In [6]:

```
pip install pandas
```

```
Requirement already satisfied: pandas in /home/nzovia/anaconda3/lib/
python3.11/site-packages (1.5.3)
Requirement already satisfied: python-dateutil>=2.8.1 in /home/nzovi
a/anaconda3/lib/python3.11/site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /home/nzovia/anaconda
3/lib/python3.11/site-packages (from pandas) (2022.7)
Requirement already satisfied: numpy>=1.21.0 in /home/nzovia/anacond
a3/lib/python3.11/site-packages (from pandas) (1.24.3)
Requirement already satisfied: six>=1.5 in /home/nzovia/anaconda3/li
b/python3.11/site-packages (from python-dateutil>=2.8.1->pandas) (1.
16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [7]:

```
import pandas as pd
import numpy as np
```

In [8]:

```
df = pd.read_csv('labelled.csv', delimiter=';')
```

In [9]:

```
df.head()
```

Out[9]:

	id	created_at	text	sentiment
0	77522	2020-04-15 01:03:46+00:00	RT @RobertBeadles: Yo💥\nEnter to WIN 1,000 Mon...	positive
1	661634	2020-06-25 06:20:06+00:00	#SriLanka surcharge on fuel removed!\n⛽📉\nThe ...	negative
2	413231	2020-06-04 15:41:45+00:00	Net issuance increases to fund fiscal programs...	positive
3	760262	2020-07-03 19:39:35+00:00	RT @bentboolean: How much of Amazon's traffic ...	positive
4	830153	2020-07-09 14:39:14+00:00	$AMD Ryzen 4000 desktop CPUs looking ‘great’ a...	positive

In [10]:

```
pip install clean-text
```

```
Requirement already satisfied: clean-text in /home/nzovia/anaconda3/
lib/python3.11/site-packages (0.6.0)
Requirement already satisfied: emoji<2.0.0,>=1.0.0 in /home/nzovia/a
naconda3/lib/python3.11/site-packages (from clean-text) (1.7.0)
Requirement already satisfied: ftfy<7.0,>=6.0 in /home/nzovia/anacon
da3/lib/python3.11/site-packages (from clean-text) (6.1.1)
Requirement already satisfied: wcwidth>=0.2.5 in /home/nzovia/anacon
da3/lib/python3.11/site-packages (from ftfy<7.0,>=6.0->clean-text)
(0.2.5)
Note: you may need to restart the kernel to use updated packages.
```

In [11]:

```
from cleantext import clean
```

In [12]:

```
!pip install regex
```

```
Requirement already satisfied: regex in /home/nzovia/anaconda3/lib/p
ython3.11/site-packages (2022.7.9)
```

In [13]:

```

import regex as re


```

In [14]:

```
# Define the remove_emojis function (if you haven't defined it yet)
```

In [15]:

```
def remove_emojis(text):
    emoji_pattern = re.compile("[\U0001F600-\U0001F64F\U0001F300-\U0001F5FF\U0001
    return emoji_pattern.sub(r'', text)
```

In [16]:

```
 # Apply the function to the 'text' column
```

In [17]:

```
df['text'] = df['text'].apply(remove_emojis)
```

In [18]:

```
df.head()
```

Out[18]:

	id	created_at	text	sentiment
0	77522	2020-04-15 01:03:46+00:00	RT @RobertBeadles: Yo\nEnter to WIN 1,000 Mona...	positive
1	661634	2020-06-25 06:20:06+00:00	#SriLanka surcharge on fuel removed!\n\nThe su...	negative
2	413231	2020-06-04 15:41:45+00:00	Net issuance increases to fund fiscal programs...	positive
3	760262	2020-07-03 19:39:35+00:00	RT @bentboolean: How much of Amazon's traffic ...	positive
4	830153	2020-07-09 14:39:14+00:00	$AMD Ryzen 4000 desktop CPUs looking ‘great’ a...	positive

In [19]:

```
def remove_user_mentions(text):
      # Define the regular expression pattern to match user mentions
    user_mention_pattern = re.compile(r'@\w+')
     # Replace user mentions with an empty string
    cleaned_text = user_mention_pattern.sub('', text)

    return cleaned_text
```

In [20]:

```
# Define a function to username from the text
```

In [21]:

```
df['text'] = df['text'].apply(remove_user_mentions)
```

In [22]:

```
df.head()
```

Out[22]:

	id	created_at	text	sentiment
0	77522	2020-04-15 01:03:46+00:00	RT : Yo\nEnter to WIN 1,000 Monarch Tokens\n\n...	positive
1	661634	2020-06-25 06:20:06+00:00	#SriLanka surcharge on fuel removed!\n\nThe su...	negative
2	413231	2020-06-04 15:41:45+00:00	Net issuance increases to fund fiscal programs...	positive
3	760262	2020-07-03 19:39:35+00:00	RT : How much of Amazon's traffic is served by...	positive
4	830153	2020-07-09 14:39:14+00:00	$AMD Ryzen 4000 desktop CPUs looking ‘great’ a...	positive

In [23]:

```
#To remove The #tag from my data
```

In [24]:

```
def remove_hashtags(text):
    # Define the regular expression patterns to match user mentions and hashtags
    hashtag_pattern = re.compile(r'#\w+')

    # Replace hashtags with an empty string
    cleaned_text = hashtag_pattern.sub('', text)

    return cleaned_text
```

In [25]:

```
df['text'] = df['text'].apply(remove_hashtags)
```

In [26]:

```
df.head()
```

Out[26]:

	id	created_at	text	sentiment
0	77522	2020-04-15 01:03:46+00:00	RT : Yo\nEnter to WIN 1,000 Monarch Tokens\n\n...	positive
1	661634	2020-06-25 06:20:06+00:00	surcharge on fuel removed!\n\nThe surcharge o...	negative
2	413231	2020-06-04 15:41:45+00:00	Net issuance increases to fund fiscal programs...	positive
3	760262	2020-07-03 19:39:35+00:00	RT : How much of Amazon's traffic is served by...	positive
4	830153	2020-07-09 14:39:14+00:00	$AMD Ryzen 4000 desktop CPUs looking ‘great’ a...	positive

In [27]:

```
def remove_special_characters(text):
     special_chars_pattern = re.compile(r'[^a-zA-Z0-9\s]')
     return special_chars_pattern.sub('', text)
```

In [28]:

```
df['text'] = df['text'].apply(remove_special_characters)
```

In [29]:

```
df.head()
```

Out[29]:

	id	created_at	text	sentiment
0	77522	2020-04-15 01:03:46+00:00	RT Yo\nEnter to WIN 1000 Monarch Tokens\n\nUS...	positive
1	661634	2020-06-25 06:20:06+00:00	surcharge on fuel removed\n\nThe surcharge of...	negative
2	413231	2020-06-04 15:41:45+00:00	Net issuance increases to fund fiscal programs...	positive
3	760262	2020-07-03 19:39:35+00:00	RT How much of Amazons traffic is served by F...	positive
4	830153	2020-07-09 14:39:14+00:00	AMD Ryzen 4000 desktop CPUs looking great and ...	positive

In [67]:

```
df1 = pd.read_csv('twitter-stocks (1).csv')
```

In [68]:

```
df.head()
```

Out[68]:

	id	Date	text	sentiment
Index				
1501	874716	2020-07-13 19:04:37+00:00	GOOGL GOOG Searches for Chainlink Hits Record ...	NaN
2586	110418	2020-04-18 19:22:33+00:00	We watch NFLX AMZN and on our ROKU device Pl...	NaN
2653	888454	2020-07-13 02:54:01+00:00	RT Lol at ES futures Record number of coronav...	NaN
1055	711593	2020-06-30 22:48:16+00:00	RT SPX SPY Can we close out this quarter alre...	negative
705	155304	2020-04-22 17:42:14+00:00	RT JNJ AdVac looks like best nearterm prospe...	positive

In [32]:

```
df1.head()
```

Out[32]:

	Date	Open	High	Low	Close	Adj Close	Volume
0	2013-11-07	45.099998	50.090000	44.000000	44.900002	44.900002	117701670.0
1	2013-11-08	45.930000	46.939999	40.685001	41.650002	41.650002	27925307.0
2	2013-11-11	40.500000	43.000000	39.400002	42.900002	42.900002	16113941.0
3	2013-11-12	43.660000	43.779999	41.830002	41.900002	41.900002	6316755.0
4	2013-11-13	41.029999	42.869999	40.759998	42.599998	42.599998	8688325.0

In [33]:

```
#looking at the dimension of the two data
```

In [34]:

```
df.shape
```

Out[34]:

```
(5000, 4)
```

In [35]:

```
df1.shape
```

Out[35]:

```
(2259, 7)
```

In [36]:

```
#To reduce data to 2259 rows
```

In [37]:

```
import numpy as np
```

In [38]:

```
sample_size = 2259
random_sample_df = df.sample(n=sample_size, random_state=42)
```

In [39]:

```
df = random_sample_df
```

In [40]:

```
print(df.shape)
(2259, 4)
```

```
(2259, 4)
```

In [41]:

```
df.shape
```

Out[41]:

```
(2259, 4)
```

In [42]:

```
#naming my first column
```

In [87]:

```
df = df.rename_axis('Index')
```

In [88]:

```
df.head()
```

Out[88]:

	id	text	sentiment
Index			
0	301411	Amedisys Inc AMED COO Christopher Gerard Sells...	NaN
1	890123	DIS it could break the 120 pin then 125gt 130 ...	NaN
2	62318	RT Well another point to add to dent the curr...	NaN
3	411380	With ad revenues falling whats the impact on s...	NaN
4	766908	Your ordinary person would focus on buying pur...	NaN

In [89]:

```
df.reset_index(drop=True, inplace=True)
```

In [90]:

```
df = df.rename_axis('Index')
```

In [91]:

```
df.head()
```

Out[91]:

	id	text	sentiment
Index			
0	301411	Amedisys Inc AMED COO Christopher Gerard Sells...	NaN
1	890123	DIS it could break the 120 pin then 125gt 130 ...	NaN
2	62318	RT Well another point to add to dent the curr...	NaN
3	411380	With ad revenues falling whats the impact on s...	NaN
4	766908	Your ordinary person would focus on buying pur...	NaN

In [92]:

```
df1 = df1.rename_axis('Index')
```

In [93]:

```
df1.head()
```

Out[93]:

	Date	Open	High	Low	Close	Adj Close	Volume
Index							
0	2013-11-07	45.099998	50.090000	44.000000	44.900002	44.900002	117701670.0
1	2013-11-08	45.930000	46.939999	40.685001	41.650002	41.650002	27925307.0
2	2013-11-11	40.500000	43.000000	39.400002	42.900002	42.900002	16113941.0
3	2013-11-12	43.660000	43.779999	41.830002	41.900002	41.900002	6316755.0
4	2013-11-13	41.029999	42.869999	40.759998	42.599998	42.599998	8688325.0

In [ ]:

```
# Drop Column created_at
```

In [105]:

```
df.head()
```

Out[105]:

	id	text	sentiment
Index			
0	301411	Amedisys Inc AMED COO Christopher Gerard Sells...	NaN
1	890123	DIS it could break the 120 pin then 125gt 130 ...	NaN
2	62318	RT Well another point to add to dent the curr...	NaN
3	411380	With ad revenues falling whats the impact on s...	NaN
4	766908	Your ordinary person would focus on buying pur...	NaN

In [106]:

```
pd.merge(df1, df, on='Index')
```

Out[106]:

	Date	Open	High	Low	Close	Adj Close	Volume	id	
Index									
0	2013-11-07	45.099998	50.090000	44.000000	44.900002	44.900002	117701670.0	301411	(
1	2013-11-08	45.930000	46.939999	40.685001	41.650002	41.650002	27925307.0	890123	I
2	2013-11-11	40.500000	43.000000	39.400002	42.900002	42.900002	16113941.0	62318	ar a
3	2013-11-12	43.660000	43.779999	41.830002	41.900002	41.900002	6316755.0	411380	re
4	2013-11-13	41.029999	42.869999	40.759998	42.599998	42.599998	8688325.0	766908	fo
...	...	...	...	...	...	...	...	...	
2254	2022-10-21	50.000000	50.750000	49.549999	49.889999	49.889999	51209029.0	436313	S
2255	2022-10-24	50.709999	51.860001	50.520000	51.520000	51.520000	22987553.0	313771	N h
2256	2022-10-25	52.415001	53.180000	52.200001	52.779999	52.779999	35077848.0	392845	F
2257	2022-10-26	52.950001	53.500000	52.770000	53.349998	53.349998	28064973.0	472959	8(Sa 1
2258	2022-10-27	53.910000	54.000000	53.700001	53.700001	53.700001	136345128.0	77522	RT To

2259 rows × 10 columns

In [107]:

```
print(merged_df.columns)
```

```
Index(['Date_left', 'Open', 'High', 'Low', 'Close', 'Adj Close', 'Vo
lume',
       'id', 'Date_right', 'text', 'sentiment'],
      dtype='object')
```

In [110]:

```
pd.merge(df1, df, on='Index', suffixes=('_left', '_right'))
```

Out[110]:

	Date	Open	High	Low	Close	Adj Close	Volume	id	
Index									
0	2013-11-07	45.099998	50.090000	44.000000	44.900002	44.900002	117701670.0	301411	(
1	2013-11-08	45.930000	46.939999	40.685001	41.650002	41.650002	27925307.0	890123	I
2	2013-11-11	40.500000	43.000000	39.400002	42.900002	42.900002	16113941.0	62318	ar a
3	2013-11-12	43.660000	43.779999	41.830002	41.900002	41.900002	6316755.0	411380	re
4	2013-11-13	41.029999	42.869999	40.759998	42.599998	42.599998	8688325.0	766908	fo
...	...	...	...	...	...	...	...	...	
2254	2022-10-21	50.000000	50.750000	49.549999	49.889999	49.889999	51209029.0	436313	S
2255	2022-10-24	50.709999	51.860001	50.520000	51.520000	51.520000	22987553.0	313771	N h
2256	2022-10-25	52.415001	53.180000	52.200001	52.779999	52.779999	35077848.0	392845	F
2257	2022-10-26	52.950001	53.500000	52.770000	53.349998	53.349998	28064973.0	472959	8(S: 1
2258	2022-10-27	53.910000	54.000000	53.700001	53.700001	53.700001	136345128.0	77522	R' Tc

2259 rows × 10 columns

In [111]:

```
print(merged_df.columns)
```

```
Index(['Date', 'Open', 'High', 'Low', 'Close', 'Adj Close', 'Volum
e', 'id',
       'text', 'sentiment'],
      dtype='object')
```

In [112]:

```
merged_df = merged_df[['Date', 'id', 'text', 'Open', 'High', 'Low', 'Close', 'Adj
```

In [113]:

```
merged_df.head()
```

Out[113]:

	Date	id	text	Open	High	Low	Close	Adj Close	
Index									
0	2013-11-07	301411	Amedisys Inc AMED COO Christopher Gerard Sells...	45.099998	50.090000	44.000000	44.900002	44.900002	117
1	2013-11-08	890123	DIS it could break the 120 pin then 125gt 130 ...	45.930000	46.939999	40.685001	41.650002	41.650002	27
2	2013-11-11	62318	RT Well another point to add to dent the curr...	40.500000	43.000000	39.400002	42.900002	42.900002	16
3	2013-11-12	411380	With ad revenues falling whats the impact on s...	43.660000	43.779999	41.830002	41.900002	41.900002	6
4	2013-11-13	766908	Your ordinary person would focus on buying pur...	41.029999	42.869999	40.759998	42.599998	42.599998	8

In [114]:

```
merged_df.to_csv('modified_data.csv', index=False)
```

In [115]:

```
from IPython import display
```

In [116]:

```
display.FileLink('modified_data.csv')
```

Out[116]:

modified_data.csv (modified_data.csv)

In [ ]:

```

```