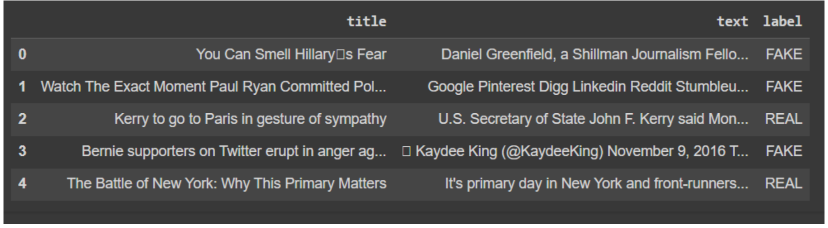
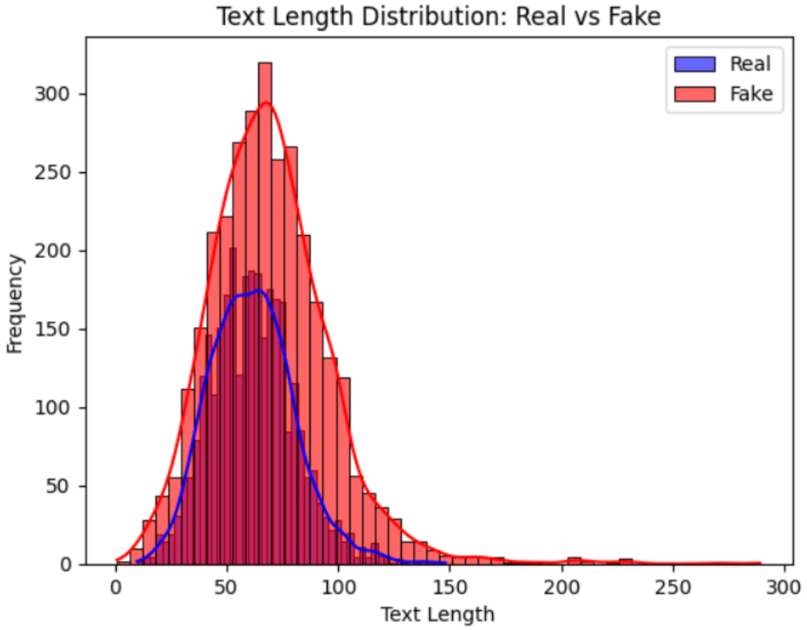
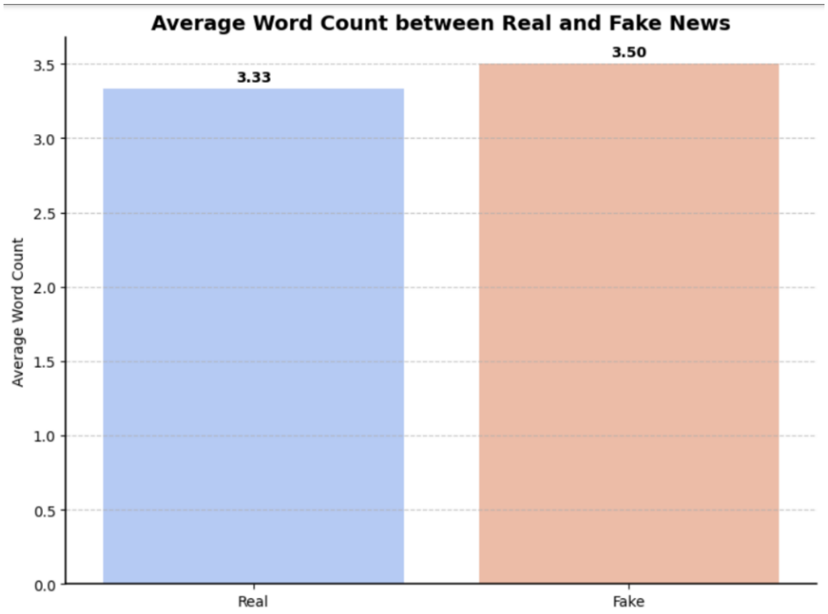
# Data Collection and Preprocessing Phase

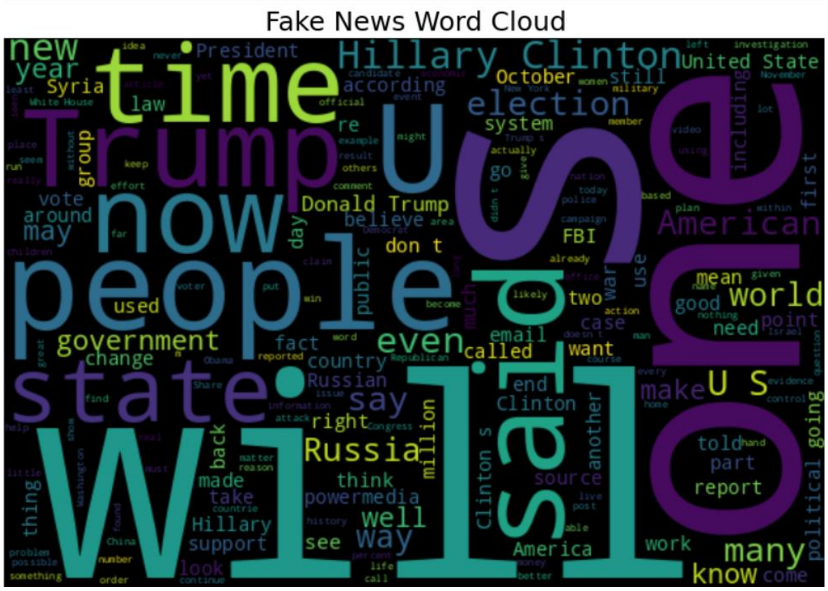
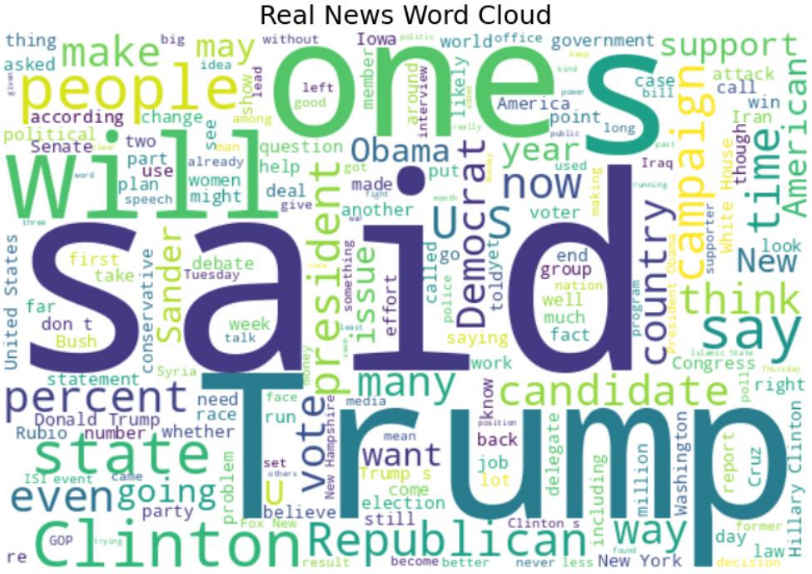
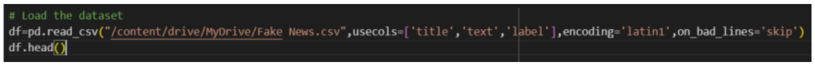
| | |
|---|---|
| Date | 23 September 2024 |
| Team ID | LTVIP2024TMID25030 |
| Project Title | FAKE NEWS ANALYSIS IN SOCIAL MEDIA |
| Maximum Marks | 6 Marks |

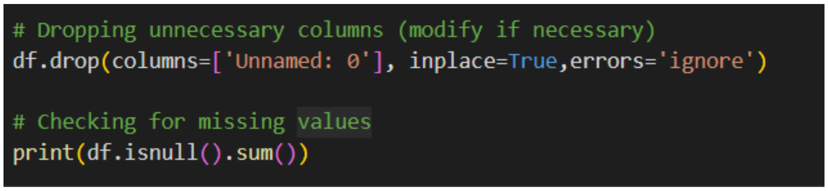
### Data Exploration and Preprocessing Template

Dataset variables will be statistically analysed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modelling, and forming a strong foundation for insights and predictions.

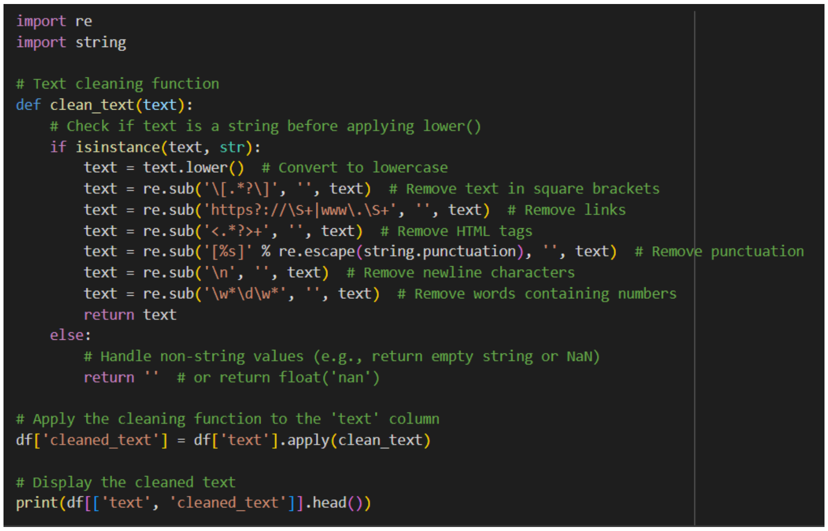
<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>
<u>Dimension:</u><br>
9285 Rows X 4 Columns<br>
<u>Descriptive statistics:</u><br>

<table>
<tr><th></th><th>title</th><th>text</th><th>label</th></tr>
<tr><td>0</td><td>You Can Smell Hillary□s Fear</td><td>Daniel Greenfield, a Shillman Journalism Fello...</td><td>FAKE</td></tr>
<tr><td>1</td><td>Watch The Exact Moment Paul Ryan Committed Pol...</td><td>Google Pinterest Digg Linkedin Reddit Stumbleu...</td><td>FAKE</td></tr>
<tr><td>2</td><td>Kerry to go to Paris in gesture of sympathy</td><td>U.S. Secretary of State John F. Kerry said Mon...</td><td>REAL</td></tr>
<tr><td>3</td><td>Bernie supporters on Twitter erupt in anger ag...</td><td>□ Kaydee King (@KaydeeKing) November 9, 2016 T...</td><td>FAKE</td></tr>
<tr><td>4</td><td>The Battle of New York: Why This Primary Matters</td><td>It's primary day in New York and front-runners...</td><td>REAL</td></tr>
</table>
</td></tr>
</table>

| | |
|---|---|
| Univariate Analysis |  |
| Bivariate Analysis |  |

<table>
<tr><td>Multivariate Analysis</td><td>Fake News Word Cloud<br><br> <br>Real News Word Cloud<br><br> </td></tr>
<tr><td>Outliers and Anomalies</td><td>-</td></tr>
<tr><td colspan="2"><b>Data Preprocessing Code Screenshots</b></td></tr>
<tr><td>Loading Data</td><td><br># Load the dataset<br>df=pd.read_csv("/content/drive/MyDrive/Fake News.csv",usecols=['title','text','label'],encoding='latin1',on_bad_lines='skip')<br>df.head()</td></tr>
</table>

<table>
<tr><td></td><td></td></tr>
<tr><td>Handling Missing Data</td><td>


```
# Dropping unnecessary columns (modify if necessary)
df.drop(columns=['Unnamed: 0'], inplace=True,errors='ignore')

# Checking for missing values
print(df.isnull().sum())
```
</td></tr>
<tr><td>Data Transformation</td><td>


```
import re
import string

# Text cleaning function
def clean_text(text):
    # Check if text is a string before applying lower()
    if isinstance(text, str):
        text = text.lower()  # Convert to lowercase
        text = re.sub('\[.*?\]', '', text)  # Remove text in square brackets
        text = re.sub('https?://\S+|www\.\S+', '', text)  # Remove links
        text = re.sub('<.*?>+', '', text)  # Remove HTML tags
        text = re.sub('[%s]' % re.escape(string.punctuation), '', text)  # Remove punctuation
        text = re.sub('\n', '', text)  # Remove newline characters
        text = re.sub('\w*\d\w*', '', text)  # Remove words containing numbers
        return text
    else:
        # Handle non-string values (e.g., return empty string or NaN)
        return ''  # or return float('nan')

# Apply the cleaning function to the 'text' column
df['cleaned_text'] = df['text'].apply(clean_text)

# Display the cleaned text
print(df[['text', 'cleaned_text']].head())
```
</td></tr>
<tr><td>Feature Engineering</td><td>Attached the code in the final submission</td></tr>
<tr><td>Save Processed Data</td><td>-</td></tr>
</table>