CapX

Name: Yatharth Asthana

1.) Data collection and importing necessary libraries

```
import tweepy
import re
import pandas as pd
import numpy as np
!pip install praw
import praw
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import nltk
nltk.download('punkt')
nltk.download('stopwords')

# Reddit API credentials
reddit = praw.Reddit(client_id='CF5EOmlCOrpL7qbc3GhxmA',
                     client_secret='zRq-FDzbq0xQ4gXGzZ_EmpdppbqxOg',
                     user_agent='StockSentimentAnalysis')

subreddit = reddit.subreddit('wallstreetbets')
posts = subreddit.search('GME', limit=1000)
```

2.)

 # Store post data in DataFrame

posts_data = [[post.title, post.selftext, post.score, post.created_utc] for post in posts]

df_posts = pd.DataFrame(posts_data, columns=['Title', 'Body', 'Score', 'Created'])

df_posts.to_csv('/content/stocks_posts.csv', index=False)

df = pd.read_csv('/content/stocks_posts.csv')

	Title	Body	Score	Created
0	Gme 5/17 yolo	NaN	1005	1.714681e+09
1	My best GME gain porn. $1,295 --> $959,663 in ...	NaN	45209	1.612819e+09
2	Whatever happened with the GME fiasco?	I tried searching this subreddit for some clos...	85	1.658715e+09
3	DEAD FUCKING BABA BRINGS YOU A NEW YOLO	NaN	1396	1.727117e+09
4	(GME) Gamestop earnings. Let's take a closer l...	**Gamestop (GME) Earnings Summary:**\n\n* EPS ...	9376	1.679439e+09
...	...	...	...	...
232	GME YOLO UPDATE: BACK OVER $5,000,000! (20,000...	NaN	27290	1.615487e+09
233	GME liquidy is drying up - causing the share t...	[https://i.imgur.com/DxM4SwP.png](https://i.im...	19215	1.612303e+09
234	🚨🚀🚨 Linus just ran a stream where he would 5x ...	NaN	40635	1.611976e+09
235	GME 2/11	GME/Jim Cramer megathread	8942	1.613049e+09
236	To those that recently joined WSB, this GME ev...	I hope new retards that just joined WSB needs ...	23047	1.612210e+09

237 rows × 4 columns

Function to clean text

def clean_text(text):

```
    text = re.sub(r'http\S+', '', text)  # Remove URLs
    text = re.sub(r'@\w+', '', text)  # Remove mentions
    text = re.sub(r'#', '', text)  # Remove hashtags
    text = re.sub(r'[^A-Za-z\s]', '', text)  # Remove special characters
    text = text.lower()  # Convert to lowercase
    text = ' '.join([word for word in word_tokenize(text) if word not in
stopwords.words('english')])
    return text

# Apply cleaning function to the dataset
df['Cleaned_Text'] = df['Title'].apply(clean_text)

df.to_csv('/content/cleaned_stock_tweets.csv', index=False)

!pip install vaderSentiment
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

df1=pd.read_csv('/content/cleaned_stock_tweets.csv')
```

```
[ ] df1
```

	Title	Body	Score	Created	Cleaned_Text
0	Gme 5/17 yolo	NaN	1005	1.714681e+09	gme yolo
1	My best GME gain porn. $1,295 --> $959,663 in ...	NaN	45209	1.612819e+09	best gme gain porn weeks sold brokers blocking...
2	Whatever happened with the GME fiasco?	I tried searching this subreddit for some clos...	85	1.658715e+09	whatever happened gme fiasco
3	DEAD FUCKING BABA BRINGS YOU A NEW YOLO	NaN	1396	1.727117e+09	dead fucking baba brings new yolo
4	(GME) Gamestop earnings. Let's take a closer l...	**Gamestop (GME) Earnings Summary:**\n\n* EPS ...	9376	1.679439e+09	gme gamestop earnings lets take closer look
...	...	...	...	...	...
232	GME YOLO UPDATE: BACK OVER $5,000,000! (20,000...	NaN	27290	1.615487e+09	gme yolo update back shares roller coaster fun...
233	GME liquidy is drying up - causing the share t...	[https://i.imgur.com/DxM4SwP.png](https://i.im...	19215	1.612303e+09	gme liquidy drying causing share become volatile
234	🚨🚀🚨 Linus just ran a stream where he would 5x ...	NaN	40635	1.611976e+09	linus ran stream would x donations buy gme k r...
235	GME 2/11	GME/Jim Cramer megathread	8942	1.613049e+09	gme
236	To those that recently joined WSB, this GME ev...	I hope new retards that just joined WSB needs ...	23047	1.612210e+09	recently joined wsb gme event may take

237 rows × 5 columns

```
[ ] # Initialize VADER sentiment analyzer
    analyzer = SentimentIntensityAnalyzer()
```

 # Function to get sentiment scores

```
def get_sentiment(text):
    score = analyzer.polarity_scores(text)
    return score['compound']
```

 # Apply sentiment analysis

```
df['Sentiment_Score'] = df['Cleaned_Text'].apply(get_sentiment)
```

 # Label sentiment as Positive, Negative, or Neutral

```
df['Sentiment_Label'] = df['Sentiment_Score'].apply(lambda x: 'Positive' if x > 0 else ('Negative' if x < 0 else 'Neutral'))
```

 # Save results

```
df.to_csv('sentiment_stock_tweets.csv', index=False)
```

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation

# Count Vectorizer to find mentions of specific stocks
vectorizer = CountVectorizer(stop_words='english', ngram_range=(1, 1))
X = vectorizer.fit_transform(df['Cleaned_Text'])

# Identify words related to stocks
stock_words = ['AAPL', 'GOOGL', 'TSLA', 'GME']  # Add stock symbols here
mention_matrix = X[:, [vectorizer.vocabulary_.get(w.lower()) for w in
stock_words if w.lower() in vectorizer.vocabulary_]]

# Get top words per topic
def print_top_words(model, feature_names, n_top_words):
    for topic_idx, topic in enumerate(model.components_):
        print(f"Topic #{topic_idx}: ", " ".join([feature_names[i] for i in
topic.argsort()[:-n_top_words - 1:-1]]))

print_top_words(lda, vectorizer.get_feature_names_out(), 10)



import matplotlib.pyplot as plt
from scipy.stats import pearsonr

# Load sentiment data and stock price data
sentiment_df = pd.read_csv('/content/sentiment_stock_tweets.csv')
```

stock_price_df = pd.read_csv('/content/stocks_posts.csv') # Assuming you have stock price data

 # Merge data on time or another relevant feature

merged_df = pd.merge(sentiment_df, stock_price_df, on='Created')

print(stock_price_df.columns)

```
[ ] stock_price_df
```

	Title	Body	Score	Created
0	Gme 5/17 yolo	NaN	1005	1.714681e+09
1	My best GME gain porn. $1,295 --> $959,663 in ...	NaN	45209	1.612819e+09
2	Whatever happened with the GME fiasco?	I tried searching this subreddit for some clos...	85	1.658715e+09
3	DEAD FUCKING BABA BRINGS YOU A NEW YOLO	NaN	1396	1.727117e+09
4	(GME) Gamestop earnings. Let's take a closer l...	**Gamestop (GME) Earnings Summary:**\n\n* EPS ...	9376	1.679439e+09
...	...	...	...	...
232	GME YOLO UPDATE: BACK OVER $5,000,000! (20,000...	NaN	27290	1.615487e+09
233	GME liquidy is drying up - causing the share t...	[https://i.imgur.com/DxM4SwP.png](https://i.im...	19215	1.612303e+09
234	🚨🚀🚨 Linus just ran a stream where he would 5x ...	NaN	40635	1.611976e+09
235	GME 2/11	GME/Jim Cramer megathread	8942	1.613049e+09
236	To those that recently joined WSB, this GME ev...	I hope new retards that just joined WSB needs ...	23047	1.612210e+09

237 rows × 4 columns

```
[ ] #stock_price_df['Stock_Price_Change'] = stock_price_df['Close'].pct_change()
    merged_df.columns
```

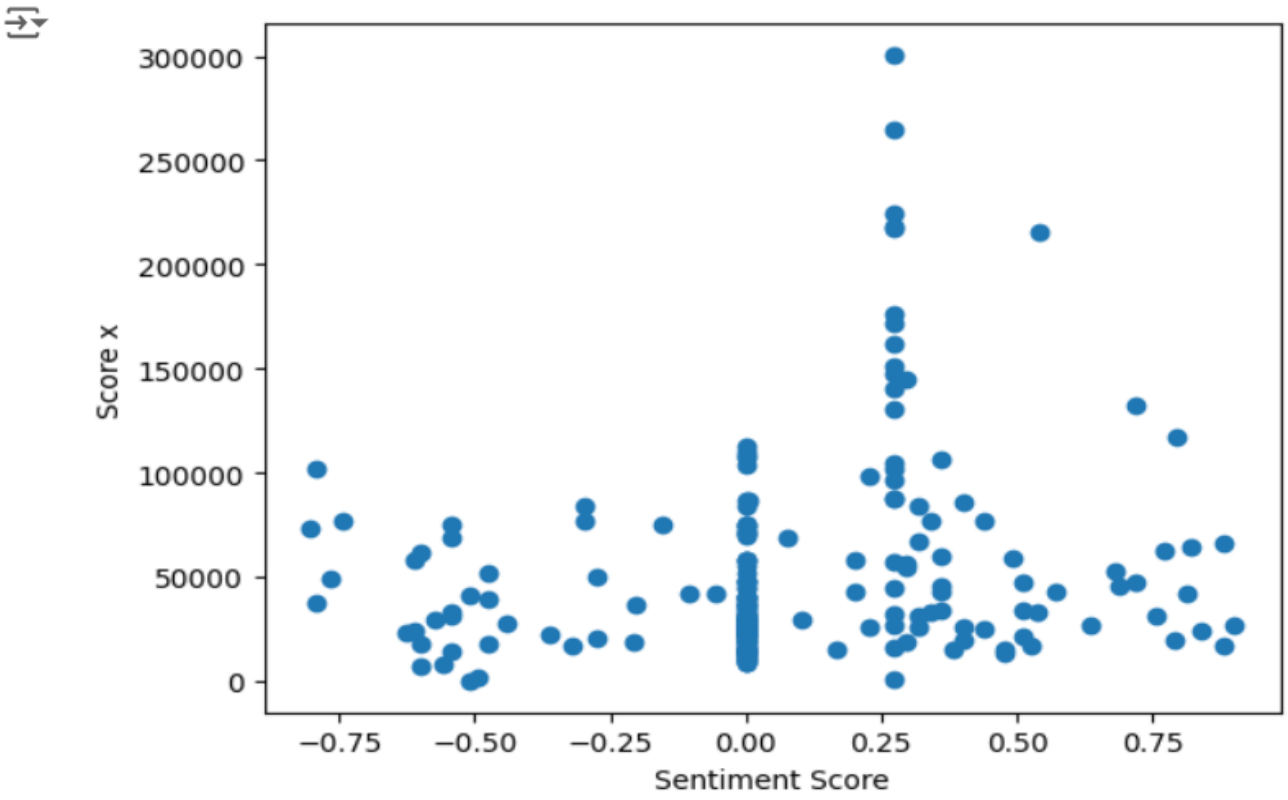
corr, _ = pearsonr(merged_df['Sentiment_Score'], merged_df['Score_x'])

print(f'Correlation between sentiment score and score x: {corr}')

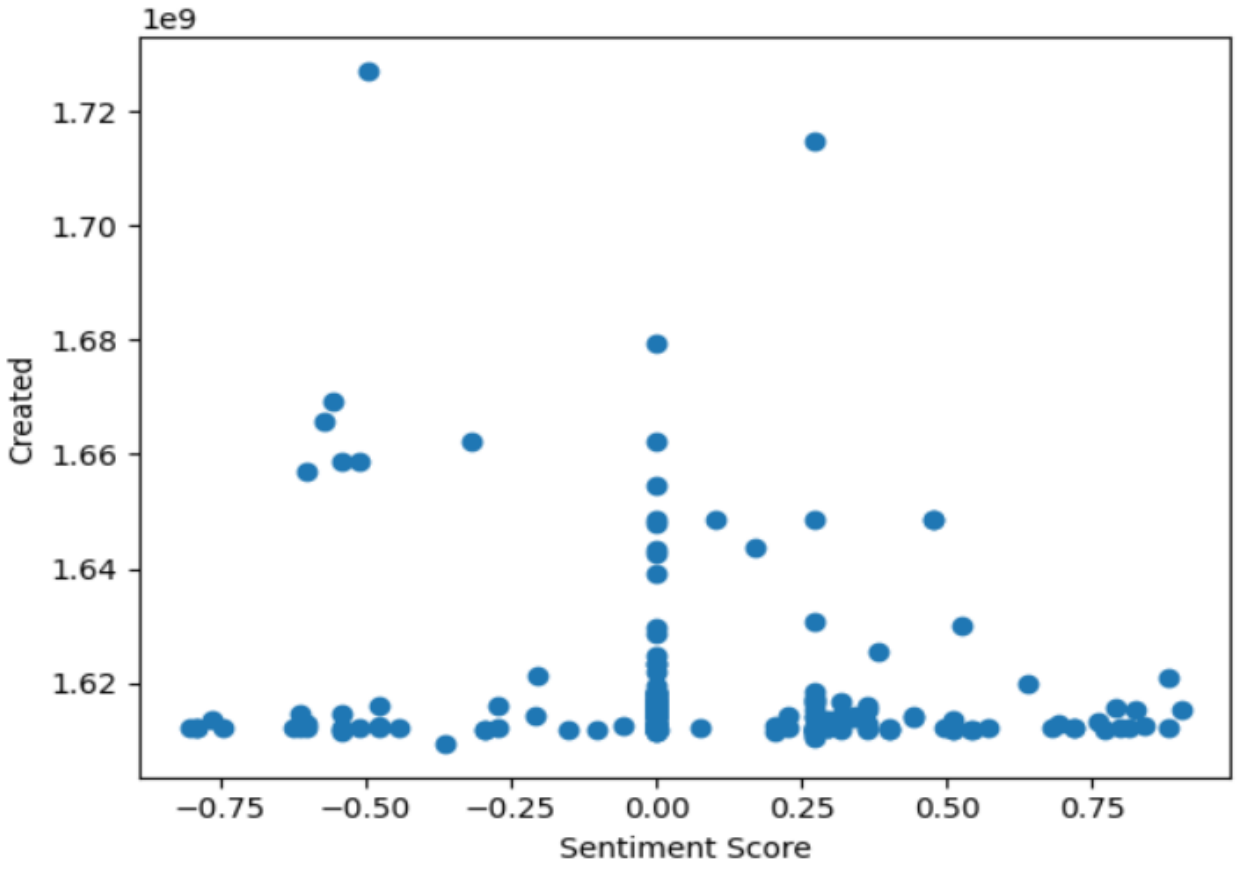
corr1, _ = pearsonr(merged_df['Sentiment_Score'], merged_df['Created'])

print(f'Correlation between sentiment score and Created time: {corr1}')

```
[ ]  # Plot sentiment vs stock price change
     plt.scatter(merged_df['Sentiment_Score'], merged_df['Score_x'])
     plt.xlabel('Sentiment Score')
     plt.ylabel('Score x')
     plt.show()
```



```
# Plot sentiment vs created
plt.scatter(merged_df['Sentiment_Score'], merged_df['Created'])
plt.xlabel('Sentiment Score')
plt.ylabel('Created')
plt.show()
```




```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

# Prepare features (Sentiment Score) and target (Stock Price Change)
X = merged_df[['Sentiment_Score']]
y = merged_df['Score_x']
```

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

 # Build Linear Regression model

model = LinearRegression()

model.fit(X_train, y_train)

 # Predict and evaluate

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)

print(f'Mean Squared Error: {mse}')

output: Mean Squared Error: 2314710322.6136622