Fake and Real News Detection System

Fake news from social media
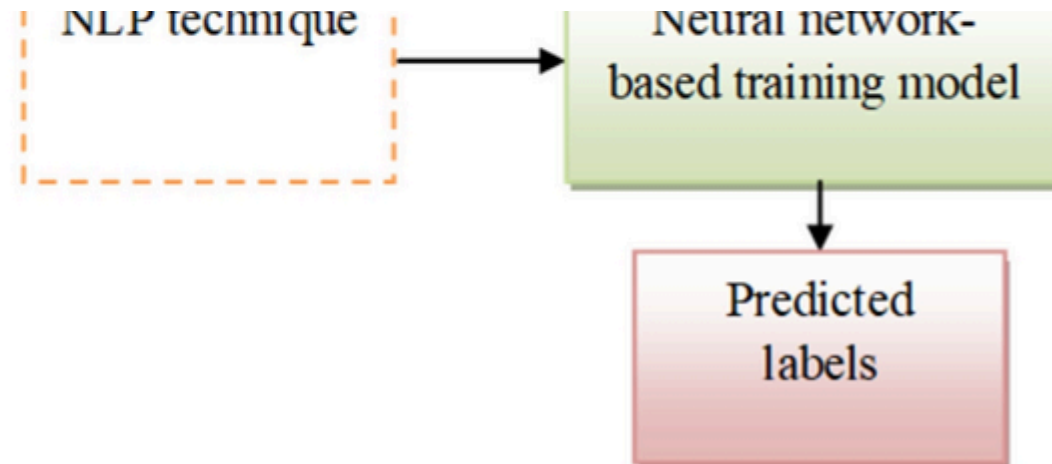
News articles

Data collection

Data pre-processing

NLP techniques

Neural networks

About Dataset Fake News detection dataset Dataset separated in two files:

Fake.csv (23502 fake news article) True.csv (21417 true news article) Dataset columns:

Title: title of news article Text: body text of news article Subject: subject of news article Date: publish date of news article

```python
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        import warnings
        warnings.filterwarnings('ignore')
```

Load both fake and real news datasets

```python
In [2]: # Load both fake and real news datasets
        fake_df = pd.read_csv('Fake.csv')
        True_df = pd.read_csv('True.csv')
```

```python
In [3]: fake_df.head()
```

Out[3]:

	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

In [4]: `fake_df.tail()`

Out[4]:

	title	text	subject	date
23476	McPain: John McCain Furious That Iran Treated …	21st Century Wire says As 21WIRE reported earl…	Middle-east	January 16, 2016
23477	JUSTICE? Yahoo Settles E-mail Privacy Class-ac…	21st Century Wire says It s a familiar theme. …	Middle-east	January 16, 2016
23478	Sunnistan: US and Allied 'Safe Zone' Plan to T…	Patrick Henningsen 21st Century WireRemember …	Middle-east	January 15, 2016
23479	How to Blow $700 Million: Al Jazeera America F…	21st Century Wire says Al Jazeera America will…	Middle-east	January 14, 2016
23480	10 U.S. Navy Sailors Held by Iranian Military …	21st Century Wire says As 21WIRE predicted in …	Middle-east	January 12, 2016

In [5]: `fake_df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23481 entries, 0 to 23480
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    23481 non-null  object
 1   text     23481 non-null  object
 2   subject  23481 non-null  object
 3   date     23481 non-null  object
dtypes: object(4)
memory usage: 733.9+ KB
```

In [6]: `fake_df.columns`

Out[6]: `Index(['title', 'text', 'subject', 'date'], dtype='object')`

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

Out[7]:
```
title      0
text       0
subject    0
date       0
dtype: int64
```

In [8]: `fake_df.duplicated().sum()`

Out[8]: `3`

In [9]: `fake_df=fake_df.drop_duplicates()`

In [10]: `fake_df.isnull().sum()`

Out[10]:
```
title      0
text       0
subject    0
date       0
dtype: int64
```

In [11]: `True_df.head()`

Out[11]:

	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

In [12]: `True_df.tail()`

Out[12]:

	title	text	subject	date
21412	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017
21413	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017
21415	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017

In [13]: `True_df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21417 entries, 0 to 21416
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    21417 non-null  object
 1   text     21417 non-null  object
 2   subject  21417 non-null  object
 3   date     21417 non-null  object
dtypes: object(4)
memory usage: 669.4+ KB
```

In [14]: `True_df.columns`

Out[14]: `Index(['title', 'text', 'subject', 'date'], dtype='object')`

In [15]: `True_df.isnull().sum()`

Out[15]:
```
title      0
text       0
subject    0
date       0
dtype: int64
```

In [16]: `True_df.shape`

Out[16]: (21417, 4)

In [17]: `True_df.duplicated().sum()`

Out[17]: 206

In [18]: `True_df=True_df.drop_duplicates()`

In [19]: `True_df.isnull().sum()`

Out[19]:
```
title      0
text       0
subject    0
date       0
dtype: int64
```

In [20]: `True_df.info()`

```
<class 'pandas.core.frame.DataFrame'>
Index: 21211 entries, 0 to 21416
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    21211 non-null  object
 1   text     21211 non-null  object
 2   subject  21211 non-null  object
 3   date     21211 non-null  object
dtypes: object(4)
memory usage: 828.6+ KB
```

