Complete Text Mining and Classification Pipeline for Fake News Detection Module: LD7185 - Programming for AI

This script covers:

1. Data loading and exploration
2. Text preprocessing
3. Feature extraction (TF-IDF)
4. Rule-based classification
5. Machine learning models (Naive Bayes, Logistic Regression, Random Forest)
6. Model evaluation and visualization

Author: Okoh Collins Date: December 2025

In [8]:
```
!python -m pip install pandas
!python -m pip install numpy
!python -m pip install matplotlib
!python -m pip install seaborn
!python -m pip install nltk
!python -m pip install re
!python -m pip install scikit-learn
!python -m pip install wordcloud
```

```
Requirement already satisfied: pandas in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (2.3.3)
Requirement already satisfied: numpy>=1.26.0 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from pandas) (2.3.5)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: six>=1.5 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
Requirement already satisfied: numpy in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (2.3.5)
Requirement already satisfied: matplotlib in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (3.10.7)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib) (1.3.3)
Requirement already satisfied: cycler>=0.10 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib) (4.61.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib) (1.4.9)
Requirement already satisfied: numpy>=1.23 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib) (2.3.5)
Requirement already satisfied: packaging>=20.0 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib) (25.0)
Requirement already satisfied: pillow>=8 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib) (12.0.0)
Requirement already satisfied: pyparsing>=3 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib) (3.2.5)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib) (2.9.0.post0)
Requirement already satisfied: six>=1.5 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from python-dateutil>=2.7->matplotlib) (1.17.0)
Requirement already satisfied: seaborn in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (0.13.2)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from seaborn) (2.3.5)
Requirement already satisfied: pandas>=1.2 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from seaborn) (2.3.3)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from seaborn) (3.10.7)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (1.3.3)
Requirement already satisfied: cycler>=0.10 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (4.61.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (1.4.9)
Requirement already satisfied: packaging>=20.0 in c:\users\dhavid\appdata\local\python\pythoncore-3.14-64\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (25.0)
```

```
Requirement already satisfied: pillow>=8 in c:\users\dhavid\appdata\local\python\pytho
ncore-3.14-64\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (12.0.0)
Requirement already satisfied: pyparsing>=3 in c:\users\dhavid\appdata\local\python\py
thoncore-3.14-64\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (3.2.5)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\dhavid\appdata\local\p
ython\pythoncore-3.14-64\lib\site-packages (from matplotlib!=3.6.1,>=3.4->seaborn) (2.
9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\dhavid\appdata\local\python\py
thoncore-3.14-64\lib\site-packages (from pandas>=1.2->seaborn) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in c:\users\dhavid\appdata\local\python
\pythoncore-3.14-64\lib\site-packages (from pandas>=1.2->seaborn) (2025.2)
Requirement already satisfied: six>=1.5 in c:\users\dhavid\appdata\local\python\python
core-3.14-64\lib\site-packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=3.4->se
aborn) (1.17.0)
Requirement already satisfied: nltk in c:\users\dhavid\appdata\local\python\pythoncore
-3.14-64\lib\site-packages (3.9.2)
Requirement already satisfied: click in c:\users\dhavid\appdata\local\python\pythoncor
e-3.14-64\lib\site-packages (from nltk) (8.3.1)
Requirement already satisfied: joblib in c:\users\dhavid\appdata\local\python\pythonco
re-3.14-64\lib\site-packages (from nltk) (1.5.2)
Requirement already satisfied: regex>=2021.8.3 in c:\users\dhavid\appdata\local\python
\pythoncore-3.14-64\lib\site-packages (from nltk) (2025.11.3)
Requirement already satisfied: tqdm in c:\users\dhavid\appdata\local\python\pythoncore
-3.14-64\lib\site-packages (from nltk) (4.67.1)
Requirement already satisfied: colorama in c:\users\dhavid\appdata\local\python\python
core-3.14-64\lib\site-packages (from click->nltk) (0.4.6)
ERROR: Could not find a version that satisfies the requirement re (from versions: non
e)
ERROR: No matching distribution found for re
```

```
Requirement already satisfied: scikit-learn in c:\users\dhavid\appdata\local\python\py
thoncore-3.14-64\lib\site-packages (1.8.0)
Requirement already satisfied: numpy>=1.24.1 in c:\users\dhavid\appdata\local\python\p
ythoncore-3.14-64\lib\site-packages (from scikit-learn) (2.3.5)
Requirement already satisfied: scipy>=1.10.0 in c:\users\dhavid\appdata\local\python\p
ythoncore-3.14-64\lib\site-packages (from scikit-learn) (1.16.3)
Requirement already satisfied: joblib>=1.3.0 in c:\users\dhavid\appdata\local\python\p
ythoncore-3.14-64\lib\site-packages (from scikit-learn) (1.5.2)
Requirement already satisfied: threadpoolctl>=3.2.0 in c:\users\dhavid\appdata\local\p
ython\pythoncore-3.14-64\lib\site-packages (from scikit-learn) (3.6.0)
Requirement already satisfied: wordcloud in c:\users\dhavid\appdata\local\python\pytho
ncore-3.14-64\lib\site-packages (1.9.4)
Requirement already satisfied: numpy>=1.6.1 in c:\users\dhavid\appdata\local\python\py
thoncore-3.14-64\lib\site-packages (from wordcloud) (2.3.5)
Requirement already satisfied: pillow in c:\users\dhavid\appdata\local\python\pythonco
re-3.14-64\lib\site-packages (from wordcloud) (12.0.0)
Requirement already satisfied: matplotlib in c:\users\dhavid\appdata\local\python\pyth
oncore-3.14-64\lib\site-packages (from wordcloud) (3.10.7)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\dhavid\appdata\local\pytho
n\pythoncore-3.14-64\lib\site-packages (from matplotlib->wordcloud) (1.3.3)
Requirement already satisfied: cycler>=0.10 in c:\users\dhavid\appdata\local\python\py
thoncore-3.14-64\lib\site-packages (from matplotlib->wordcloud) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\dhavid\appdata\local\pyth
on\pythoncore-3.14-64\lib\site-packages (from matplotlib->wordcloud) (4.61.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\dhavid\appdata\local\pyth
on\pythoncore-3.14-64\lib\site-packages (from matplotlib->wordcloud) (1.4.9)
Requirement already satisfied: packaging>=20.0 in c:\users\dhavid\appdata\local\python
\pythoncore-3.14-64\lib\site-packages (from matplotlib->wordcloud) (25.0)
Requirement already satisfied: pyparsing>=3 in c:\users\dhavid\appdata\local\python\py
thoncore-3.14-64\lib\site-packages (from matplotlib->wordcloud) (3.2.5)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\dhavid\appdata\local\p
ython\pythoncore-3.14-64\lib\site-packages (from matplotlib->wordcloud) (2.9.0.post0)
Requirement already satisfied: six>=1.5 in c:\users\dhavid\appdata\local\python\python
core-3.14-64\lib\site-packages (from python-dateutil>=2.7->matplotlib->wordcloud) (1.1
7.0)
```

In [9]:

```
# ============================================================================
# SECTION 1: IMPORT LIBRARIES
# ============================================================================

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Text preprocessing
import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

# Feature extraction
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer

# Machine learning models
```

```
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC

# Evaluation metrics
from sklearn.metrics import (accuracy_score, precision_score, recall_score,
                             f1_score, confusion_matrix, classification_report,
                             roc_auc_score, roc_curve)

# Visualization
from wordcloud import WordCloud

# Download required NLTK data
try:
    nltk.data.find('tokenizers/punkt')
except LookupError:
    nltk.download('punkt')
    nltk.download('stopwords')
    nltk.download('wordnet')
    nltk.download('omw-1.4')

print("All libraries imported successfully!")
```

```
All libraries imported successfully!
```

In [13]:

```
#load dataset
real = pd.read_csv("news/True.csv")
fake = pd.read_csv("news/Fake.csv")

#check for missing values
print("Missing Values real:")
print(real.isnull().sum())
print("\nMissing Values fake:")
print(fake.isnull().sum())

real['label'] = 1
fake['label'] = 0

df = pd.concat([real, fake]).sample(frac=1).reset_index(drop=True)
df.head()
```

```
Missing Values real:
title      0
text       0
subject    0
date       0
dtype: int64

Missing Values fake:
title      0
text       0
subject    0
date       0
dtype: int64
```

Out[13]:

	title	text	subject	date	label
0	BREAKING: PLANNED PARENTHOOD PULLS A LAME PR S...	Desperation has set in and Planned parenthood ...	politics	Jul 30, 2015	0
1	Conservative Supreme Court Ruling Just Gave T...	In what is considered a massive set-back not j...	News	February 9, 2016	0
2	OOPS! WAS ANTIFA TERRORIST Who Threatened Acid...	Antifa member, Paul Luke Kuhn who was busted...	left-news	Apr 23, 2017	0
3	Philippines arrests Indonesian pro-Islamist mi...	MANILA (Reuters) - Philippine security forces ...	worldnews	November 1, 2017	1
4	Country star Garth Brooks in talks for Trump i...	NEW YORK (Reuters) - Country star Garth Brooks...	politicsNews	December 9, 2016	1

In [16]:

```
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))

def clean_text(text):
    text = re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)
    text = re.sub(r'[^a-zA-Z ]', '', text)
    text = text.lower()
    words = text.split()
    words = [lemmatizer.lemmatize(w) for w in words if w not in stop_words]
    return " ".join(words)

df['clean_text'] = df['text'].apply(clean_text)
df.head()
```

Out[16]:

	title	text	subject	date	label	clean_text
0	BREAKING: PLANNED PARENTHOOD PULLS A LAME PR S...	Desperation has set in and Planned parenthood ...	politics	Jul 30, 2015	0	desperation set planned parenthood resorting m...
1	Conservative Supreme Court Ruling Just Gave T...	In what is considered a massive set-back not j...	News	February 9, 2016	0	considered massive setback obama environmental...
2	OOPS! WAS ANTIFA TERRORIST Who Threatened Acid...	Antifa member, Paul Luke Kuhn who was busted...	left-news	Apr 23, 2017	0	antifa member paul luke kuhn busted project ve...
3	Philippines arrests Indonesian pro-Islamist mi...	MANILA (Reuters) - Philippine security forces ...	worldnews	November 1, 2017	1	manila reuters philippine security force wedne...
4	Country star Garth Brooks in talks for Trump i...	NEW YORK (Reuters) - Country star Garth Brooks...	politicsNews	December 9, 2016	1	new york reuters country star garth brook disc...

In [18]:
```
tfidf = TfidfVectorizer(max_features=5000)
X = tfidf.fit_transform(df['clean_text'])
y = df['label']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state

model = LogisticRegression(max_iter=300)
model.fit(X_train, y_train)

pred = model.predict(X_test)
print(classification_report(y_test, pred))
```

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      4756
           1       0.98      0.99      0.99      4224

    accuracy                           0.99      8980
   macro avg       0.99      0.99      0.99      8980
weighted avg       0.99      0.99      0.99      8980
```

In [ ]:
```
```