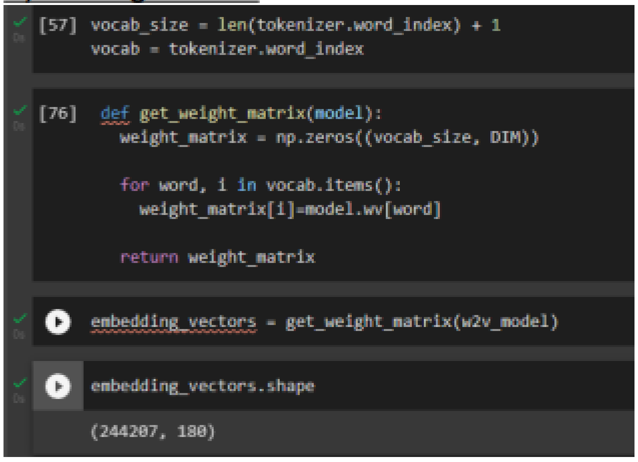
Histogram for different kinds or genres of news:



Word cloud implementation on data set:

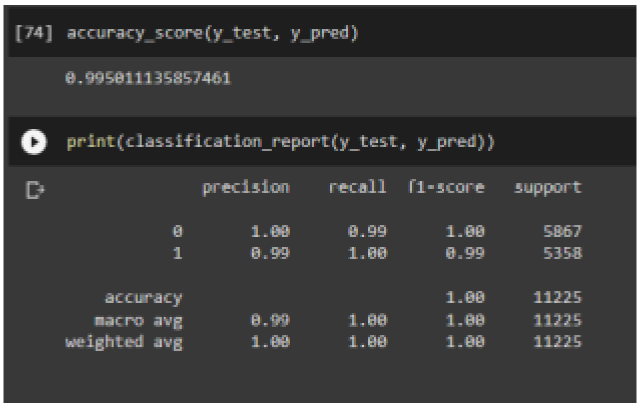


iv) Training Method:



```
[57] vocab_size = len(tokenizer.word_index) + 1
     vocab = tokenizer.word_index

[76]  def get_weight_matrix(model):
        weight_matrix = np.zeros((vocab_size, DIM))

        for word, i in vocab.items():
          weight_matrix[i]=model.wv[word]

        return weight_matrix

     embedding_vectors = get_weight_matrix(w2v_model)

     embedding_vectors.shape

     (244207, 100)
```

v)Accuracy Test:

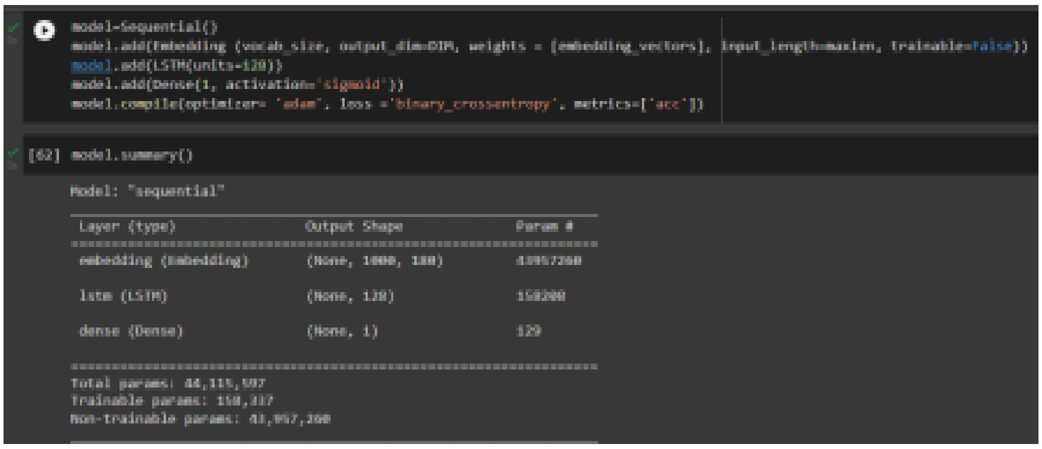


```
[74] accuracy_score(y_test, y_pred)

     0.995011135857461

     print(classification_report(y_test, y_pred))

                   precision    recall  f1-score   support

                0       1.00      0.99      1.00      5867
                1       0.99      1.00      0.99      5358

         accuracy                           1.00     11225
        macro avg       0.99      1.00      1.00     11225
     weighted avg       1.00      1.00      1.00     11225
```

V. Results and Discussion

The weights of the trained model are updated by training the model across a number of epochs. The weights and biases of the network are trained across the epochs to generate the final network. Epochs are used to increase the ultimate accuracy. Out of the LSTM models we created, LSTM had the best accuracy. So, LSTM will be the model we use.

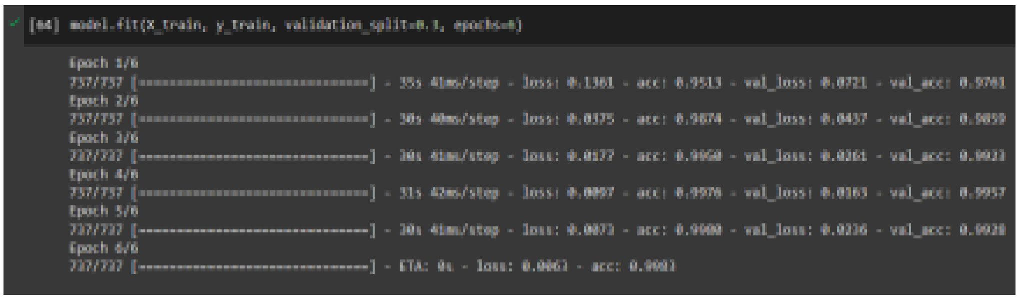
Here, following model training on the dataset. Text input that has been transformed into sequences will be provided. Then, we use our model to identify whether anything is true or false.



```
model=Sequential()
model.add(Embedding (vocab_size, output_dim=DIM, weights = [embedding_vectors], input_length=maxlen, trainable=False))
model.add(LSTM(units=128))
model.add(Dense(1, activation='sigmoid'))
model.compile(optimizer= 'adam', loss ='binary_crossentropy', metrics=['acc'])
```

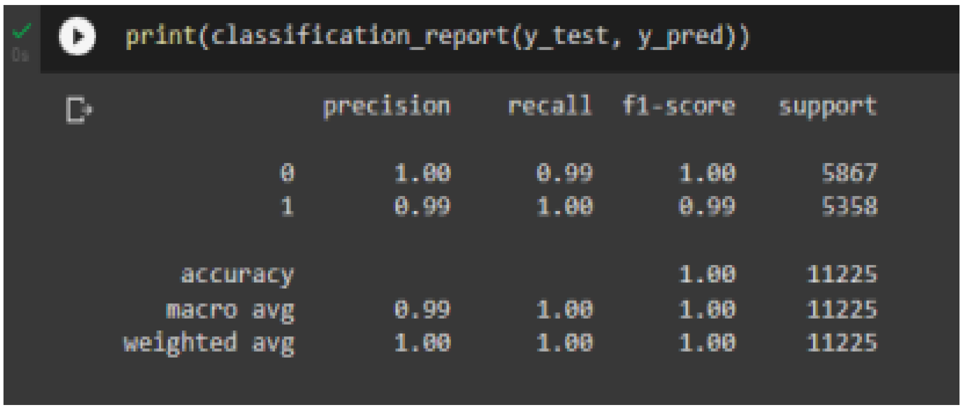
```
[62] model.summary()
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 1000, 100)         43957200

 lstm (LSTM)                 (None, 128)               158208

 dense (Dense)               (None, 1)                 129

=================================================================
Total params: 44,115,537
Trainable params: 158,337
Non-trainable params: 43,957,200
_________________________________________________________________
```



```
[64] model.fit(X_train, y_train, validation_split=0.1, epochs=6)
```

```
Epoch 1/6
737/737 [==============================] - 35s 41ms/step - loss: 0.1361 - acc: 0.9513 - val_loss: 0.0721 - val_acc: 0.9761
Epoch 2/6
737/737 [==============================] - 30s 40ms/step - loss: 0.0375 - acc: 0.9874 - val_loss: 0.0437 - val_acc: 0.9859
Epoch 3/6
737/737 [==============================] - 30s 41ms/step - loss: 0.0177 - acc: 0.9950 - val_loss: 0.0261 - val_acc: 0.9923
Epoch 4/6
737/737 [==============================] - 31s 42ms/step - loss: 0.0097 - acc: 0.9970 - val_loss: 0.0163 - val_acc: 0.9957
Epoch 5/6
737/737 [==============================] - 30s 41ms/step - loss: 0.0073 - acc: 0.9980 - val_loss: 0.0226 - val_acc: 0.9928
Epoch 6/6
737/737 [==============================] - ETA: 0s - loss: 0.0063 - acc: 0.9983
```

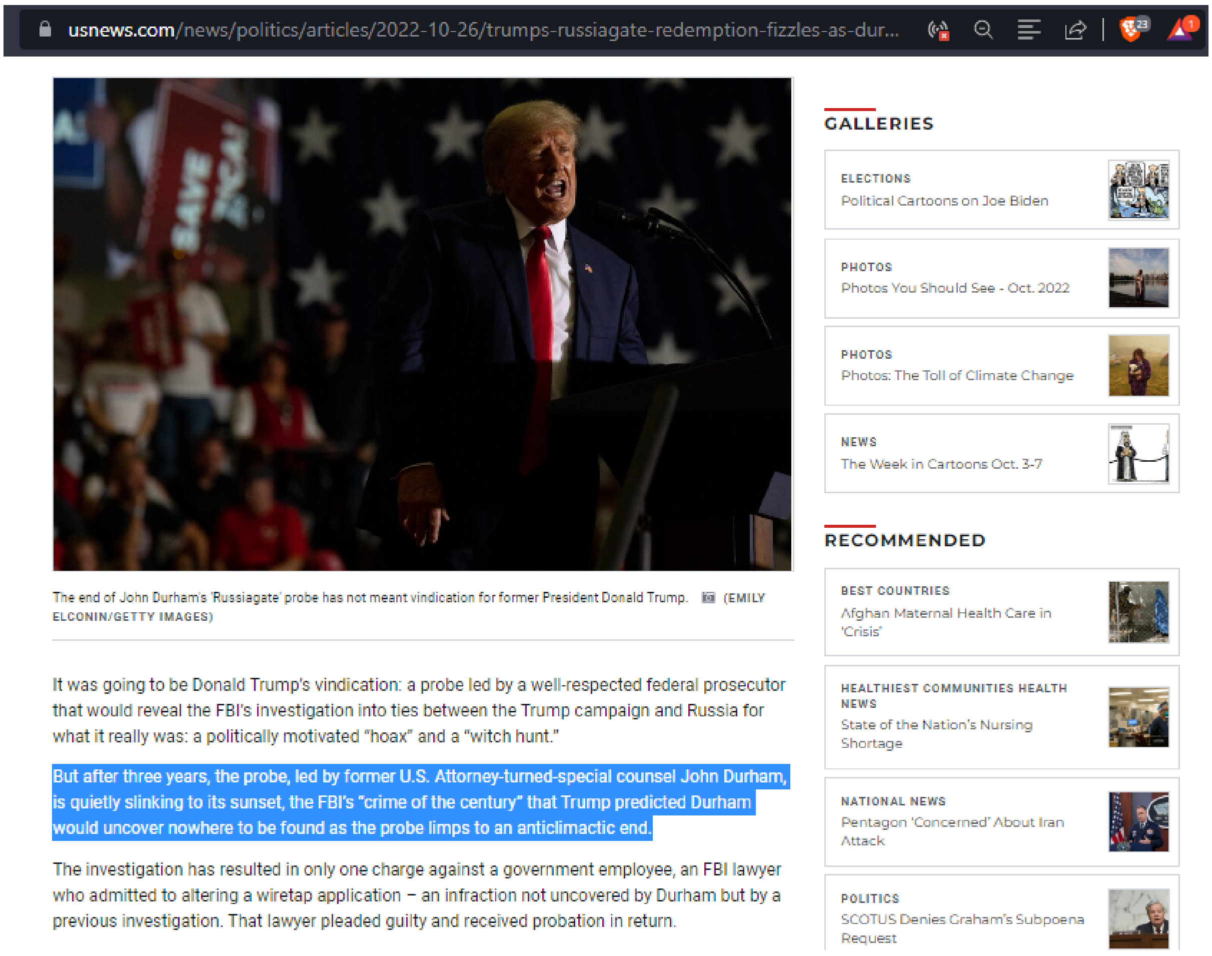


```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00      5867
           1       0.99      1.00      0.99      5358

    accuracy                           1.00     11225
   macro avg       0.99      1.00      1.00     11225
weighted avg       1.00      1.00      1.00     11225
```

<u>Testing and verification:</u>

We have taken an article of US news from a website and checking it in our classifier by pasting a part of the news in our classifier.



usnews.com/news/politics/articles/2022-10-26/trumps-russiagate-redemption-fizzles-as-dur...

The end of John Durham's 'Russiagate' probe has not meant vindication for former President Donald Trump. (EMILY ELCONIN/GETTY IMAGES)

It was going to be Donald Trump's vindication: a probe led by a well-respected federal prosecutor that would reveal the FBI's investigation into ties between the Trump campaign and Russia for what it really was: a politically motivated "hoax" and a "witch hunt."

But after three years, the probe, led by former U.S. Attorney-turned-special counsel John Durham, is quietly slinking to its sunset, the FBI's "crime of the century" that Trump predicted Durham would uncover nowhere to be found as the probe limps to an anticlimactic end.

The investigation has resulted in only one charge against a government employee, an FBI lawyer who admitted to altering a wiretap application – an infraction not uncovered by Durham but by a previous investigation. That lawyer pleaded guilty and received probation in return.

GALLERIES

ELECTIONS
Political Cartoons on Joe Biden

PHOTOS
Photos You Should See - Oct. 2022

PHOTOS
Photos: The Toll of Climate Change

NEWS
The Week in Cartoons Oct. 3-7

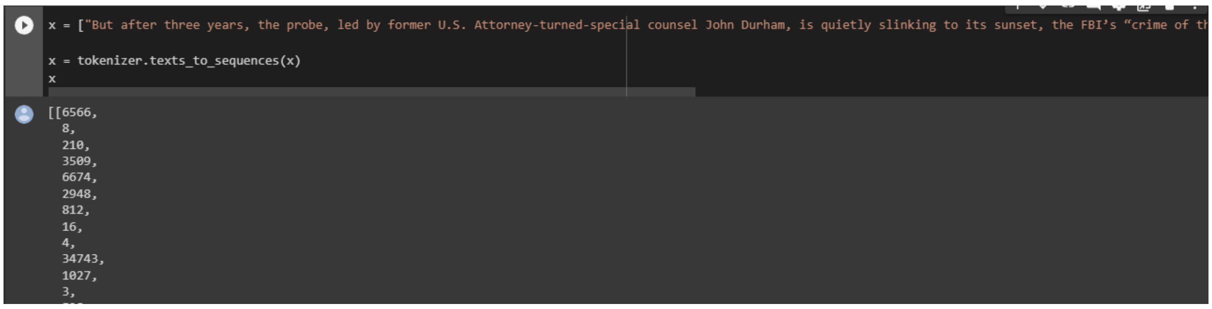
RECOMMENDED

BEST COUNTRIES
Afghan Maternal Health Care in 'Crisis'

HEALTHIEST COMMUNITIES HEALTH NEWS
State of the Nation's Nursing Shortage

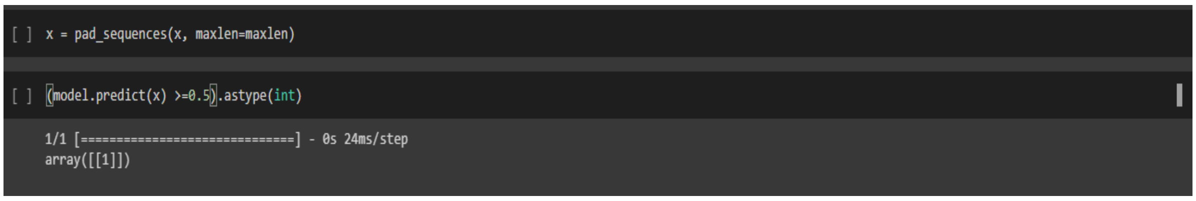
NATIONAL NEWS
Pentagon 'Concerned' About Iran Attack

POLITICS
SCOTUS Denies Graham's Subpoena Request



```
x = ["But after three years, the probe, led by former U.S. Attorney-turned-special counsel John Durham, is quietly slinking to its sunset, the FBI's "crime of th

x = tokenizer.texts_to_sequences(x)
x
```

```
[[6566,
  8,
  210,
  3509,
  6674,
  2948,
  812,
  16,
  4,
  34743,
  1027,
  3,
```



```
[ ] x = pad_sequences(x, maxlen=maxlen)

[ ] (model.predict(x) >=0.5).astype(int)

    1/1 [==============================] - 0s 24ms/step
    array([[1]])
```

We get an output of array([[1]]) which means the news is true because we have labeled real news as “1” and fake news as “0”.