In [ ]:
```
from google.colab import drive
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import torch
from torch.nn.functional import softmax
from tqdm import tqdm
import matplotlib.pyplot as plt
```

In [ ]:
```
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [ ]:
```
file_path = '/content/drive/My Drive/DJIA Dataset/upload_DJIA_table.csv'

stock_data = pd.read_csv(file_path)
```

In [ ]:
```
stock_data.head()
```

Out[ ]:

	Date	Open	High	Low	Close	Volume	Adj Close
0	2016-07-01	17924.240234	18002.380859	17916.910156	17949.369141	82160000	17949.369141
1	2016-06-30	17712.759766	17930.609375	17711.800781	17929.990234	133030000	17929.990234
2	2016-06-29	17456.019531	17704.509766	17456.019531	17694.679688	106380000	17694.679688
3	2016-06-28	17190.509766	17409.720703	17190.509766	17409.720703	112190000	17409.720703
4	2016-06-27	17355.210938	17355.210938	17063.080078	17140.240234	138740000	17140.240234

In [ ]:
```
file_path1 = '/content/drive/My Drive/DJIA Dataset/RedditNews.csv'

reddit_data = pd.read_csv(file_path1)
```

In [ ]:
```
reddit_data.head()
```

Out[ ]:

	Date	News
0	2016-07-01	A 117-year-old woman in Mexico City finally re...
1	2016-07-01	IMF chief backs Athens as permanent Olympic host
2	2016-07-01	The president of France says if Brexit won, so...
3	2016-07-01	British Man Who Must Give Police 24 Hours' Not...
4	2016-07-01	100+ Nobel laureates urge Greenpeace to stop o...

In [ ]:
```
file_path2 = '/content/drive/My Drive/DJIA Dataset/Combined_News_DJIA.csv'

combined_data = pd.read_csv(file_path2)
```

In [ ]:
```
combined_data.head()
```

Out[ ]:

	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6
0	2008-08-08	0	b"Georgia 'downs two Russian warplanes' as cou...	b'BREAKING: Musharraf to be impeached.'	b'Russia Today: Columns of troops roll into So...	b'Russian tanks are moving towards the capital...	b"Afghan children raped with 'impunity,' U.N. ...	b'150 Russian tanks have entered South Ossetia...
1	2008-08-11	1	b'Why wont America and Nato help us? If they w...	b'Bush puts foot down on Georgian conflict'	b"Jewish Georgian minister: Thanks to Israeli ...	b'Georgian army flees in disarray as Russians ...	b"Olympic opening ceremony fireworks 'faked'"	b'What were the Mossad with fraudulent New Zea...
2	2008-08-12	0	b'Remember that adorable 9-year-old who sang a...	b"Russia 'ends Georgia operation'"	b"'If we had no sexual harassment we would hav...	b"Al-Qa'eda is losing support in Iraq because ...	b'Ceasefire in Georgia: Putin Outmaneuvers the...	b'Why Microsoft and Intel tried to kill the XO...
3	2008-08-13	0	b' U.S. refuses Israel weapons to attack Iran:...	b"When the president ordered to attack Tskhinv...	b' Israel clears troops who killed Reuters cam...	b'Britain\'s policy of being tough on drugs is...	b'Body of 14 year old found in trunk; Latest (...	b'China has moved 10 *million* quake survivors...
4	2008-08-14	1	b'All the experts admit that we should legalis...	b'War in South Osetia - 89 pictures made by a ...	b'Swedish wrestler Ara Abrahamian throws away ...	b'Russia exaggerated the death toll in South O...	b'Missile That Killed 9 Inside Pakistan May Ha...	b"Rushdie Condemns Random House's Refusal to P...

5 rows × 27 columns

In [ ]:
```
stock_data.isnull().sum()
reddit_data.isnull().sum()
combined_data.isnull().sum()
```

Out[ ]:

	0
Date	0
Label	0
Top1	0
Top2	0
Top3	0
Top4	0
Top5	0
Top6	0
Top7	0
Top8	0
Top9	0
Top10	0
Top11	0
Top12	0
Top13	0
Top14	0
Top15	0
Top16	0
Top17	0
Top18	0
Top19	0
Top20	0
Top21	0
Top22	0
Top23	1
Top24	3
Top25	3

dtype: int64

In [ ]:
```
stock_data.info()
reddit_data.info()
combined_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1989 entries, 0 to 1988
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   Date       1989 non-null   object 
 1   Open       1989 non-null   float64
 2   High       1989 non-null   float64
 3   Low        1989 non-null   float64
 4   Close      1989 non-null   float64
 5   Volume     1989 non-null   int64  
 6   Adj Close  1989 non-null   float64
dtypes: float64(5), int64(1), object(1)
memory usage: 108.9+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 73608 entries, 0 to 73607
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   Date    73608 non-null  object
 1   News    73608 non-null  object
dtypes: object(2)
memory usage: 1.1+ MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1989 entries, 0 to 1988
Data columns (total 27 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   Date    1989 non-null   object
 1   Label   1989 non-null   int64 
 2   Top1    1989 non-null   object
 3   Top2    1989 non-null   object
 4   Top3    1989 non-null   object
 5   Top4    1989 non-null   object
 6   Top5    1989 non-null   object
 7   Top6    1989 non-null   object
 8   Top7    1989 non-null   object
 9   Top8    1989 non-null   object
 10  Top9    1989 non-null   object
 11  Top10   1989 non-null   object
 12  Top11   1989 non-null   object
 13  Top12   1989 non-null   object
 14  Top13   1989 non-null   object
 15  Top14   1989 non-null   object
 16  Top15   1989 non-null   object
 17  Top16   1989 non-null   object
 18  Top17   1989 non-null   object
 19  Top18   1989 non-null   object
 20  Top19   1989 non-null   object
 21  Top20   1989 non-null   object
 22  Top21   1989 non-null   object
 23  Top22   1989 non-null   object
 24  Top23   1988 non-null   object
 25  Top24   1986 non-null   object
 26  Top25   1986 non-null   object
dtypes: int64(1), object(26)
memory usage: 419.7+ KB
```

In [ ]:
```
# filled missing values by a blank space for text pre-processing
combined_data.fillna('', inplace=True)
```

In [ ]:
```
combined_data.isnull().sum()
```

Out[ ]:

	0
Date	0
Label	0
Top1	0
Top2	0
Top3	0
Top4	0
Top5	0
Top6	0
Top7	0
Top8	0
Top9	0
Top10	0
Top11	0
Top12	0
Top13	0
Top14	0
Top15	0
Top16	0
Top17	0
Top18	0
Top19	0
Top20	0
Top21	0
Top22	0
Top23	0
Top24	0
Top25	0

dtype: int64

In [ ]:
```
# combining all the Top1 to Top25 news headlines of each day into one singl
e string as FinBERT works best on one single large text/sentece
combined_data['combined_news'] = combined_data[[f'Top{i}' for i in range(1,
26)]].apply(lambda row: ' '.join(row.values.astype(str)), axis=1)
```

In [ ]:
```
# sorted data to follow ascending order as the other dataframes do
combined_data = combined_data.sort_values(by='Date', ascending=True).reset_
index(drop=True)
```

In [ ]:
```
combined_data.head()
```

Out[ ]:

	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6
0	2008-08-08	0	b"Georgia 'downs two Russian warplanes' as cou...	b'BREAKING: Musharraf to be impeached.'	b'Russia Today: Columns of troops roll into So...	b'Russian tanks are moving towards the capital...	b"Afghan children raped with 'impunity,' U.N. ...	b'150 Russian tanks have entered South Ossetia...
1	2008-08-11	1	b'Why wont America and Nato help us? If they w...	b'Bush puts foot down on Georgian conflict'	b"Jewish Georgian minister: Thanks to Israeli ...	b'Georgian army flees in disarray as Russians ...	b"Olympic opening ceremony fireworks 'faked'"	b'What were the Mossad with fraudulent New Zea...
2	2008-08-12	0	b'Remember that adorable 9-year-old who sang a...	b"Russia 'ends Georgia operation'"	b"'If we had no sexual harassment we would hav...	b"Al-Qa'eda is losing support in Iraq because ...	b'Ceasefire in Georgia: Putin Outmaneuvers the...	b'Why Microsoft and Intel tried to kill the XO...
3	2008-08-13	0	b' U.S. refuses Israel weapons to attack Iran:...	b"When the president ordered to attack Tskhinv...	b' Israel clears troops who killed Reuters cam...	b'Britain\'s policy of being tough on drugs is...	b'Body of 14 year old found in trunk; Latest (...	b'China has moved 10 *million* quake survivors...
4	2008-08-14	1	b'All the experts admit that we should legalis...	b'War in South Osetia - 89 pictures made by a ...	b'Swedish wrestler Ara Abrahamian throws away ...	b'Russia exaggerated the death toll in South O...	b'Missile That Killed 9 Inside Pakistan May Ha...	b"Rushdie Condemns Random House's Refusal to P...

5 rows × 28 columns

In [ ]:
```
print(stock_data['Date'].dtype)
print(reddit_data['Date'].dtype)
print(combined_data['Date'].dtype)
```

```
object
object
object
```

In [ ]:
```
# changing to datetime format
stock_data['Date'] = pd.to_datetime(stock_data['Date'])
reddit_data['Date'] = pd.to_datetime(reddit_data['Date'])
combined_data['Date'] = pd.to_datetime(combined_data['Date'])
```

In [ ]:
```
print(stock_data['Date'].dtype)
print(reddit_data['Date'].dtype)
print(combined_data['Date'].dtype)
```

```
datetime64[ns]
datetime64[ns]
datetime64[ns]
```

In [ ]:
```
print(stock_data['Date'].min(), stock_data['Date'].max())
print(reddit_data['Date'].min(), reddit_data['Date'].max())
print(combined_data['Date'].min(), combined_data['Date'].max())
```

```
2008-08-08 00:00:00 2016-07-01 00:00:00
2008-06-08 00:00:00 2016-07-01 00:00:00
2008-08-08 00:00:00 2016-07-01 00:00:00
```

In [ ]:
```
# removed every sentence starting with letter 'b'
for i in range(1, 26):
    combined_data[f'Top{i}'] = combined_data[f'Top{i}'].apply(lambda x: x.r
eplace("b'", "").replace("'", "") if isinstance(x, str) else x)
```

In [ ]:
```
# removed spaces
combined_data = combined_data.applymap(lambda x: x.strip() if isinstance(x,
str) else x)
```

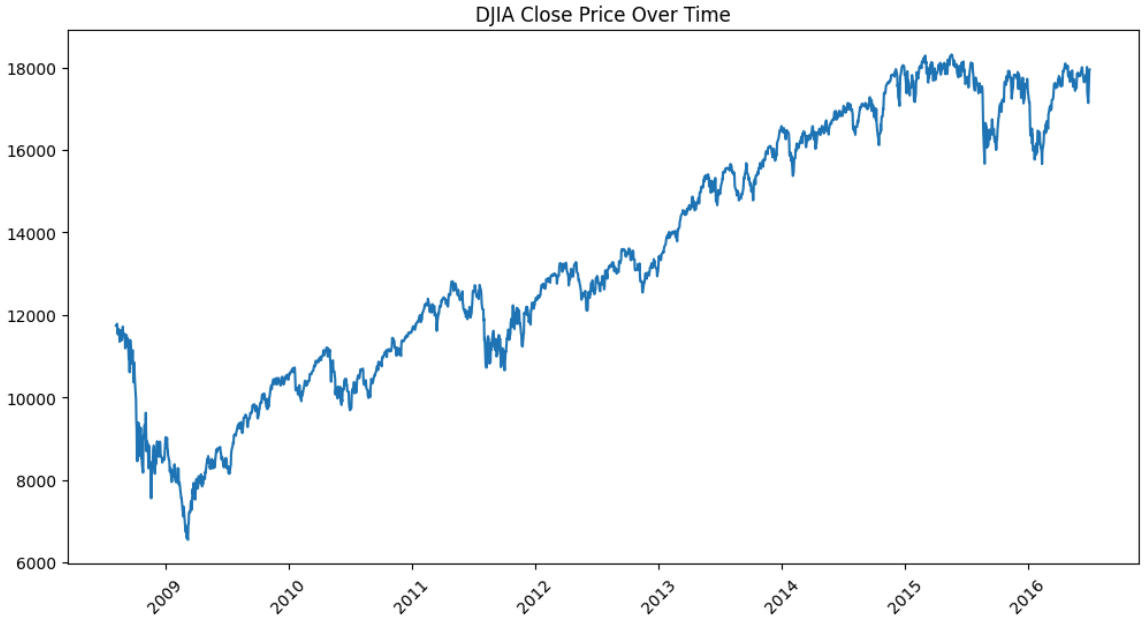
```
<ipython-input-21-72922a461d81>:2: FutureWarning: DataFrame.applymap has b
een deprecated. Use DataFrame.map instead.
  combined_data = combined_data.applymap(lambda x: x.strip() if isinstance
(x, str) else x)
```

In [ ]:
```
# lowercasing text data
for i in range(1, 26):
    combined_data[f'Top{i}'] = combined_data[f'Top{i}'].str.lower()
```

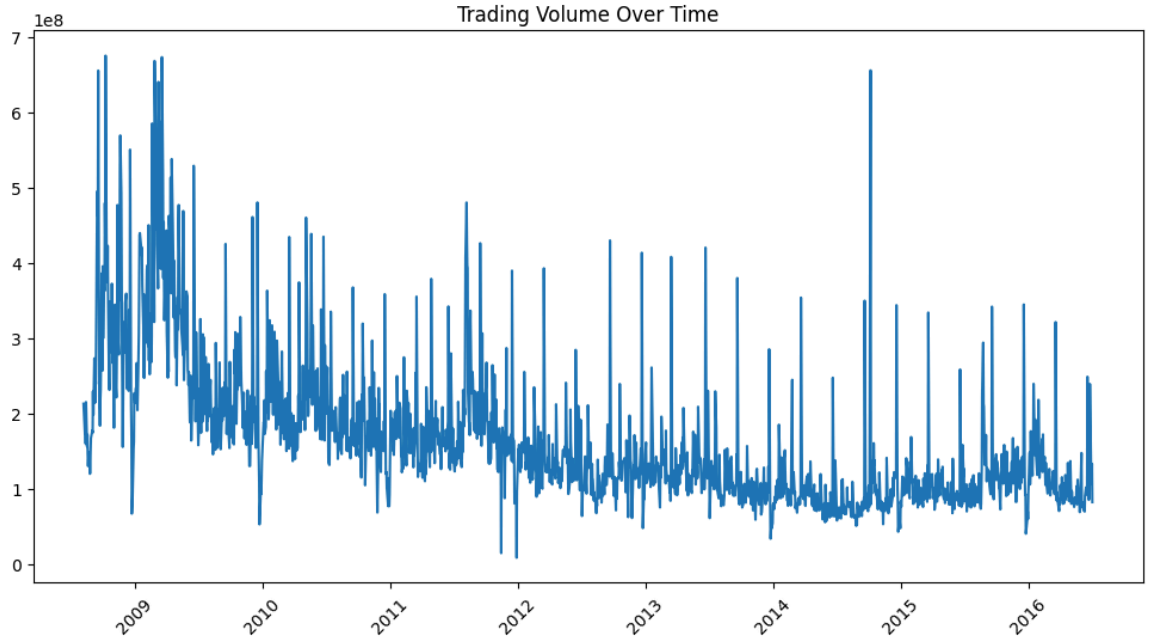
In [ ]:
```
combined_data.duplicated().sum()
reddit_data.duplicated().sum()
stock_data.duplicated().sum()
combined_data['Date'].duplicated().sum()
```

Out[ ]:
```
np.int64(0)
```

In [ ]:
```
# Closing Price over time
plt.figure(figsize=(12,6))
plt.plot(stock_data['Date'], stock_data['Close'])
plt.title('DJIA Close Price Over Time')
plt.xticks(rotation=45)
plt.show()
```

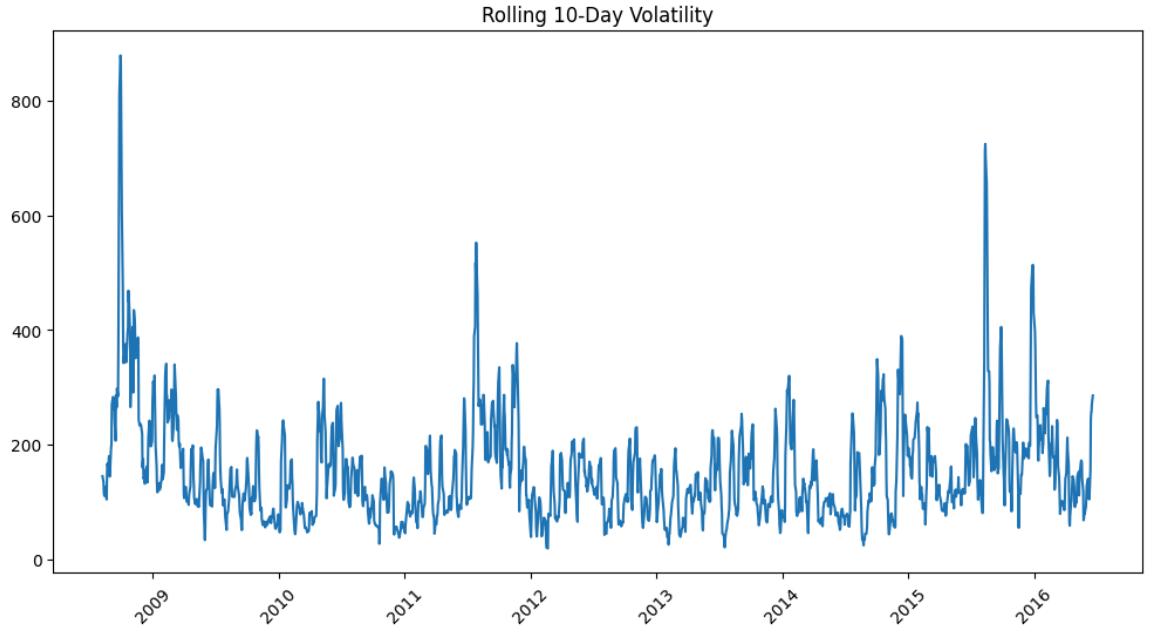


In [ ]:
```
plt.figure(figsize=(12,6))
plt.plot(stock_data['Date'], stock_data['Volume'])
plt.title('Trading Volume Over Time')
plt.xticks(rotation=45)
plt.show()
```

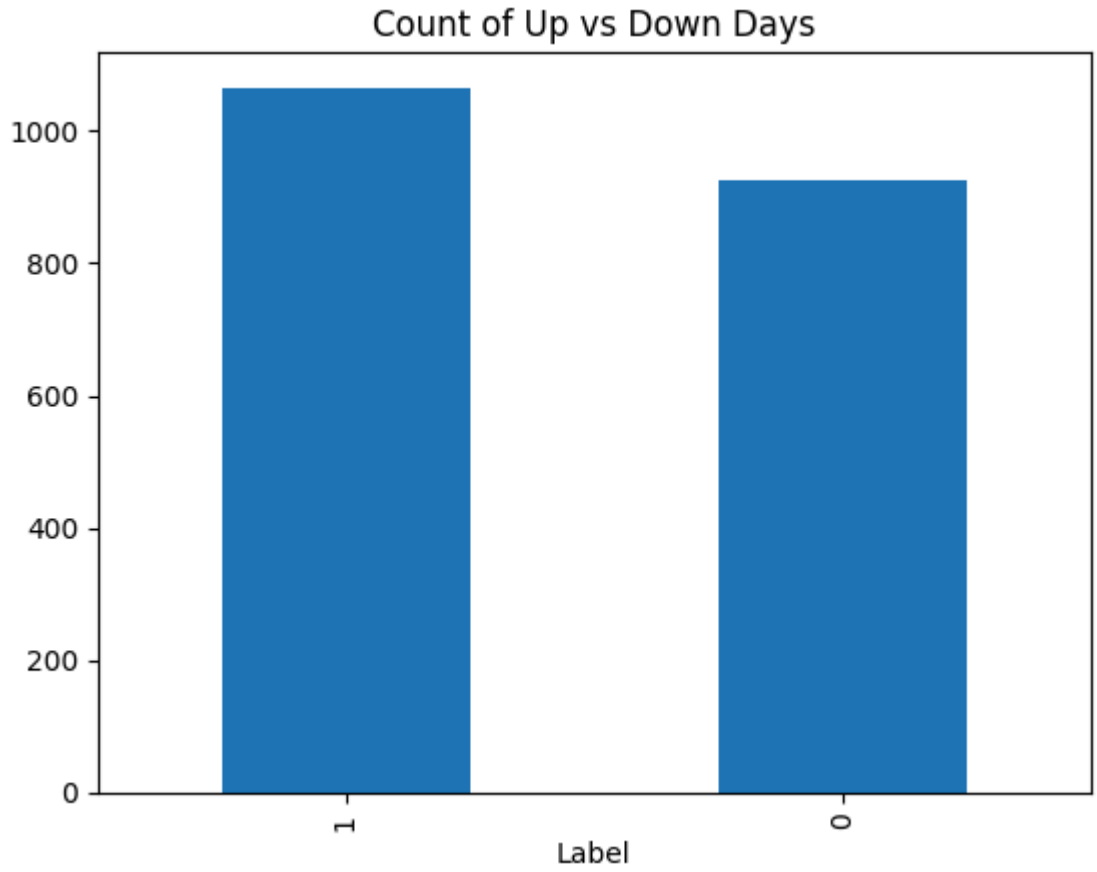


In [ ]:
```
stock_data['Volatility'] = stock_data['Close'].rolling(window=10).std()

plt.figure(figsize=(12,6))
plt.plot(stock_data['Date'], stock_data['Volatility'])
plt.title('Rolling 10-Day Volatility')
plt.xticks(rotation=45)
plt.show()
```



In [ ]:
```
combined_data['Label'].value_counts().plot(kind='bar')
plt.title('Count of Up vs Down Days')
plt.show()
```



In [ ]:
```
stock_features = stock_data[['Open', 'High', 'Low', 'Close', 'Volume']]

correlation_matrix = stock_features.corr()

print(correlation_matrix)
```

```
            Open      High       Low     Close    Volume
Open    1.000000  0.999592  0.999436  0.998991 -0.691621
High    0.999592  1.000000  0.999373  0.999546 -0.686997
Low     0.999436  0.999373  1.000000  0.999595 -0.699572
Close   0.998991  0.999546  0.999595  1.000000 -0.694281
Volume -0.691621 -0.686997 -0.699572 -0.694281  1.000000
```

In [ ]:
```
stock_features = stock_data[['Open', 'High', 'Low', 'Close', 'Volume']]

covariance_matrix = stock_features.cov()

print(covariance_matrix)
```

```
                Open          High           Low         Close        Volu
me
Open    9.880219e+06  9.854163e+06  9.897074e+06  9.872527e+06 -2.041858e+
11
High    9.854163e+06  9.836200e+06  9.874379e+06  9.855979e+06 -2.023682e+
11
Low     9.897074e+06  9.874379e+06  9.925152e+06  9.900935e+06 -2.070022e+
11
Close   9.872527e+06  9.855979e+06  9.900935e+06  9.884780e+06 -2.050184e+
11
Volume -2.041858e+11 -2.023682e+11 -2.070022e+11 -2.050184e+11  8.821610e+
15
```

In [ ]:
```
!pip install transformers
!pip install torch
```

```
Requirement already satisfied: transformers in /usr/local/lib/python3.11/d
ist-packages (4.51.3)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/
lib/python3.11/dist-packages (from transformers) (0.31.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/di
st-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.1
1/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/di
st-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python
3.11/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-
packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/py
thon3.11/dist-packages (from transformers) (0.21.1)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python
3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dis
t-packages (from transformers) (4.67.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.
11/dist-packages (from huggingface-hub<1.0,>=0.30.0->transformers) (2025.
3.2)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/li
b/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0->transformer
s) (4.13.2)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.0 in /usr/local/lib/pyth
on3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0->transformers) (1.
1.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/
python3.11/dist-packages (from requests->transformers) (3.4.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/d
ist-packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python
3.11/dist-packages (from requests->transformers) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python
3.11/dist-packages (from requests->transformers) (2025.4.26)
Requirement already satisfied: torch in /usr/local/lib/python3.11/dist-pac
kages (2.6.0+cu124)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from torch) (3.18.0)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/li
b/python3.11/dist-packages (from torch) (4.13.2)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-
packages (from torch) (3.4.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-pa
ckages (from torch) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-pa
ckages (from torch) (2025.3.2)
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch)
  Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_6
4.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch)
  Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86
_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch)
  Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_6
4.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch)
```

```
  Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.wh
l.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch)
  Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.wh
l.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.wh
l.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.
whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.
whl.metadata (1.6 kB)
Collecting nvidia-cusparse-cu12==12.3.1.170 (from torch)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_6
4.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/loca
l/lib/python3.11/dist-packages (from torch) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/
python3.11/dist-packages (from torch) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/li
b/python3.11/dist-packages (from torch) (12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_6
4.whl.metadata (1.5 kB)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/
dist-packages (from torch) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/
dist-packages (from torch) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python
3.11/dist-packages (from sympy==1.13.1->torch) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.1
1/dist-packages (from jinja2->torch) (3.0.2)
Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl
(363.4 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 3.8 MB/s eta 0:
00:00
Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.
whl (13.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 142.4 MB/s eta 0:
00:00
Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.
whl (24.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 99.7 MB/s eta 0:0
0:00
Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_6
4.whl (883 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 65.1 MB/s eta
0:00:00
Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl
(664.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 2.1 MB/s eta 0:
00:00
Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl
(211.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 5.4 MB/s eta 0:
00:00
Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.wh
l (56.3 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 15.5 MB/s eta 0:0
```

```
0:00
Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.wh
l (127.9 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 10.5 MB/s eta
0:00:00
Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.
whl (207.5 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 5.7 MB/s eta 0:
00:00
Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.w
hl (21.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 121.1 MB/s eta 0:
00:00
Installing collected packages: nvidia-nvjitlink-cu12, nvidia-curand-cu12,
nvidia-cufft-cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidi
a-cuda-cupti-cu12, nvidia-cublas-cu12, nvidia-cusparse-cu12, nvidia-cudnn-
cu12, nvidia-cusolver-cu12
  Attempting uninstall: nvidia-nvjitlink-cu12
    Found existing installation: nvidia-nvjitlink-cu12 12.5.82
    Uninstalling nvidia-nvjitlink-cu12-12.5.82:
      Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
  Attempting uninstall: nvidia-curand-cu12
    Found existing installation: nvidia-curand-cu12 10.3.6.82
    Uninstalling nvidia-curand-cu12-10.3.6.82:
      Successfully uninstalled nvidia-curand-cu12-10.3.6.82
  Attempting uninstall: nvidia-cufft-cu12
    Found existing installation: nvidia-cufft-cu12 11.2.3.61
    Uninstalling nvidia-cufft-cu12-11.2.3.61:
      Successfully uninstalled nvidia-cufft-cu12-11.2.3.61
  Attempting uninstall: nvidia-cuda-runtime-cu12
    Found existing installation: nvidia-cuda-runtime-cu12 12.5.82
    Uninstalling nvidia-cuda-runtime-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-runtime-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-nvrtc-cu12
    Found existing installation: nvidia-cuda-nvrtc-cu12 12.5.82
    Uninstalling nvidia-cuda-nvrtc-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.5.82
  Attempting uninstall: nvidia-cuda-cupti-cu12
    Found existing installation: nvidia-cuda-cupti-cu12 12.5.82
    Uninstalling nvidia-cuda-cupti-cu12-12.5.82:
      Successfully uninstalled nvidia-cuda-cupti-cu12-12.5.82
  Attempting uninstall: nvidia-cublas-cu12
    Found existing installation: nvidia-cublas-cu12 12.5.3.2
    Uninstalling nvidia-cublas-cu12-12.5.3.2:
      Successfully uninstalled nvidia-cublas-cu12-12.5.3.2
  Attempting uninstall: nvidia-cusparse-cu12
    Found existing installation: nvidia-cusparse-cu12 12.5.1.3
    Uninstalling nvidia-cusparse-cu12-12.5.1.3:
      Successfully uninstalled nvidia-cusparse-cu12-12.5.1.3
  Attempting uninstall: nvidia-cudnn-cu12
    Found existing installation: nvidia-cudnn-cu12 9.3.0.75
    Uninstalling nvidia-cudnn-cu12-9.3.0.75:
      Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
  Attempting uninstall: nvidia-cusolver-cu12
    Found existing installation: nvidia-cusolver-cu12 11.6.3.83
    Uninstalling nvidia-cusolver-cu12-11.6.3.83:
      Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
Successfully installed nvidia-cublas-cu12-12.4.5.8 nvidia-cuda-cupti-cu12-
12.4.127 nvidia-cuda-nvrtc-cu12-12.4.127 nvidia-cuda-runtime-cu12-12.4.127
nvidia-cudnn-cu12-9.1.0.70 nvidia-cufft-cu12-11.2.1.3 nvidia-curand-cu12-1
```

```
0.3.5.147 nvidia-cusolver-cu12-11.6.1.9 nvidia-cusparse-cu12-12.3.1.170 nv
idia-nvjitlink-cu12-12.4.127
```

In [ ]:
```
tokenizer = AutoTokenizer.from_pretrained("ProsusAI/finbert")
model = AutoModelForSequenceClassification.from_pretrained("ProsusAI/finber
t")
```

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94:
UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings
tab (https://huggingface.co/settings/tokens), set it as secret in your Goo
gle Colab and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to acces
s public models or datasets.
  warnings.warn(
```

In [ ]:
```
!pip install huggingface_hub[hf_xet]
```

```
Requirement already satisfied: huggingface_hub[hf_xet] in /usr/local/lib/p
ython3.11/dist-packages (0.31.1)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from huggingface_hub[hf_xet]) (3.18.0)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.
11/dist-packages (from huggingface_hub[hf_xet]) (2025.3.2)
Requirement already satisfied: packaging>=20.9 in /usr/local/lib/python3.1
1/dist-packages (from huggingface_hub[hf_xet]) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/di
st-packages (from huggingface_hub[hf_xet]) (6.0.2)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-
packages (from huggingface_hub[hf_xet]) (2.32.3)
Requirement already satisfied: tqdm>=4.42.1 in /usr/local/lib/python3.11/d
ist-packages (from huggingface_hub[hf_xet]) (4.67.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/li
b/python3.11/dist-packages (from huggingface_hub[hf_xet]) (4.13.2)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.0 in /usr/local/lib/pyth
on3.11/dist-packages (from huggingface_hub[hf_xet]) (1.1.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/
python3.11/dist-packages (from requests->huggingface_hub[hf_xet]) (3.4.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/d
ist-packages (from requests->huggingface_hub[hf_xet]) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python
3.11/dist-packages (from requests->huggingface_hub[hf_xet]) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python
3.11/dist-packages (from requests->huggingface_hub[hf_xet]) (2025.4.26)
```

In [ ]:
```
def get_finbert_sentiment_with_label(text):
    inputs = tokenizer(text, return_tensors="pt", truncation=True, max_leng
th=512)
    with torch.no_grad():
        outputs = model(**inputs)
        probs = softmax(outputs.logits, dim=1)

        negative = probs[0][0].item()
        neutral = probs[0][1].item()
        positive = probs[0][2].item()

        # Improved Neutral Handling
        if 0.2 < neutral < 0.6:
          sentiment_score = 0.0
          label = 'Neutral'
        else:
          sentiment_score = positive - negative
          # Avoid extreme +1.0 or -0.5 scores
          sentiment_score = max(min(sentiment_score, 0.9), -0.4)
          # Basic label assignment (refined in smart_sentiment)
          label = 'Positive' if sentiment_score > 0 else 'Negative'


    return sentiment_score, label
```

In [ ]:
```
# def smart_sentiment(text):
#     score, label = get_finbert_sentiment_with_label(text)

#     # Keyword lists
#     negative_words = ['fall', 'crash', 'breakdown', 'down', 'drop', 'coll
apse', 'bearish', 'loss', 'plunge', 'sell']
#     extreme_negative_words = ['crash', 'collapse', 'plummet', 'massive lo
ss', 'wipeout']
#     positive_words = ['rise', 'gain', 'soar', 'up', 'surge', 'bullish',
'profit', 'growth', 'buy']
#     neutral_words = ['neutral', 'flat', 'no change', 'unchanged', 'stabl
e', 'steady']

#     text_lower = text.lower()

#     # Apply Extreme Negative Penalty (Higher)
#     if any(word in text_lower for word in extreme_negative_words):
#         score -= 0.7
#         label = 'Negative'

#     # Apply Regular Negative Penalty
#     elif any(word in text_lower for word in negative_words):
#         score -= 0.4
#         label = 'Negative'

#     # Apply Positive Boost (Reduced)
#     elif any(word in text_lower for word in positive_words):
#         score += 0.3
#         label = 'Positive'

#     # Set Explicit Neutral Handling (Zero Out)
#     elif any(word in text_lower for word in neutral_words):
#         score = 0.0
#         label = 'Neutral'

#     # Clip score between -1 and 1 (No Extreme 1.0)
#     score = max(min(score, 0.9), -0.4)

#     # Final Label Assignment (Refined Thresholds)
#     if score > 0.3:
#         label = 'Positive'
#     elif score < -0.3:
#         label = 'Negative'
#     else:
#         label = 'Neutral'

#     return score, label

# def smart_sentiment(text):
#     score, label = get_finbert_sentiment_with_label(text)

#     # Clip score between -1 and 1 (No Extreme 1.0)
#     score = max(min(score, 0.9), -0.7)

#     # Final Label Assignment (Refined Thresholds)
#     if score > 0.3:
#         label = 'Positive'
#     elif score < -0.3:
#         label = 'Negative'
#     else:
```

```
#         label = 'Neutral'

#     return score, label

def smart_sentiment(text):
    score, label = get_finbert_sentiment_with_label(text)

    # Keyword lists
    extreme_negative_words = ['crash', 'collapse', 'plummet', 'wipeout', 'b
ankruptcy', 'liquidation', 'massive loss', 'wipe out']
    negative_words = ['fall', 'down', 'drop', 'loss', 'bearish', 'sell', 'b
reakdown', 'decline']
    positive_words = ['rise', 'gain', 'soar', 'up', 'surge', 'bullish', 'pr
ofit', 'growth', 'buy', 'record high']
    neutral_words = ['neutral', 'flat', 'no change', 'unchanged', 'stable',
'steady']

    text_lower = text.lower()

    # Apply Extreme Negative Penalty (Higher)
    if any(word in text_lower for word in extreme_negative_words):
        score -= 0.7
        label = 'Negative'

    # Apply Regular Negative Penalty
    elif any(word in text_lower for word in negative_words):
        score -= 0.4
        label = 'Negative'

    # Apply Positive Boost
    elif any(word in text_lower for word in positive_words):
        score += 0.4
        label = 'Positive'

    # Set Explicit Neutral Handling
    elif any(word in text_lower for word in neutral_words):
        score = 0.0
        label = 'Neutral'

    # Clip score between -1 and 1
    score = max(min(score, 0.9), -0.7)

    # Final Label Assignment (Refined Thresholds)
    if score > 0.3:
        label = 'Positive'
    elif score < -0.3:
        label = 'Negative'
    else:
        label = 'Neutral'

    return score, label
```

In [ ]:

```
# def simple_sentiment(text):
#     score, label = get_finbert_sentiment_with_label(text)
#     return score, label
```

In [ ]:
```
test_sentences = [
    "Tesla stock will rise",
    "Massive bearish breakdown expected",
    "Neutral news for Tesla stock",
    "Tesla stock will crash badly",
    "Tesla stock might go down a little",
    "Market remains stable today",
    "Huge profit expected for Tesla",
    "Tesla stock wiped out in a massive collapse"
]

for sentence in test_sentences:
    score, label = smart_sentiment(sentence)
    print(f"Sentence: '{sentence}'")
    print(f"Smart Sentiment Score: {score} | Label: {label}")
    print("-" * 60)
```

```
Sentence: 'Tesla stock will rise'
Smart Sentiment Score: 0.9 | Label: Positive
------------------------------------------------------------
Sentence: 'Massive bearish breakdown expected'
Smart Sentiment Score: -0.3136800989508629 | Label: Negative
------------------------------------------------------------
Sentence: 'Neutral news for Tesla stock'
Smart Sentiment Score: 0.0 | Label: Neutral
------------------------------------------------------------
Sentence: 'Tesla stock will crash badly'
Smart Sentiment Score: -0.5429103549569845 | Label: Negative
------------------------------------------------------------
Sentence: 'Tesla stock might go down a little'
Smart Sentiment Score: -0.3469357039779425 | Label: Negative
------------------------------------------------------------
Sentence: 'Market remains stable today'
Smart Sentiment Score: 0.0 | Label: Neutral
------------------------------------------------------------
Sentence: 'Huge profit expected for Tesla'
Smart Sentiment Score: 0.09099831581115725 | Label: Neutral
------------------------------------------------------------
Sentence: 'Tesla stock wiped out in a massive collapse'
Smart Sentiment Score: -0.5813862260431051 | Label: Negative
------------------------------------------------------------
```

In [ ]:
```
tqdm.pandas()

combined_data[['news_sentiment', 'news_label']] = combined_data['combined_news'].progress_apply(
    lambda x: pd.Series(smart_sentiment(x))
)
print(combined_data[['Date', 'combined_news', 'news_sentiment', 'news_label']].head())
```

```
100%|██████████| 1989/1989 [39:25<00:00,  1.19s/it]

        Date                                      combined_news  \
0 2008-08-08  b"Georgia 'downs two Russian warplanes' as cou...
1 2008-08-11  b'Why wont America and Nato help us? If they w...
2 2008-08-12  b'Remember that adorable 9-year-old who sang a...
3 2008-08-13  b' U.S. refuses Israel weapons to attack Iran:...
4 2008-08-14  b'All the experts admit that we should legalis...

   news_sentiment news_label
0            -0.4   Negative
1            -0.4   Negative
2            -0.4   Negative
3             0.4   Positive
4             0.4   Positive
```

In [ ]:
```
combined_data[['combined_news', 'news_sentiment', 'news_label']].to_csv(
    'temp_sentiment_check.csv',
    index=False
)
```

In [ ]:
```
tokenizer = AutoTokenizer.from_pretrained("ProsusAI/finbert")
model = AutoModelForSequenceClassification.from_pretrained(
    "ProsusAI/finbert",
    torch_dtype=torch.float32  # Skipping safetensors to increase colab run time efficiency by forcing pytorch model weights
)
```

In [ ]:
```
def get_finbert_sentiment(text):
    inputs = tokenizer(text, return_tensors="pt", truncation=True, max_length=512)
    with torch.no_grad():
        outputs = model(**inputs)
        probs = softmax(outputs.logits, dim=1)
        sentiment_score = probs[0][2] - probs[0][0]  # Positive - Negative
    return sentiment_score.item()
```

In [ ]:
```
tqdm.pandas()

reddit_data['reddit_sentiment'] = reddit_data['News'].progress_apply(get_finbert_sentiment)
```

```
100%|██████████| 73608/73608 [2:02:59<00:00,  9.97it/s]
```

In [ ]:
```
reddit_data[['Date', 'News', 'reddit_sentiment']].head()
combined_data[['Date', 'combined_news', 'news_sentiment']].head()
```

Out[ ]:

	Date	combined_news	news_sentiment
0	2008-08-08	b"Georgia 'downs two Russian warplanes' as cou...	-0.4
1	2008-08-11	b'Why wont America and Nato help us? If they w...	-0.4
2	2008-08-12	b'Remember that adorable 9-year-old who sang a...	-0.4
3	2008-08-13	b' U.S. refuses Israel weapons to attack Iran:...	0.4
4	2008-08-14	b'All the experts admit that we should legalis...	0.4

In [ ]:
```
merged_data = stock_data.merge(
    combined_data[['Date', 'news_sentiment', 'Label']],
    on='Date',
    how='inner'
)

merged_data = merged_data.merge(
    reddit_data.groupby('Date')['reddit_sentiment'].mean().reset_index(),
    on='Date',
    how='inner'
)
```

In [ ]:
```
merged_data['Target'] = (merged_data['Close'].shift(-1) > merged_data['Clos
e']).astype(int)
```

In [ ]:
```
merged_data['pct_change'] = merged_data['Close'].pct_change()
```

In [ ]:
```
merged_data['final_sentiment'] = merged_data['news_sentiment'] + merged_dat
a['reddit_sentiment']
```

In [ ]:
```
(merged_data['Target'] == merged_data['Label']).mean()
```

Out[ ]:
```
np.float64(0.0015082956259426848)
```

In [ ]:
```
merged_data.to_csv('multimodal_dataset_final2.csv', index=False)
```

In [ ]:
```
merged_data.to_csv('/content/drive/MyDrive/DJIA Dataset/multimodal_dataset_
final2.csv', index=False)
```