In [1]:
```
pip install datasets
```

```
Collecting datasets
  Downloading datasets-3.0.2-py3-none-any.whl.metadata (20 kB)
Requirement already satisfied: filelock in c:\users\jon29\anaconda3\lib\site-pack
ages (from datasets) (3.13.1)
Requirement already satisfied: numpy>=1.17 in c:\users\jon29\anaconda3\lib\site-p
ackages (from datasets) (1.26.4)
Collecting pyarrow>=15.0.0 (from datasets)
  Downloading pyarrow-17.0.0-cp311-cp311-win_amd64.whl.metadata (3.4 kB)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in c:\users\jon29\anaconda3\lib
\site-packages (from datasets) (0.3.7)
Requirement already satisfied: pandas in c:\users\jon29\anaconda3\lib\site-packag
es (from datasets) (2.1.4)
Collecting requests>=2.32.2 (from datasets)
  Downloading requests-2.32.3-py3-none-any.whl.metadata (4.6 kB)
Collecting tqdm>=4.66.3 (from datasets)
  Downloading tqdm-4.66.5-py3-none-any.whl.metadata (57 kB)
     ---------------------------------------- 0.0/57.6 kB ? eta -:--:--
     ---------------------- ----------------- 30.7/57.6 kB 1.3 MB/s eta 0:00:01
     ---------------------------------------- 57.6/57.6 kB 751.8 kB/s eta 0:00:00
Collecting xxhash (from datasets)
  Downloading xxhash-3.5.0-cp311-cp311-win_amd64.whl.metadata (13 kB)
Collecting multiprocess<0.70.17 (from datasets)
  Downloading multiprocess-0.70.16-py311-none-any.whl.metadata (7.2 kB)
Requirement already satisfied: fsspec<=2024.9.0,>=2023.1.0 in c:\users\jon29\anac
onda3\lib\site-packages (from fsspec[http]<=2024.9.0,>=2023.1.0->datasets) (2023.
10.0)
Requirement already satisfied: aiohttp in c:\users\jon29\anaconda3\lib\site-packa
ges (from datasets) (3.9.3)
Collecting huggingface-hub>=0.23.0 (from datasets)
  Downloading huggingface_hub-0.26.1-py3-none-any.whl.metadata (13 kB)
Requirement already satisfied: packaging in c:\users\jon29\anaconda3\lib\site-pac
kages (from datasets) (23.1)
Requirement already satisfied: pyyaml>=5.1 in c:\users\jon29\anaconda3\lib\site-p
ackages (from datasets) (6.0.1)
Requirement already satisfied: aiosignal>=1.1.2 in c:\users\jon29\anaconda3\lib\s
ite-packages (from aiohttp->datasets) (1.2.0)
Requirement already satisfied: attrs>=17.3.0 in c:\users\jon29\anaconda3\lib\site
-packages (from aiohttp->datasets) (23.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in c:\users\jon29\anaconda3\lib
\site-packages (from aiohttp->datasets) (1.4.0)
Requirement already satisfied: multidict<7.0,>=4.5 in c:\users\jon29\anaconda3\li
b\site-packages (from aiohttp->datasets) (6.0.4)
Requirement already satisfied: yarl<2.0,>=1.0 in c:\users\jon29\anaconda3\lib\sit
e-packages (from aiohttp->datasets) (1.9.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\jon29\anaco
nda3\lib\site-packages (from huggingface-hub>=0.23.0->datasets) (4.9.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\jon29\anacond
a3\lib\site-packages (from requests>=2.32.2->datasets) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\users\jon29\anaconda3\lib\site-
packages (from requests>=2.32.2->datasets) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\jon29\anaconda3\lib
\site-packages (from requests>=2.32.2->datasets) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\jon29\anaconda3\lib
\site-packages (from requests>=2.32.2->datasets) (2024.2.2)
Requirement already satisfied: colorama in c:\users\jon29\anaconda3\lib\site-pack
ages (from tqdm>=4.66.3->datasets) (0.4.6)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\jon29\anaconda3
\lib\site-packages (from pandas->datasets) (2.8.2)
```

```
Requirement already satisfied: pytz>=2020.1 in c:\users\jon29\anaconda3\lib\site-
packages (from pandas->datasets) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\jon29\anaconda3\lib\sit
e-packages (from pandas->datasets) (2023.3)
Requirement already satisfied: six>=1.5 in c:\users\jon29\anaconda3\lib\site-pack
ages (from python-dateutil>=2.8.2->pandas->datasets) (1.16.0)
Downloading datasets-3.0.2-py3-none-any.whl (472 kB)
   ---------------------------------------- 0.0/472.7 kB ? eta -:--:--
   ------- -------------------------------- 92.2/472.7 kB 5.1 MB/s eta 0:00:01
   ------------------ --------------------- 225.3/472.7 kB 2.3 MB/s eta 0:00:01
   ----------------------- ---------------- 286.7/472.7 kB 2.0 MB/s eta 0:00:01
   ---------------------------------------- 472.7/472.7 kB 3.0 MB/s eta 0:00:00
Downloading huggingface_hub-0.26.1-py3-none-any.whl (447 kB)
   ---------------------------------------- 0.0/447.4 kB ? eta -:--:--
   ---------------------------------------- 447.4/447.4 kB 9.3 MB/s eta 0:00:00
Downloading multiprocess-0.70.16-py311-none-any.whl (143 kB)
   ---------------------------------------- 0.0/143.5 kB ? eta -:--:--
   ---------------------------------------- 143.5/143.5 kB 8.9 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
   ---------------------------------------- 0.0/116.3 kB ? eta -:--:--
   ---------------------------------------- 116.3/116.3 kB ? eta 0:00:00
Downloading pyarrow-17.0.0-cp311-cp311-win_amd64.whl (25.2 MB)
   ---------------------------------------- 0.0/25.2 MB ? eta -:--:--
    --------------------------------------- 0.4/25.2 MB 23.2 MB/s eta 0:00:02
    --------------------------------------- 0.4/25.2 MB 23.2 MB/s eta 0:00:02
   -- ------------------------------------- 1.4/25.2 MB 11.1 MB/s eta 0:00:03
   -- ------------------------------------- 1.5/25.2 MB 8.2 MB/s eta 0:00:03
   --- ------------------------------------ 2.1/25.2 MB 9.0 MB/s eta 0:00:03
   ---- ----------------------------------- 2.6/25.2 MB 9.9 MB/s eta 0:00:03
   ---- ----------------------------------- 3.1/25.2 MB 9.5 MB/s eta 0:00:03
   ----- ---------------------------------- 3.6/25.2 MB 9.7 MB/s eta 0:00:03
   ------ --------------------------------- 4.2/25.2 MB 9.9 MB/s eta 0:00:03
   ------- -------------------------------- 4.7/25.2 MB 9.9 MB/s eta 0:00:03
   -------- ------------------------------- 5.2/25.2 MB 10.1 MB/s eta 0:00:02
   -------- ------------------------------- 5.5/25.2 MB 9.8 MB/s eta 0:00:03
   --------- ------------------------------ 6.0/25.2 MB 9.9 MB/s eta 0:00:02
   ---------- ----------------------------- 6.6/25.2 MB 10.1 MB/s eta 0:00:02
   ----------- ---------------------------- 7.1/25.2 MB 10.1 MB/s eta 0:00:02
   ----------- ---------------------------- 7.2/25.2 MB 10.2 MB/s eta 0:00:02
   ----------- ---------------------------- 7.4/25.2 MB 9.3 MB/s eta 0:00:02
   ------------ --------------------------- 8.0/25.2 MB 9.5 MB/s eta 0:00:02
   ------------- -------------------------- 8.5/25.2 MB 9.7 MB/s eta 0:00:02
   ------------- -------------------------- 8.5/25.2 MB 9.7 MB/s eta 0:00:02
   -------------- ------------------------- 9.3/25.2 MB 9.4 MB/s eta 0:00:02
   ---------------- ----------------------- 10.1/25.2 MB 9.8 MB/s eta 0:00:02
   ---------------- ----------------------- 10.6/25.2 MB 10.6 MB/s eta 0:00:02
   ----------------- ---------------------- 11.0/25.2 MB 10.2 MB/s eta 0:00:02
   ----------------- ---------------------- 11.5/25.2 MB 9.8 MB/s eta 0:00:02
   ------------------ --------------------- 12.0/25.2 MB 10.1 MB/s eta 0:00:02
   ------------------ --------------------- 12.4/25.2 MB 9.9 MB/s eta 0:00:02
   ------------------- -------------------- 12.9/25.2 MB 9.9 MB/s eta 0:00:02
   -------------------- ------------------- 13.4/25.2 MB 9.9 MB/s eta 0:00:02
   -------------------- ------------------- 13.9/25.2 MB 9.9 MB/s eta 0:00:02
   --------------------- ------------------ 14.4/25.2 MB 9.8 MB/s eta 0:00:02
   ---------------------- ----------------- 14.9/25.2 MB 9.9 MB/s eta 0:00:02
   ---------------------- ----------------- 15.3/25.2 MB 9.8 MB/s eta 0:00:02
   ----------------------- ---------------- 15.8/25.2 MB 9.9 MB/s eta 0:00:01
   ----------------------- ---------------- 16.2/25.2 MB 9.8 MB/s eta 0:00:01
   ------------------------ --------------- 16.7/25.2 MB 9.8 MB/s eta 0:00:01
   ------------------------ --------------- 17.1/25.2 MB 9.6 MB/s eta 0:00:01
```

```
   --------------------------- ------------ 17.6/25.2 MB 10.4 MB/s eta 0:00:01
   ---------------------------- ----------- 18.1/25.2 MB 10.4 MB/s eta 0:00:01
   ----------------------------- ---------- 18.4/25.2 MB 10.4 MB/s eta 0:00:01
   ------------------------------ --------- 18.8/25.2 MB 10.9 MB/s eta 0:00:01
   ------------------------------- -------- 19.3/25.2 MB 10.6 MB/s eta 0:00:01
   -------------------------------- ------- 19.8/25.2 MB 10.1 MB/s eta 0:00:01
   --------------------------------- ------ 20.2/25.2 MB 9.9 MB/s eta 0:00:01
   --------------------------------- ------ 20.7/25.2 MB 9.9 MB/s eta 0:00:01
   ---------------------------------- ----- 21.3/25.2 MB 10.1 MB/s eta 0:00:01
   ---------------------------------- ----- 21.7/25.2 MB 9.9 MB/s eta 0:00:01
   ----------------------------------- ---- 22.2/25.2 MB 10.1 MB/s eta 0:00:01
   ----------------------------------- ---- 22.7/25.2 MB 10.1 MB/s eta 0:00:01
   ------------------------------------ --- 23.2/25.2 MB 10.1 MB/s eta 0:00:01
   ------------------------------------- -- 23.6/25.2 MB 10.1 MB/s eta 0:00:01
   ------------------------------------- -- 24.1/25.2 MB 10.1 MB/s eta 0:00:01
   -------------------------------------- - 24.6/25.2 MB 10.1 MB/s eta 0:00:01
   -------------------------------------- - 25.1/25.2 MB 10.1 MB/s eta 0:00:01
   ---------------------------------------- 25.2/25.2 MB 9.8 MB/s eta 0:00:00
Downloading requests-2.32.3-py3-none-any.whl (64 kB)
   ---------------------------------------- 0.0/64.9 kB ? eta -:--:--
   ---------------------------------------- 64.9/64.9 kB 3.4 MB/s eta 0:00:00
Downloading tqdm-4.66.5-py3-none-any.whl (78 kB)
   ---------------------------------------- 0.0/78.4 kB ? eta -:--:--
   ---------------------------------------- 78.4/78.4 kB 4.3 MB/s eta 0:00:00
Downloading xxhash-3.5.0-cp311-cp311-win_amd64.whl (30 kB)
Installing collected packages: xxhash, tqdm, requests, pyarrow, dill, multiproces
s, huggingface-hub, datasets
  Attempting uninstall: tqdm
    Found existing installation: tqdm 4.65.0
    Uninstalling tqdm-4.65.0:
      Successfully uninstalled tqdm-4.65.0
  Attempting uninstall: requests
    Found existing installation: requests 2.31.0
    Uninstalling requests-2.31.0:
      Successfully uninstalled requests-2.31.0
  Attempting uninstall: pyarrow
    Found existing installation: pyarrow 14.0.2
    Uninstalling pyarrow-14.0.2:
      Successfully uninstalled pyarrow-14.0.2
  Attempting uninstall: dill
    Found existing installation: dill 0.3.7
    Uninstalling dill-0.3.7:
      Successfully uninstalled dill-0.3.7
Successfully installed datasets-3.0.2 dill-0.3.8 huggingface-hub-0.26.1 multiproc
ess-0.70.16 pyarrow-17.0.0 requests-2.32.3 tqdm-4.66.5 xxhash-3.5.0
Note: you may need to restart the kernel to use updated packages.
ERROR: pip's dependency resolver does not currently take into account all the pac
kages that are installed. This behaviour is the source of the following dependenc
y conflicts.
conda-repo-cli 1.0.75 requires requests_mock, which is not installed.
conda-repo-cli 1.0.75 requires clyent==1.2.1, but you have clyent 1.2.2 which is
incompatible.
conda-repo-cli 1.0.75 requires requests==2.31.0, but you have requests 2.32.3 whi
ch is incompatible.
```

In [2]:
```
from datasets import load_dataset
dataset = load_dataset("Senem/Nostalgic_Sentiment_Analysis_of_YouTube_Comments_D
print(dataset['train'])
```

```
README.md:   0%|          | 0.00/978 [00:00<?, ?B/s]
```

```
C:\Users\jon29\anaconda3\Lib\site-packages\huggingface_hub\file_download.py:139:
UserWarning: `huggingface_hub` cache-system uses symlinks by default to efficient
ly store duplicated files but your machine does not support them in C:\Users\jon2
9\.cache\huggingface\hub\datasets--Senem--Nostalgic_Sentiment_Analysis_of_YouTube
_Comments_Data. Caching files will still work but in a degraded version that migh
t require more space on your disk. This warning can be disabled by setting the `H
F_HUB_DISABLE_SYMLINKS_WARNING` environment variable. For more details, see http
s://huggingface.co/docs/huggingface_hub/how-to-cache#limitations.
To support symlinks on Windows, you either need to activate Developer Mode or to
run Python as an administrator. In order to activate developer mode, see this art
icle: https://docs.microsoft.com/en-us/windows/apps/get-started/enable-your-devic
e-for-development
  warnings.warn(message)
(…)nt_Analysis_of_YouTube_Comments_Data.csv:   0%|          | 0.00/219k [00:00<?
?B/s]
Generating train split:   0%|          | 0/1500 [00:00<?, ? examples/s]
Dataset({
    features: ['sentiment', 'comment'],
    num_rows: 1500
})
```

In [3]:
```python
import pandas as pd

# 將數據集轉換為 DataFrame 格式
df = pd.DataFrame(dataset['train'])

# 查看數據集的前幾行
print(df.head())
```

```
       sentiment                                            comment
0  not nostalgia  He was a singer with a golden voice that I lov...
1      nostalgia  The mist beautiful voice ever I listened to hi...
2      nostalgia  I have most of Mr. Reeves songs.  Always love ...
3  not nostalgia  30 day leave from 1st tour in Viet Nam to conv...
4      nostalgia  listening to his songs reminds me of my mum wh...
```

In [4]:
```python
print("數據集大小:", df.shape)
print("數據集特徵:", df.columns.tolist())
print("數據集示例:")
print(df.head())
```

```
數據集大小: (1500, 2)
數據集特徵: ['sentiment', 'comment']
數據集示例:
       sentiment                                            comment
0  not nostalgia  He was a singer with a golden voice that I lov...
1      nostalgia  The mist beautiful voice ever I listened to hi...
2      nostalgia  I have most of Mr. Reeves songs.  Always love ...
3  not nostalgia  30 day leave from 1st tour in Viet Nam to conv...
4      nostalgia  listening to his songs reminds me of my mum wh...
```

In [ ]:
```python
#資料準備
```

In [5]:
```python
# 確認數據類型
print(df.dtypes)

# 清理數據：將評論轉換為小寫，去除多餘的空格
df['comment'] = df['comment'].str.lower().str.strip()
```

```
# 確認清理結果
print(df.head())
```

```
sentiment    object
comment      object
dtype: object
       sentiment                                            comment
0  not nostalgia  he was a singer with a golden voice that i lov...
1      nostalgia  the mist beautiful voice ever i listened to hi...
2      nostalgia  i have most of mr. reeves songs.  always love ...
3  not nostalgia  30 day leave from 1st tour in viet nam to conv...
4      nostalgia  listening to his songs reminds me of my mum wh...
```

In [6]:
```
# 查看數據集的基本信息
print(df.info())
print(df.describe())
print("前五條記錄:")
print(df.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   sentiment  1500 non-null   object
 1   comment    1500 non-null   object
dtypes: object(2)
memory usage: 23.6+ KB
None
           sentiment                                            comment
count           1500                                               1500
unique             2                                               1499
top     not nostalgia  never heard this song before... wow what an am...
freq             750                                                  2
前五條記錄:
       sentiment                                            comment
0  not nostalgia  he was a singer with a golden voice that i lov...
1      nostalgia  the mist beautiful voice ever i listened to hi...
2      nostalgia  i have most of mr. reeves songs.  always love ...
3  not nostalgia  30 day leave from 1st tour in viet nam to conv...
4      nostalgia  listening to his songs reminds me of my mum wh...
```

In [ ]:
```
#Data Mining using Pandas
```

In [7]:
```
# 檢查缺失值
print("缺失值檢查:")
print(df.isnull().sum())

# 如果有缺失值，刪除缺失行
df = df.dropna()
```

```
缺失值檢查:
sentiment    0
comment      0
dtype: int64
```

In [8]:
```
# 檢查重複行
print("重複行數量:", df.duplicated().sum())

# 刪除重複行
df = df.drop_duplicates()
```

重複行數量: 1

In [ ]:
```
#Data Preprocessing
```

In [9]:
```
sample_df = df.sample(500, random_state=42)
```

In [10]:
```
# 計算字符數和單詞數
df['char_count'] = df['comment'].apply(len)
df['word_count'] = df['comment'].apply(lambda x: len(x.split()))
```

In [11]:
```
# 只保留相關的特徵
selected_features = ['sentiment', 'comment', 'char_count', 'word_count']
df = df[selected_features]
```

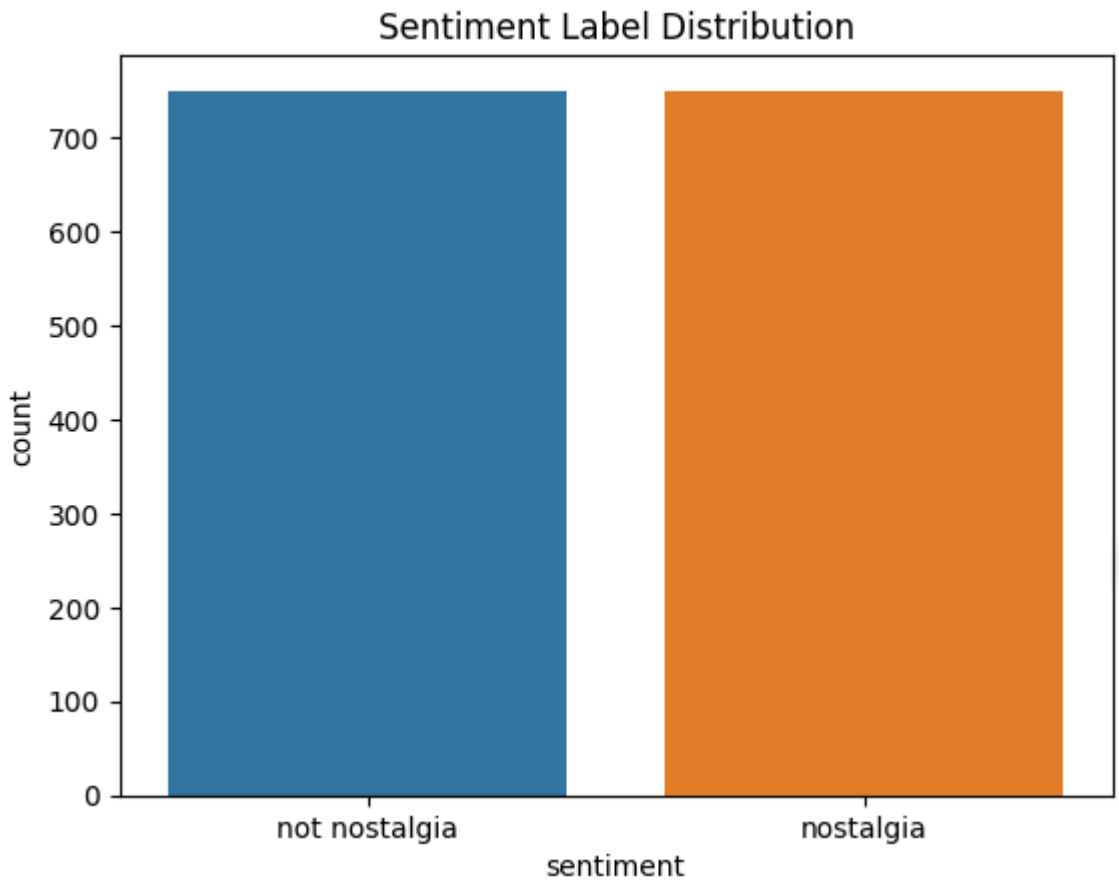
In [12]:
```
from sklearn.preprocessing import MinMaxScaler

# 對字符數和單詞數進行標準化
scaler = MinMaxScaler()
df[['char_count', 'word_count']] = scaler.fit_transform(df[['char_count', 'word_
```

In [13]:
```
from sklearn.decomposition import PCA

# 使用 PCA 降維至 2 維
pca = PCA(n_components=2)
df_pca = pca.fit_transform(df[['char_count', 'word_count']])
```

In [14]:
```
from sklearn.preprocessing import LabelBinarizer

# 將情感標籤進行二值化
lb = LabelBinarizer()
df['sentiment_binary'] = lb.fit_transform(df['sentiment'])
```

In [ ]:
```
#Data Exploration
```

In [15]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# 情感標籤分佈
sns.countplot(x='sentiment', data=df)
plt.title("Sentiment Label Distribution")
plt.show()
```



In [ ]:
```
#Data Classification
```

In [16]:
```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report

# 分割數據集
X_train, X_test, y_train, y_test = train_test_split(df['comment'], df['sentiment

# TF-IDF 特徵
tfidf = TfidfVectorizer()
X_train_tfidf = tfidf.fit_transform(X_train)
X_test_tfidf = tfidf.transform(X_test)

# 訓練 Naive Bayes 模型
nb = MultinomialNB()
nb.fit(X_train_tfidf, y_train)

# 模型評估
y_pred = nb.predict(X_test_tfidf)
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.80      0.92      0.86       158
           1       0.90      0.74      0.81       142

    accuracy                           0.84       300
   macro avg       0.85      0.83      0.83       300
weighted avg       0.84      0.84      0.83       300
```

In [18]:
```
!pip install wordcloud
```

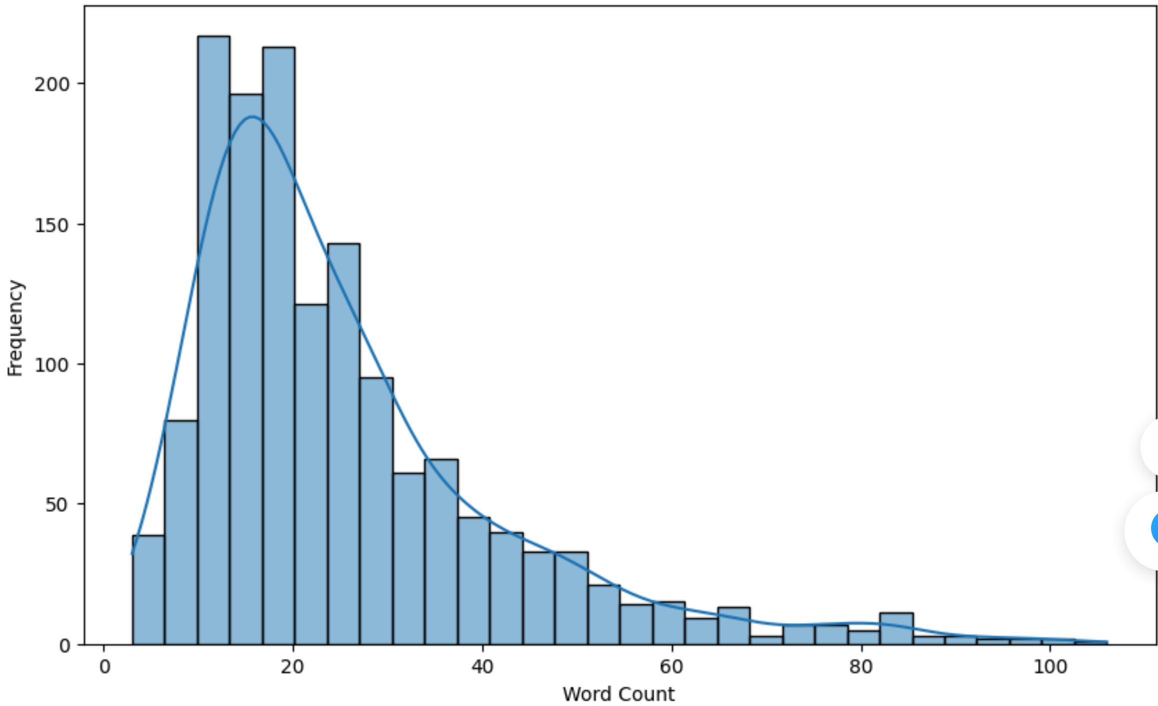
```
Collecting wordcloud
  Downloading wordcloud-1.9.3-cp311-cp311-win_amd64.whl.metadata (3.5 kB)
Requirement already satisfied: numpy>=1.6.1 in c:\users\jon29\anaconda3\lib\site-
packages (from wordcloud) (1.26.4)
Requirement already satisfied: pillow in c:\users\jon29\anaconda3\lib\site-packag
es (from wordcloud) (10.2.0)
Requirement already satisfied: matplotlib in c:\users\jon29\anaconda3\lib\site-pa
ckages (from wordcloud) (3.7.3)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\jon29\anaconda3\lib\s
ite-packages (from matplotlib->wordcloud) (1.2.0)
Requirement already satisfied: cycler>=0.10 in c:\users\jon29\anaconda3\lib\site-
packages (from matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\jon29\anaconda3\lib
\site-packages (from matplotlib->wordcloud) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\jon29\anaconda3\li
\site-packages (from matplotlib->wordcloud) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\jon29\anaconda3\lib\si
te-packages (from matplotlib->wordcloud) (23.1)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\jon29\anaconda3\lib\s
ite-packages (from matplotlib->wordcloud) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\jon29\anaconda3\l
ib\site-packages (from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\jon29\anaconda3\lib\site-pack
ages (from python-dateutil>=2.7->matplotlib->wordcloud) (1.16.0)
Downloading wordcloud-1.9.3-cp311-cp311-win_amd64.whl (300 kB)
   ---------------------------------------- 0.0/300.2 kB ? eta -:--:--
   - -------------------------------------- 10.2/300.2 kB ? eta -:--:--
   ----- ---------------------------------- 41.0/300.2 kB 495.5 kB/s eta 0:00:01
   ---------------- ----------------------- 133.1/300.2 kB 1.1 MB/s eta 0:00:01
   ---------------------------------- ----- 276.5/300.2 kB 1.7 MB/s eta 0:00:01
   ---------------------------------------- 300.2/300.2 kB 1.6 MB/s eta 0:00:00
Installing collected packages: wordcloud
Successfully installed wordcloud-1.9.3
```

In [19]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB, ComplementNB
from sklearn.metrics import classification_report
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

# 讀取數據
df = pd.DataFrame(dataset['train'])

# 計算單詞數量並可視化
df['word_count'] = df['comment'].apply(lambda x: len(x.split()))
plt.figure(figsize=(10, 6))
sns.histplot(df['word_count'], bins=30, kde=True)
plt.title('Word Count Distribution')
plt.xlabel('Word Count')
plt.ylabel('Frequency')
plt.show()

# 根據情感標籤生成詞雲
for sentiment in df['sentiment'].unique():
    text = ' '.join(df[df['sentiment'] == sentiment]['comment'])
```

```
    wordcloud = WordCloud(width=800, height=400, background_color='white').gener
    plt.figure(figsize=(10, 6))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.title(f'Word Cloud for {sentiment}')
    plt.axis('off')
    plt.show()

# 拆分數據
X_train, X_test, y_train, y_test = train_test_split(df['comment'], df['sentiment

# 建立 CountVectorizer 和 TfidfVectorizer 特徵
count_vectorizer = CountVectorizer()
tfidf_vectorizer = TfidfVectorizer()

X_train_count = count_vectorizer.fit_transform(X_train)
X_test_count = count_vectorizer.transform(X_test)

X_train_tfidf = tfidf_vectorizer.fit_transform(X_train)
X_test_tfidf = tfidf_vectorizer.transform(X_test)

# MultinomialNB （適合詞頻特徵）
nb_count = MultinomialNB()
nb_count.fit(X_train_count, y_train)
y_pred_count = nb_count.predict(X_test_count)
print("MultinomialNB with CountVectorizer:")
print(classification_report(y_test, y_pred_count))

# ComplementNB （適合 TF-IDF 特徵）
nb_tfidf = ComplementNB()
nb_tfidf.fit(X_train_tfidf, y_train)
y_pred_tfidf = nb_tfidf.predict(X_test_tfidf)
print("ComplementNB with TfidfVectorizer:")
print(classification_report(y_test, y_pred_tfidf))
```

```
C:\Users\jon29\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarnin
g: use_inf_as_na option is deprecated and will be removed in a future version. Co
nvert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

Word Count Distribution



Word Cloud for not nostalgia

Word Cloud for nostalgia

```
MultinomialNB with CountVectorizer:
                precision    recall  f1-score   support

    nostalgia       0.78      0.95      0.86       152
not nostalgia       0.94      0.72      0.81       148

     accuracy                           0.84       300
    macro avg       0.86      0.84      0.83       300
 weighted avg       0.86      0.84      0.83       300

ComplementNB with TfidfVectorizer:
                precision    recall  f1-score   support

    nostalgia       0.77      0.96      0.85       152
not nostalgia       0.95      0.70      0.81       148

     accuracy                           0.83       300
    macro avg       0.86      0.83      0.83       300
 weighted avg       0.86      0.83      0.83       300
```

In [24]:
```
print("CountVectorizer 與 TF-IDF 的影響：TF-IDF 通過降低高頻詞的權重，強化了某些不常
print("模型選擇：如果目標是更高的整體準確性並且重視 nostalgia 類別的完整識別，Compleme
```

CountVectorizer 與 TF-IDF 的影響：TF-IDF 通過降低高頻詞的權重，強化了某些不常見但重要的詞語，對 nostalgia 類別尤其有效，使得模型在該類別的召回率有所提升。
模型選擇：如果目標是更高的整體準確性並且重視 nostalgia 類別的完整識別，ComplementNB 與 TF-IDF 特徵可能更適合；但如果更重視平衡表現且略微偏向 not nostalgia 類別，MultinomialNB 與 CountVectorizer 會更合適。

In [ ]:
```

```

In [ ]:
```
# 4. Fourth: In the lab, we applied each step really quickly just to illustrate
```

In [27]:
```
text = """
數據處理優化：

1. 停用詞處理：目前，模型的特徵集可能包含了一些高頻且無意義的詞（如 "the"、"is" 等），
2. 文本清理：增加對文本的清理，比如去除特殊符號、還原縮寫（如將 "can't" 還原為 "cannot
3. 加入雙詞組特徵：考慮增加雙詞組（bigrams）特徵。對於情感分類，一些詞的組合可能比單詞更
4. 類別平衡：檢查數據中 `nostalgia` 和 `not nostalgia` 的樣本數量是否均衡。若不平衡，
5. 自定義情感詞庫：針對情感分析，可以考慮建立一些自定義詞庫，加入特定情感的詞或詞組，幫助

"""

print(text)
```

數據處理優化：

1. 停用詞處理：目前，模型的特徵集可能包含了一些高頻且無意義的詞（如 "the"、"is" 等），這些詞不利於情感分類。可以在 CountVectorizer 和 TF-IDF 中直接去掉這些停用詞，讓模型更專注於具有辨識度的詞語，減少無用維度。

2. 文本清理：增加對文本的清理，比如去除特殊符號、還原縮寫（如將 "can't" 還原為 "cannot"）。這樣可以讓詞語更一致，提高模型的表現。這部分可以在預處理時用正則表達式來實現。

3. 加入雙詞組特徵：考慮增加雙詞組（bigrams）特徵。對於情感分類，一些詞的組合可能比單詞更有區分性，比如 “good old” 在 nostalgia 中可能出現頻率更高。這樣可以幫助模型捕捉到一些關鍵的情感搭配。

4. 類別平衡：檢查數據中 `nostalgia` 和 `not nostalgia` 的樣本數量是否均衡。若不平衡，可以考慮下采樣或上采樣來平衡數據，以防止模型過度偏向樣本多的類別。

5. 自定義情感詞庫：針對情感分析，可以考慮建立一些自定義詞庫，加入特定情感的詞或詞組，讓模型更準確地抓取某些情感特徵，比如“懷舊”可能包含一些具體的年份、音樂或人物的詞語。

In [ ]:

In [ ]: