Spam News Detection System

Introduction

In the digital era, the spread of fake or misleading news has become a major concern. It's often hard to tell if the news we read online is trustworthy or just clickbait meant to misinform. This project aims to build a **Spam News Detection System** that uses machine learning to help people distinguish between real and spam news articles.

We'll clean and analyze a dataset of news stories, train machine learning models to spot fake news, and evaluate how well these models perform.

In this project, we will:

1. Preprocess and clean the dataset to extract meaningful information.
2. Implement various machine learning algorithms such as Logistic Regression, Naive Bayes, and Random Forest to classify news articles.
3. Evaluate the performance of the models using metrics like accuracy, precision, recall, and F1-score.
4. Deploy the system as a useful tool to help users identify and avoid fake or spam news.

This system could be integrated with news aggregation platforms, social media websites, or as a browser extension to provide real-time feedback on the credibility of online articles.

Import Libraries

In [1]:
```
import pandas as pd
import numpy as np
```

Import datasets

In [2]:
```
True_news = pd.read_csv(r'C:\Users\bhara\Desktop\Pythonset\True.csv')
Fake_news = pd.read_csv(r'C:\Users\bhara\Desktop\Pythonset\Fake.csv')\

True_news.head()
Fake_news.head()
```

Out[2]:

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

Assign labels to recognize as TRUE/FAKE news

In [3]:
```
True_news['label'] = 0
Fake_news['label'] = 1
```

In [4]:
```
True_news.head()
Fake_news.head()
```

Out[4]:

	title	text	subject	date	label
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	1
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	1
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	1
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	1
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	1

Combining the datasets

In [5]:
```
dataset1 = True_news[['text','label']]
dataset2 = Fake_news[['text','label']]

dataset = pd.concat([dataset1,dataset2])
```

In [6]:
```
dataset
```

Out[6]:

	text	label
0	WASHINGTON (Reuters) - The head of a conservat...	0
1	WASHINGTON (Reuters) - Transgender people will...	0
2	WASHINGTON (Reuters) - The special counsel inv...	0
3	WASHINGTON (Reuters) - Trump campaign adviser ...	0
4	SEATTLE/WASHINGTON (Reuters) - President Donal...	0
...	...	...
23476	21st Century Wire says As 21WIRE reported earl...	1
23477	21st Century Wire says It s a familiar theme. ...	1
23478	Patrick Henningsen 21st Century WireRemember ...	1
23479	21st Century Wire says Al Jazeera America will...	1
23480	21st Century Wire says As 21WIRE predicted in ...	1

44898 rows × 2 columns

In [7]:
```
dataset.isnull().sum()
```

Out[7]:
```
text     0
label    0
dtype: int64
```

Shuffling Dataset

In [8]:
```
dataset = dataset.sample(frac=1)
dataset
```

Out[8]:

	text	label
12261		1
13137	KABUL (Reuters) - A top leader of militant gro...	0
20842	COPENHAGEN (Reuters) - Denmark s Prince Henrik...	0
12714	Former Bernie campaign director endorses Trump...	1
16405	House Republicans are gathering closely around...	1
...	...	...
15326	Has the Obama regime really sunk to a new low ...	1
13098		1
3509	If you haven t heard, Donald Trump has joined ...	1
18751	SEOUL (Reuters) - The United Nations nuclear w...	0
11078	Bruce Springsteen aka The Boss is showing al...	1

44898 rows × 2 columns

NLP

In [9]:
```
import nltk
import re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\bhara\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\bhara\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[9]: True

In [10]:
```
ps = WordNetLemmatizer()
stopwords = stopwords.words('english')
```

In [11]:
```
def clean_row(row):

    row = row.lower()
    row = re.sub('[^a-zA-Z]',' ',row)
    token = row.split()

    news = [ps.lemmatize(word) for word in token if not word in stopwords]
```

```
    cleaned_news = ' '.join(news)
    return cleaned_news
```

In [12]:
```
dataset['text'] = dataset['text'].apply(lambda x : clean_row(x))
dataset['text']
```

Out[12]:
```
12261
13137    kabul reuters top leader militant group al qae...
20842    copenhagen reuters denmark prince henrik husba...
12714    former bernie campaign director endorses trump...
16405    house republican gathering closely around bill...
                               ...
15326    obama regime really sunk new low level even po...
13098
3509     heard donald trump joined adolf hitler latest ...
18751    seoul reuters united nation nuclear watchdog c...
11078    bruce springsteen aka bos showing fan touch la...
Name: text, Length: 44898, dtype: object
```

In [13]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [14]:
```
vectorizer = TfidfVectorizer(max_features = 50000, lowercase = False, ngram_
```

In [15]:
```
x = dataset.iloc[:35000,0]
y = dataset.iloc[:35000, 1]
```

Train-Test split

In [16]:
```
from sklearn.model_selection import train_test_split
```

In [17]:
```
train_data, test_data, train_label, test_label = train_test_split(x,y, test_
train_data
```

Out[17]:
```
19849    huh could one damning email yet reveals true c...
3823     washington reuters president donald trump crea...
7034     black widow delivered super smackdown republic...
7773     washington reuters u president barack obama sa...
11221    charleston c washington reuters republican can...
                               ...
1778     donald trump might want stop creating public p...
3107     liberal third world hell hole ann coulter call...
11156    border patrol council president brandon judd v...
2332     washington reuters president donald trump doub...
13853    thank goodness mandatory financial disclosure ...
Name: text, Length: 28000, dtype: object
```

In [18]:
```
vec_train_data = vectorizer.fit_transform(train_data)
vec_train_data=vec_train_data.toarray()

vec_test_data = vectorizer.fit_transform(test_data)
vec_test_data = vec_test_data.toarray()
```

In [19]:
```
training_data = pd.DataFrame(vec_train_data, columns=vectorizer.get_feature_
testing_data = pd.DataFrame(vec_test_data, columns=vectorizer.get_feature_na
```

In [20]:
```
training_data
```

Out[20]:

	aa	aaf	aapl	aaron	ab	aba	aba aslani	ababa	ababa reuters	aback	...	zor
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
27995	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
27996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
27997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
27998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
27999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0

28000 rows × 50000 columns

Naive Bayes Model

In [21]:
```
from sklearn.naive_bayes import MultinomialNB
```

In [22]:
```
clf = MultinomialNB()
clf.fit(training_data,train_label)
```

Out[22]:
```
MultinomialNB
MultinomialNB()
```

In [27]:
```
from sklearn.metrics import accuracy_score

y_pred = clf.predict(testing_data)

# Calculate accuracy for test data
test_accuracy = accuracy_score(test_label, y_pred)
print(f"Testing accuracy: {test_accuracy}")

y_pred_train = clf.predict(training_data)

# Calculate accuracy for training data
```

```
train_accuracy = accuracy_score(train_label, y_pred_train)
print(f"Training accuracy: {train_accuracy}")
```

```
Testing accuracy: 0.6761428571428572
Training accuracy: 0.9586785714285714
```

Prediction

In [29]:
```
txt = 'The following statementsÂ were posted to the verified Twitter account
```

In [30]:
```
news = clean_row(txt)
news
```

Out[30]:
```
'following statement posted verified twitter account u president donald tru
mp realdonaldtrump potus opinion expressed reuters edited statement confirm
ed accuracy realdonaldtrump day trump inaugurated estimated isi fighter hel
d approx square mile territory iraq syria u military estimate remaining fig
hter occupy roughly square mile via jamiejmcintyre est left west palm beach
fire rescue met great men woman representative much u firefighter paramedic
first responder amazing people est day trump inaugurated estimated isi figh
ter held approx square mile territory iraq syria u military est remaining f
ighter occupy roughly square mile jamiejmcintyre dcexaminer est arrest m me
mber associate trump bit ly lirh b est source link bit ly jbh lu bit ly jpe
xyr'
```

In [31]:
```
pred = clf.predict(vectorizer.transform([news]).toarray())
pred
```

```
c:\Users\bhara\Desktop\Pythonset\my_venv\Lib\site-packages\sklearn\base.py:4
93: UserWarning: X does not have valid feature names, but MultinomialNB was
fitted with feature names
  warnings.warn(
```

Out[31]:
```
array([0], dtype=int64)
```

Taking input from user

In [32]:
```
txt = input("Enter news")

news = clean_row(str(txt))
pred = clf.predict(vectorizer.transform([news]).toarray())

if pred == 0:
    print("News is correct")
else:
    print("News is fake")
```

```
News is fake
```

```
c:\Users\bhara\Desktop\Pythonset\my_venv\Lib\site-packages\sklearn\base.py:4
93: UserWarning: X does not have valid feature names, but MultinomialNB was
fitted with feature names
  warnings.warn(
```

In [28]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Initialize the Logistic Regression model
clf = LogisticRegression()
clf.fit(training_data, train_label)

# Predict on the test data
y_pred = clf.predict(testing_data)
print("Logistic Regression Test Accuracy:", accuracy_score(test_label, y_pre

# Predict on the training data
y_pred_train = clf.predict(training_data)
print("Train Accuracy:", accuracy_score(train_label, y_pred_train))
```

```
Logistic Regression Test Accuracy: 0.5314285714285715
Train Accuracy: 0.9922857142857143
```

Conclusion

In this project, we developed a spam news detection system using two machine learning models: **Naive Bayes** and **Logistic Regression**. While both models performed adequately, the **Naive Bayes** model significantly outperformed the Logistic Regression model in terms of accuracy.

Given its efficiency in handling high-dimensional data, and faster training times, we have chosen to adopt **Naive Bayes** as the final model for our spam news detection system. This approach demonstrates its effectiveness in accurately classifying spam news, making it a valuable tool for future applications.