In [21]:
```
True_df['label'] = 1
fake_df['label'] = 0
```

In [22]: `df=pd.concat([fake_df,True_df])`

In [79]: `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44689 entries, 0 to 44688
Data columns (total 6 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    44689 non-null  object
 1   text     44689 non-null  object
 2   subject  44689 non-null  object
 3   date     44689 non-null  object
 4   label    44689 non-null  int64
 5   news     44689 non-null  object
dtypes: int64(1), object(5)
memory usage: 2.0+ MB
```

In [23]: `df`

Out[23]:

	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	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 Obama's Name…	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
…	…	…	…	…	…
21412	'Fully committed' NATO backs new U.S. approach…	BRUSSELS (Reuters) - NATO allies on Tuesday we…	worldnews	August 22, 2017	1
21413	LexisNexis withdrew two products from Chinese …	LONDON (Reuters) - LexisNexis, a provider of l…	worldnews	August 22, 2017	1
21414	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov…	worldnews	August 22, 2017	1
21415	Vatican upbeat on possibility of Pope Francis …	MOSCOW (Reuters) - Vatican Secretary of State …	worldnews	August 22, 2017	1
21416	Indonesia to buy $1.14 billion worth of Russia…	JAKARTA (Reuters) - Indonesia will buy 11 Sukh…	worldnews	August 22, 2017	1

44689 rows × 5 columns

In [24]:
```python
df = df.sample(frac=1).reset_index(drop=True)
```

In [25]:
```python
df
```

Out[25]:

	title	text	subject	date	label
0	23 YR-OLD "FEMALE" ANTIFA THUG Charged With St...	These are literally the same people carrying s...	left-news	Jul 9, 2017	0
1	Pope candidly admits Church 'arrived late' in ...	VATICAN CITY (Reuters) - Pope Francis, in some...	worldnews	September 21, 2017	1
2	High court emissions ruling won't deter clean ...	WASHINGTON (Reuters) - The U.S. power sector's...	politicsNews	February 11, 2016	1
3	LEFTIST Resentment Peddler, Larry David (Net W...	And why wouldn t he? The leftist comedy writer...	left-news	Nov 8, 2015	0
4	NZ kingmaker calls for inquiry over China spy ...	(Reuters) - A populist politician whose party ...	worldnews	September 14, 2017	1
...	...	...	...	...	...
44684	SHOCKING! HERE'S HOW OFTEN Airlines Bump Passe...	Who knew that the Department of Transportation...	Government News	Apr 11, 2017	0
44685	Texas Republican Introduces Bill To Put Women...	Women in Texas may have to literally fight for...	News	January 24, 2017	0
44686	JUDGE JEANINE PIRRO: "We're tired of being lec...	Next stop after BREXIT is the US! Judge Jeanin...	politics	Jun 27, 2016	0
44687	Mainstream Media Completely SILENT As Obamaca...	President Obama made history yet again by maki...	News	April 15, 2016	0
44688	Poland's PiS may sack PM Szydlo, replace with ...	WARSAW (Reuters) - Poland s ruling Law and Jus...	worldnews	December 5, 2017	1

44689 rows × 5 columns

In [26]: `df.shape`

Out[26]: (44689, 5)

In [27]: `df.columns`

```
Out[27]:  Index(['title', 'text', 'subject', 'date', 'label'], dtype='object')
```

```
In [28]:  df['label'].unique()
```

```
Out[28]:  array([0, 1], dtype=int64)
```

```
In [29]:  df['label'].value_counts()
```

```
Out[29]:  label
          0    23478
          1    21211
          Name: count, dtype: int64
```

```
In [30]:  import seaborn as sns
          import matplotlib.pyplot as plt

          sns.set_theme(style="whitegrid")

          plt.figure(figsize=(10, 6))
          ax = sns.countplot(data=df, x='label', palette='pastel', edgecolor='black')

          total = len(df)
          for p in ax.patches:
              count = p.get_height()
              percentage = f'{100 * count / total:.1f}%'
              x = p.get_x() + p.get_width() / 2
              y = p.get_height()
              ax.annotate(percentage, (x, y), ha='center', va='bottom', fontsize=12, fontweight='bold')

          plt.title('Distribution of Labels', fontsize=16, fontweight='bold')
          plt.xlabel('Label', fontsize=14)
          plt.ylabel('Count', fontsize=14)
          plt.xticks(fontsize=12)
          plt.yticks(fontsize=12)
          plt.tight_layout()
          plt.show()

          plt.figure(figsize=(8, 8))
          colors = sns.color_palette('pastel')
          label_counts = df['label'].value_counts()
```
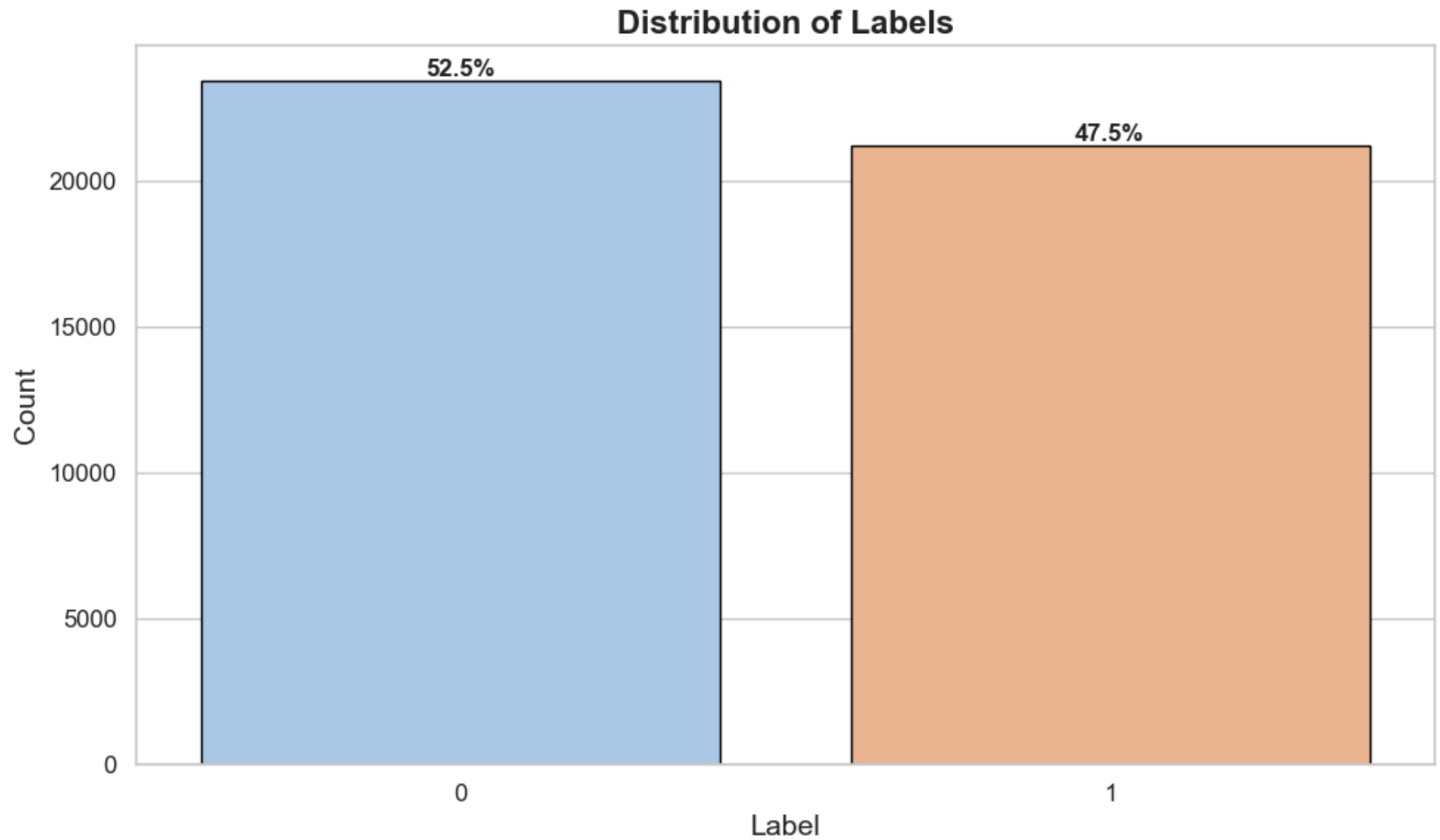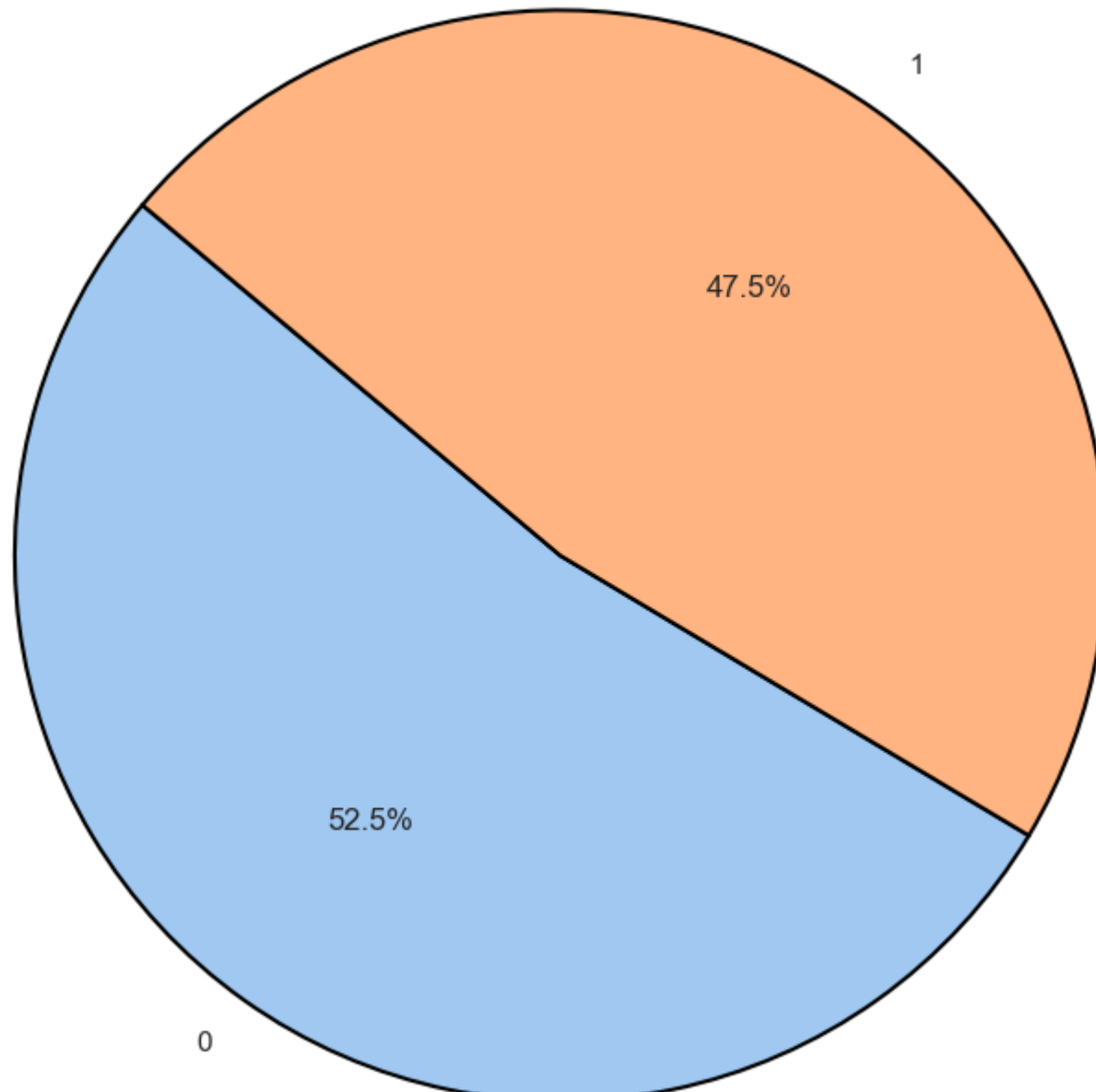
```
plt.pie(label_counts, labels=label_counts.index, autopct='%1.1f%%', startangle=140, colors=colors,
        wedgeprops={'edgecolor': 'black', 'linewidth': 1.5})
plt.title('Label Proportion', fontsize=16, fontweight='bold')
plt.tight_layout()
plt.show()
```

Distribution of Labels

Label Proportion

In [31]: `df['subject'].unique()`

Out[31]: `array(['left-news', 'worldnews', 'politicsNews', 'politics', 'US_News',`
         `       'News', 'Government News', 'Middle-east'], dtype=object)`

In [32]: `df['subject'].value_counts()`

Out[32]:
```
subject
politicsNews       11220
worldnews           9991
News                9050
politics            6838
left-news           4459
Government News     1570
US_News              783
Middle-east          778
Name: count, dtype: int64
```
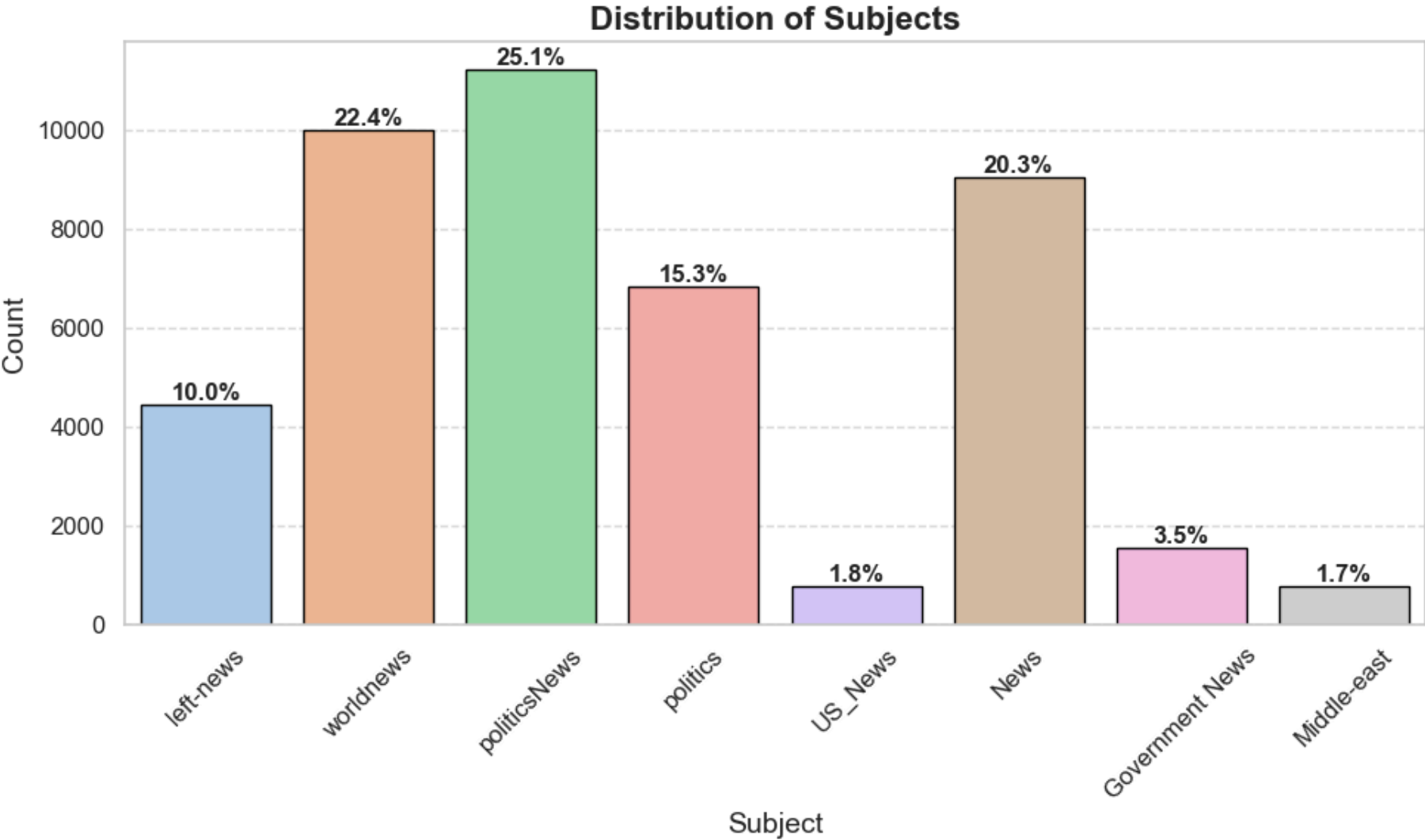
In [33]:
```python
plt.figure(figsize=(10, 6))
ax = sns.countplot(data=df, x='subject', palette='pastel', edgecolor='black')

total = len(df)
for p in ax.patches:
    count = p.get_height()
    percentage = f'{100 * count / total:.1f}%'
    x = p.get_x() + p.get_width() / 2
    y = p.get_height()
    ax.annotate(percentage, (x, y), ha='center', va='bottom', fontsize=12, fontweight='bold')

plt.title('Distribution of Subjects', fontsize=16, fontweight='bold')
plt.xlabel('Subject', fontsize=14)
plt.ylabel('Count', fontsize=14)
plt.xticks(rotation=45, fontsize=12)
plt.yticks(fontsize=12)
```

```python
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

plt.figure(figsize=(15, 15))
colors = sns.color_palette('pastel')
subject_counts = df['subject'].value_counts()

plt.pie(subject_counts, labels=subject_counts.index, autopct='%1.1f%%', startangle=140, colors=colors,
        wedgeprops={'edgecolor': 'black', 'linewidth': 1.5})
plt.title('Subject Proportion', fontsize=16, fontweight='bold')
plt.tight_layout()
plt.show()
```
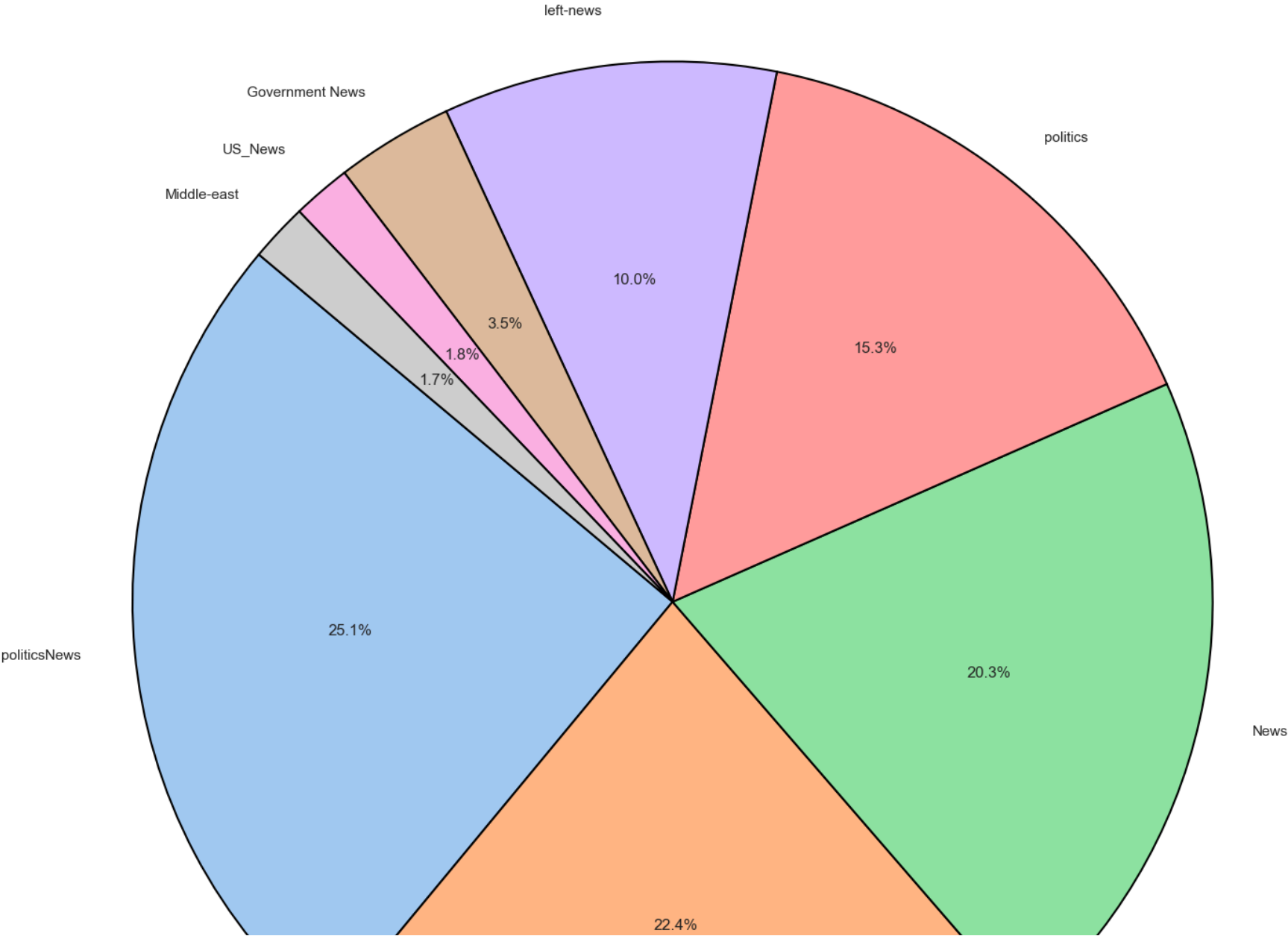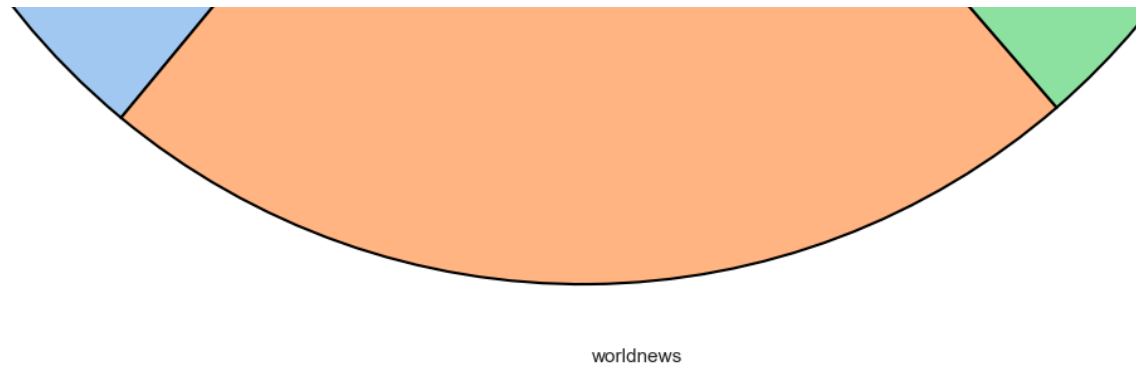
Distribution of Subjects

Subject Proportion

worldnews

In [34]: ```python
df['news']=df['title'] + ' ' +df['text']
```

In [35]: ```python
df
```

Out[35]:

	title	text	subject	date	label	news
0	23 YR-OLD "FEMALE" ANTIFA THUG Charged With St…	These are literally the same people carrying s…	left-news	Jul 9, 2017	0	23 YR-OLD "FEMALE" ANTIFA THUG Charged With St…
1	Pope candidly admits Church 'arrived late' in …	VATICAN CITY (Reuters) - Pope Francis, in some…	worldnews	September 21, 2017	1	Pope candidly admits Church 'arrived late' in …
2	High court emissions ruling won't deter clean …	WASHINGTON (Reuters) - The U.S. power sector's…	politicsNews	February 11, 2016	1	High court emissions ruling won't deter clean …
3	LEFTIST Resentment Peddler, Larry David (Net W…	And why wouldn t he? The leftist comedy writer…	left-news	Nov 8, 2015	0	LEFTIST Resentment Peddler, Larry David (Net W…
4	NZ kingmaker calls for inquiry over China spy …	(Reuters) - A populist politician whose party …	worldnews	September 14, 2017	1	NZ kingmaker calls for inquiry over China spy …
...	...	...	...	...	...	...
44684	SHOCKING! HERE'S HOW OFTEN Airlines Bump Passe…	Who knew that the Department of Transportation…	Government News	Apr 11, 2017	0	SHOCKING! HERE'S HOW OFTEN Airlines Bump Passe…
44685	Texas Republican Introduces Bill To Put Women…	Women in Texas may have to literally fight for…	News	January 24, 2017	0	Texas Republican Introduces Bill To Put Women…
44686	JUDGE JEANINE PIRRO: "We're tired of being lec…	Next stop after BREXIT is the US! Judge Jeanin…	politics	Jun 27, 2016	0	JUDGE JEANINE PIRRO: "We're tired of being lec…
44687	Mainstream Media Completely SILENT As Obamaca…	President Obama made history yet again by maki…	News	April 15, 2016	0	Mainstream Media Completely SILENT As Obamaca…
44688	Poland's PiS may sack PM Szydlo, replace with …	WARSAW (Reuters) - Poland s ruling Law and Jus…	worldnews	December 5, 2017	1	Poland's PiS may sack PM Szydlo, replace with …

44689 rows × 6 columns

In [36]: ```df['news'].head()```

Out[36]:   0   23 YR-OLD "FEMALE" ANTIFA THUG Charged With St...
            1   Pope candidly admits Church 'arrived late' in ...
            2   High court emissions ruling won't deter clean ...
            3   LEFTIST Resentment Peddler, Larry David (Net W...
            4   NZ kingmaker calls for inquiry over China spy ...
            Name: news, dtype: object

In [37]:
```python
df_new=df.copy()
```

In [38]:
```python
def clean_text(text):
    text=text.lower()
    return text.strip()
```

In [39]:
```python
df_new['news']=df_new['news'].apply(lambda x: clean_text(x))
```

In [40]:
```python
df_new['news']
```

Out[40]:   0   23 yr-old "female" antifa thug charged with st...
            1   pope candidly admits church 'arrived late' in ...
            2   high court emissions ruling won't deter clean ...
            3   leftist resentment peddler, larry david (net w...
            4   nz kingmaker calls for inquiry over china spy ...
                                          ...
            44684   shocking! here's how often airlines bump passe...
            44685   texas republican introduces bill to put women ...
            44686   judge jeanine pirro: "we're tired of being lec...
            44687   mainstream media completely silent as obamacar...
            44688   poland's pis may sack pm szydlo, replace with ...
            Name: news, Length: 44689, dtype: object

In [41]:
```python
import string
string.punctuation
```

Out[41]:   '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'

In [42]:
```python
def remove_punctuation(text):
    punctuationfree="".join([i for i in text if i not in string.punctuation])
    return punctuationfree
```

```
In [43]: df_new['news']=df_new['news'].apply(lambda x:remove_punctuation(x))
```

```
In [44]: df_new['news']
```

```
Out[44]: 0        23 yrold "female" antifa thug charged with sta...
         1        pope candidly admits church arrived late in co...
         2        high court emissions ruling wont deter clean e...
         3        leftist resentment peddler larry david net wor...
         4        nz kingmaker calls for inquiry over china spy ...
                                        ...
         44684    shocking here's how often airlines bump passen...
         44685    texas republican introduces bill to put women ...
         44686    judge jeanine pirro "we're tired of being lect...
         44687    mainstream media completely silent as obamacar...
         44688    polands pis may sack pm szydlo replace with fi...
         Name: news, Length: 44689, dtype: object
```

```
In [45]: def remove_punctuation(text):
             translator = str.maketrans('', '', string.punctuation)
             return text.translate(translator)
```

```
In [46]: df_new['news']=df_new['news'].apply(lambda x:remove_punctuation(x))
```

```
In [47]: df_new['news']
```

```
Out[47]: 0        23 yrold "female" antifa thug charged with sta...
         1        pope candidly admits church arrived late in co...
         2        high court emissions ruling wont deter clean e...
         3        leftist resentment peddler larry david net wor...
         4        nz kingmaker calls for inquiry over china spy ...
                                        ...
         44684    shocking here's how often airlines bump passen...
         44685    texas republican introduces bill to put women ...
         44686    judge jeanine pirro "we're tired of being lect...
         44687    mainstream media completely silent as obamacar...
         44688    polands pis may sack pm szydlo replace with fi...
         Name: news, Length: 44689, dtype: object
```

In [48]:
```python
import re
def remove_punctuation(text):
    return re.sub(r'[^\w\s]', '', text)
```

In [49]:
```python
df_new['news']=df_new['news'].apply(lambda x:remove_punctuation(x))
```

In [50]:
```python
df_new['news']
```

Out[50]:
```
0          23 yrold female antifa thug charged with stabb...
1          pope candidly admits church arrived late in co...
2          high court emissions ruling wont deter clean e...
3          leftist resentment peddler larry david net wor...
4          nz kingmaker calls for inquiry over china spy ...
                                 ...
44684      shocking heres how often airlines bump passeng...
44685      texas republican introduces bill to put women ...
44686      judge jeanine pirro were tired of being lectur...
44687      mainstream media completely silent as obamacar...
44688      polands pis may sack pm szydlo replace with fi...
Name: news, Length: 44689, dtype: object
```

In [51]:
```python
import nltk
from wordcloud import WordCloud
```

In [52]:
```python
stopwords=nltk.corpus.stopwords.words('english')
```

In [53]:
```python
def remove_stopwords(text):
    output = " ".join(i for i in text.split() if i not in stopwords)
    return output

df_new['news'] = df_new['news'].apply(lambda x: remove_stopwords(x))
```

In [54]:
```python
df_new['news']
```

Out[54]: 0   23 yrold female antifa thug charged stabbing h...
          1   pope candidly admits church arrived late confr...
          2   high court emissions ruling wont deter clean e...
          3   leftist resentment peddler larry david net wor...
          4   nz kingmaker calls inquiry china spy report re...
                                        ...
          44684   shocking heres often airlines bump passengers ...
          44685   texas republican introduces bill put women pri...
          44686   judge jeanine pirro tired lectured fat cats wa...
          44687   mainstream media completely silent obamacare m...
          44688   polands pis may sack pm szydlo replace finance...
          Name: news, Length: 44689, dtype: object

In [55]:
```python
import nltk
nltk.download('punkt', force=True)
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\hi\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

Out[55]: True

In [56]:
```python
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer

stemmer = PorterStemmer()

def stem_sentence(sentence):
    if isinstance(sentence, str):
        words = word_tokenize(sentence, preserve_line=True)  # Avoid sentence splitting
        stemmed_words = [stemmer.stem(word) for word in words]
        return ' '.join(stemmed_words)
    return sentence
```

In [57]:
```python
df_new['news']
```

```
Out[57]: 0          23 yrold female antifa thug charged stabbing h...
         1          pope candidly admits church arrived late confr...
         2          high court emissions ruling wont deter clean e...
         3          leftist resentment peddler larry david net wor...
         4          nz kingmaker calls inquiry china spy report re...
                                         ...
         44684      shocking heres often airlines bump passengers ...
         44685      texas republican introduces bill put women pri...
         44686      judge jeanine pirro tired lectured fat cats wa...
         44687      mainstream media completely silent obamacare m...
         44688      polands pis may sack pm szydlo replace finance...
         Name: news, Length: 44689, dtype: object
```

```python
In [58]: def clean_text(text):
             text = re.sub('\[.*\]','', text).strip()
             text = re.sub('\S*\d\S*\s*','', text).strip()
             return text.strip()
```

```python
In [59]: df_new['news'] = df_new['news'].apply(lambda x: clean_text(x))
```

```python
In [60]: df_new['news']
```

```
Out[60]: 0          yrold female antifa thug charged stabbing harr...
         1          pope candidly admits church arrived late confr...
         2          high court emissions ruling wont deter clean e...
         3          leftist resentment peddler larry david net wor...
         4          nz kingmaker calls inquiry china spy report re...
                                         ...
         44684      shocking heres often airlines bump passengers ...
         44685      texas republican introduces bill put women pri...
         44686      judge jeanine pirro tired lectured fat cats wa...
         44687      mainstream media completely silent obamacare m...
         44688      polands pis may sack pm szydlo replace finance...
         Name: news, Length: 44689, dtype: object
```

```python
In [61]: def remove_urls(vTEXT):
             vTEXT = re.sub(r'(https|http)?:\/\/(\w|\.|\/|\?|\=|\&|\%)*\b', '', vTEXT, flags=re.MULTILINE)
             return(vTEXT)
```

In [62]:
```python
df_new['news'] = df_new['news'].apply(lambda x: remove_urls(x))
```

In [63]:
```python
df_new['news'] = df_new['news'].apply(lambda x: remove_urls(x))
```

In [64]:
```python
def remove_digits(text):
    clean_text = re.sub(r"\b[0-9]+\b\s*", "", text)
    return(text)
```

In [65]:
```python
df_new['news'] = df_new['news'].apply(lambda x: remove_digits(x))
```

In [66]:
```python
df_new['news']
```

Out[66]:
```
0        yrold female antifa thug charged stabbing harr...
1        pope candidly admits church arrived late confr...
2        high court emissions ruling wont deter clean e...
3        leftist resentment peddler larry david net wor...
4        nz kingmaker calls inquiry china spy report re...
                               ...
44684    shocking heres often airlines bump passengers ...
44685    texas republican introduces bill put women pri...
44686    judge jeanine pirro tired lectured fat cats wa...
44687    mainstream media completely silent obamacare m...
44688    polands pis may sack pm szydlo replace finance...
Name: news, Length: 44689, dtype: object
```

In [67]:
```python
def remove_emojis(data):
    emoji_pattern = re.compile("["
                               u"\U0001F600-\U0001F64F"
                               u"\U0001F300-\U0001F5FF"
                               u"\U0001F680-\U0001F6FF"
                               u"\U0001F1E0-\U0001F1FF"
                               "]+", flags=re.UNICODE)
    return re.sub(emoji_pattern, '', data)
```

In [68]:
```python
df_new['news'] = df_new['news'].apply(lambda x: remove_emojis(x))
```

In [69]:
```python
df_new['news']
```

```
Out[69]:  0          yrold female antifa thug charged stabbing harr...
          1          pope candidly admits church arrived late confr...
          2          high court emissions ruling wont deter clean e...
          3          leftist resentment peddler larry david net wor...
          4          nz kingmaker calls inquiry china spy report re...
                                         ...
          44684      shocking heres often airlines bump passengers ...
          44685      texas republican introduces bill put women pri...
          44686      judge jeanine pirro tired lectured fat cats wa...
          44687      mainstream media completely silent obamacare m...
          44688      polands pis may sack pm szydlo replace finance...
          Name: news, Length: 44689, dtype: object
```

```python
In [70]:  df_new['news'] = df_new['news'].apply(lambda x:re.sub(r'\s+[a-zA-Z]\s+', '', x))
```

```python
In [71]:  df_new['news'] = df_new['news'].apply(lambda x:re.sub(r'\s+', ' ', x, flags=re.I))
```

```python
In [72]:  df_new
```

Out[72]:

	title	text	subject	date	label	news
0	23 YR-OLD "FEMALE" ANTIFA THUG Charged With St…	These are literally the same people carrying s…	left-news	Jul 9, 2017	0	yrold female antifa thug charged stabbing harr…
1	Pope candidly admits Church 'arrived late' in …	VATICAN CITY (Reuters) - Pope Francis, in some…	worldnews	September 21, 2017	1	pope candidly admits church arrived late confr…
2	High court emissions ruling won't deter clean …	WASHINGTON (Reuters) - The U.S. power sector's…	politicsNews	February 11, 2016	1	high court emissions ruling wont deter clean e…
3	LEFTIST Resentment Peddler, Larry David (Net W…	And why wouldn t he? The leftist comedy writer…	left-news	Nov 8, 2015	0	leftist resentment peddler larry david net wor…
4	NZ kingmaker calls for inquiry over China spy …	(Reuters) - A populist politician whose party …	worldnews	September 14, 2017	1	nz kingmaker calls inquiry china spy report re…
…	…	…	…	…	…	…
44684	SHOCKING! HERE'S HOW OFTEN Airlines Bump Passe…	Who knew that the Department of Transportation…	Government News	Apr 11, 2017	0	shocking heres often airlines bump passengers …
44685	Texas Republican Introduces Bill To Put Women…	Women in Texas may have to literally fight for…	News	January 24, 2017	0	texas republican introduces bill put women pri…
44686	JUDGE JEANINE PIRRO: "We're tired of being lec…	Next stop after BREXIT is the US! Judge Jeanin…	politics	Jun 27, 2016	0	judge jeanine pirro tired lectured fat cats wa…
44687	Mainstream Media Completely SILENT As Obamaca…	President Obama made history yet again by maki…	News	April 15, 2016	0	mainstream media completely silent obamacare m…
44688	Poland's PiS may sack PM Szydlo, replace with …	WARSAW (Reuters) - Poland s ruling Law and Jus…	worldnews	December 5, 2017	1	polands pis may sack pm szydlo replace finance…

44689 rows × 6 columns

In [73]:
```python
from sklearn.feature_extraction.text import TfidfVectorizer


tfidf = TfidfVectorizer(max_features=5000)
X = tfidf.fit_transform(df_new['news'])
y = df_new['label']
```

In [74]:
```python
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [75]:
```python
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report

nb_model = MultinomialNB()
nb_model.fit(X_train, y_train)
y_pred_nb = nb_model.predict(X_test)

print("Naive Bayes Results:")
print(classification_report(y_test, y_pred_nb))
```

```
Naive Bayes Results:
              precision    recall  f1-score   support

           0       0.94      0.95      0.94      4751
           1       0.94      0.93      0.94      4187

    accuracy                           0.94      8938
   macro avg       0.94      0.94      0.94      8938
weighted avg       0.94      0.94      0.94      8938
```

In [76]:
```python
from sklearn.svm import LinearSVC

svm_model = LinearSVC()
svm_model.fit(X_train, y_train)
y_pred_svm = svm_model.predict(X_test)

print("SVM Results:")
print(classification_report(y_test, y_pred_svm))
```

```
SVM Results:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      4751
           1       1.00      0.99      1.00      4187

    accuracy                           1.00      8938
   macro avg       1.00      1.00      1.00      8938
weighted avg       1.00      1.00      1.00      8938
```

In [77]:
```python
from sklearn.ensemble import RandomForestClassifier

rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)

print("Random Forest Results:")
print(classification_report(y_test, y_pred_rf))
```

```
Random Forest Results:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      4751
           1       1.00      1.00      1.00      4187

    accuracy                           1.00      8938
   macro avg       1.00      1.00      1.00      8938
weighted avg       1.00      1.00      1.00      8938
```

In [78]:
```python
from sklearn.metrics import accuracy_score

print("Accuracy Scores:")
print(f"Naive Bayes: {accuracy_score(y_test, y_pred_nb):.4f}")
print(f"SVM        : {accuracy_score(y_test, y_pred_svm):.4f}")
print(f"RandomForest: {accuracy_score(y_test, y_pred_rf):.4f}")
```

```
Accuracy Scores:
Naive Bayes: 0.9410
SVM        : 0.9955
RandomForest: 0.9971
```

📄 Final Conclusion – Fake News Detection Project

In this project, I developed a machine learning pipeline to classify news articles as real or fake, addressing the critical issue of misinformation. The dataset consisted of over 44,000 news entries, which underwent extensive preprocessing including lowercasing, punctuation removal, stemming, and vectorization using TF-IDF.

Three models were trained and evaluated:

- Naive Bayes: 94.10% accuracy
- Support Vector Machine (SVM)**: 99.55% accuracy
- Random Forest: 99.71% accuracy

Among these, Random Forest achieved the highest accuracy, making it the final model of choice. The model was integrated into a user-friendly Streamlit web app, allowing users to input any news text and receive an instant classification result.

This project demonstrates strong practical knowledge of:

- Natural Language Processing (NLP)
- Text vectorization techniques
- Supervised machine learning
- Model evaluation and comparison
- Python libraries like NLTK, Scikit-learn, Pandas
- GUI development using Streamlit

Overall, this solution can be a powerful tool to assist platforms, journalists, and readers in identifying and combating the spread of fake news.

In [ ]: