DSCI 510 Final Project Report

1. Project Overview

Project Title: Do Social Media Sentiments Drive U.S. Stock Market Reactions?

Team Members:

Mengyue Niu	USC ID: 5926-7082-41	USC Email: mengyuen@usc.edu
Yichen Yang	USC ID: 4403-8375-34	USC Email: yyang944@usc.edu

Research Question:

How do investor sentiments and discussion volumes on social media platforms relate to short-term U.S. stock market behavior?

Project Description:

This project investigates whether Reddit investor sentiment predicts stock returns and trading activity. Using a six-month dataset from Reddit's most active financial subreddits, the project integrates social media metrics—such as discussion volume and sentiment polarity—with historical market data from Yahoo Finance. Our findings reveal sentiment has negligible predictive power for returns, with near-zero correlations and statistically insignificant differences. However, discussion volume shows consistent positive correlation with abnormal trading activity. This study provides empirical evidence on social media's limited predictive role in normal market conditions.

2. Data Collection

Two major datasets were collected:

2.1 Reddit Discussion Data

Source: Reddit's public JSON API.

Method: Implemented in get_data.py with three fallback modes: (1) PRAW (with credentials), (2) Pullpush mirror, (3) Public JSON endpoint. Because my Reddit API access request was not approved and the Pullpush method failed to retrieve comments, I ultimately used the public JSON endpoint to collect the data.

Scope: 10 financial subreddits (r/stocks, r/wallstreetbets, r/investing, r/StockMarket, r/options, r/RobinHood, r/pennystocks, r/shortsqueeze, r/daytrading, r/techstocks); 6-month window (2025-05-15 to 2025-11-15).

Output: data/raw/reddit_posts_full_6mo.csv containing 17686 records.

2.2 Stock Market Data

Source: Yahoo Finance API via the yfinance library.

Method: Implemented in get_data.py under the prices command. The script uses the Yahoo Finance API through the yfinance Python library to fetch historical stock data.

Scope: For the 10 most discussed stocks on Reddit (TSLA, GOOG, NVDA, MSFT, GME, PLTR, AMZN, WMT, AAPL, and INTC), daily OHLCV (Open, High, Low, Close, Adj Close, Volume) data were collected over a 6-month window (2025-05-15 to 2025-11-15).

Output: data/raw/prices_6mo.csv, containing 1280 records.

3. Data Cleaning and Preprocessing

3.1 Reddit Discussion Data

After retrieving the raw Reddit data, it was cleaned and standardized through clean_social_media.py, which performed:

a) **Text Standardization**: Merged title and body into a single content field, removed HTML entities, Markdown symbols, spam phrases, and extreme-length texts. Only key columns were retained: kind, subreddit, id, parent_id, created_utc, author, score, num_comments, and content.

b) **Timestamp Conversion**: Transformed created_utc from UNIX time to readable UTC dates (date_utc and datetime_utc). Timezone information was removed to ensure uniformity.

c) **Deduplication and Quality Filtering**: Removed duplicate IDs and identical texts; discarded advertisement-like or non-informative content. Texts that were too short (<10 characters) or too long (>5000 characters) were discarded.

d) **Ticker Recognition**: Created a reference table ticker_universe.csv including ticker, company name and aliases for fuzzy matching. Each post mentioning multiple tickers was exploded into multiple rows.

e) **Sentiment Analysis**: Applied the VADER analyzer to compute compound sentiment scores, ranging from -1 (most negative) to +1 (most positive).The negative (neg), neutral (neu), and positive (pos) proportions were also retained for potential fine-grained analysis. Extreme sentiment values were clipped to mitigate outlier effects.

f) **Aggregation and Output**:

1) Row-level: socialmediadataclean_rows_6mo.csv (1625 records)
2) Daily aggregation: socialmediadataclean_daily_6mo.csv (824 records)
3) Top tickers summary: socialmediadataclean_top_tickers_6mo.csv (39 stocks)

g) **Weekly Aggregation:** The daily Reddit data were sparse, so the next was to aggregate by week using aggregate_weekly.py, yielding socialmediadataclean_weekly_6mo.csv.

3.2 Stock Market Data

The stock price data were cleaned through the function cmd_prices() in get_data.py.

a) **Unified Column Format:** Since Yahoo Finance returns inconsistent structures (MultiIndex like ('TSLA', 'Open') or columns such as TSLA_Open, Open_TSLA), the function automatically parses and renames all columns to the standardized schema: ['date', 'ticker', 'Open', 'High', 'Low', 'Close', 'Adj Close', 'Volume'].

b) **Column Name Normalization:** Regular expressions (re.split(r"[^a-z]+", col_norm)) identify key financial terms (“open”, “high”, “adjclose”, etc.) and correctly map each column even when formatting varies (e.g., AAPL.AdjClose, Adj Close_AAPL).

c) **Missing Column Handling:** When “Adj Close” is missing for some stocks or time periods, the “Close” column is used as a substitute to prevent later computation errors.

d) **Datetime Standardization:** All timestamps are converted to timezone-free UTC datetime to ensure consistency for merging and analysis.

e) **Final Output:** All tickers’ data are concatenated into a single DataFrame and exported as a clean CSV (data/raw/prices_6mo.csv).

3.3 Merging Reddit and Stock Data

The cleaned Reddit discussion data (socialmediadataclean_weekly_6mo.csv) and weekly stock price data (prices_6mo.csv) were merged through merge_for_model.py.

a) **Alignment:** Matched on (ticker, week_start) to synchronize sentiment with market activity.

b) **Feature Construction:** Calculated new variables, such as weekly_return, next_week_return, abnormal_volume, and standardized z_return.

c) **Output:** Produced socialmedia_price_panel_6mo.csv, the final modeling dataset containing both social and financial indicators, ready for data analysis and visualization.

4. Deviations from the Original Plan

During implementation, several challenges were encountered that required adjustments to the original proposal:

a) **Reddit API Access Restriction:** The official Reddit API denied access to my application request, and the Pullpush mirror failed to retrieve comments. To address this, I switched to the public JSON endpoint, which allowed successful data collection without authentication.

b) **Insufficient Data Volume:** The initial 3-month time window and two subreddits (r/stocks, r/wallstreetbets) produced limited data. To expand coverage, the window was extended to six months (2025-05-15 to 2025-11-15) and the scope was broadened to ten subreddits including r/investing, r/StockMarket, r/options, r/RobinHood, r/pennystocks, r/shortsqueeze, r/daytrading, and r/techstocks.

c) **Data Sparsity and Temporal Resolution:** Even after expansion, daily Reddit discussions remained too sparse for reliable analysis. Therefore, the data were aggregated from daily to weekly frequency, which improved stability and enabled consistent alignment with weekly stock price data for analysis and modeling.

5. Data Analysis

5.1 Hypothesis/Premise

a) Suppose the sentiment (positive/negative) of social media is a leading indicator of price changes, that is, positive discussions about a specific stock can predict the return rate of the next day.

b) Assume that fluctuations in the volume of discussions will show a significant positive correlation with abnormal trading volumes.

5.2 Analysis Techniques

We adopted the following four main quantitative analysis methods：

a) **Descriptive Statistics:** Use .describe() method to generate descriptive_statistics.xlsx to provide an overview of the entire dataset. Calculated the mean, standard deviation, minimum and maximum values of key indicators, such as 'next_week_return' and etc. This is helpful for understanding the distribution characteristics of the data.

b) **Correlation Analysis:** Calculate the Pearson Correlation Coefficient in the correlation_matrices.xlsx. Among them, matrix Sentiment_Returns_Corr shows the relationship between the sentiment indicator and the return rate of the next week. Matrix Discussion_Volume_Corr calculates the relationship between discussion degree and weekly trading volume.

c) **Hypothesis Testing (T-Test):** Output the two-sample t-test results in hypothesis_testing.txt. The samples were divided into "high sentiment week" and "low sentiment week" to test whether there was a statistically significant difference in the average rate of return between the two groups.

d) **Strategy Simulation:** The results of the simulation decision are saved in strategy_simulation.txt. Simulate the situation where stocks are only held during a "high sentimental" period and compare it with the average return rate during all sample periods.

5.3 Findings

a) **Sentiment vs. Returns:** Statistically, the correlation is extremely weak and it lacks significant predictive ability, but the strategy back testing shows a potential increase in average returns.

   1) **The correlation is extremely low.** The correlation between the sentiment indicator and the return rate for the next week is very close to zero. Among them, the correlation coefficient between sentiment and next_week_return is only 0.0097.
   2) **The differences are not significant.** The t-test results of hypothesis_testing.txt show that the p-value is 0.3163, which is greater than 0.05, and the null hypothesis cannot be rejected. The difference in returns between high weeks and low weeks was not statistically significant.
   3) **The strategic level.** strategy_simulation.txt shows that the average weekly return rate of the "high sentiment week" (1.91%) is higher than the overall average (1.14%), increasing by approximately 0.78%. This implies the sentiment factor has brought some excess returns by capturing volatility in specific individual stocks or extreme situations.

b) **Discussion Volume vs. Abnormal Volume:** There is a positive correlation, indicating that the discussion heat is associated with the market transaction activity. According to Discussion_Volume_Corr.sheet, the correlation coefficient is 0.1502. This indicates that high attention on social media does indeed tend to be accompanied by abnormal trading activities.

6. Data Visualization

6.1 Chart description

a) **Figure 1 - Correlation Analysis Heatmaps:** It contains two subgraphs to visualize the correlation matrix.

b) **Figure 2 - Stock Price vs. Social Sentiment Trends:** A dual-axis line graph for 10 different Tickers. The left axis represents stock prices, and the right axis represents weighted sentiment scores. If social sentiment could predict stock prices, we should observe the peak of the orange line followed by the rise of the blue line.

c) **Figure 3 - Sentiment Score vs. Next Week Returns:** Combining Violin Plot and Boxplot, the samples were classified into three categories: "negative", "neutral", and "positive", with the Y-axis representing the return rate for the next week.

d) **Figure 4 - Sentiment Distribution Analysis:** Figure 4A is the overall distribution histogram of all data, while Figure 4B shows the distribution by every ticker. The red line represents the mean value, and the dotted line represents the emotion score =0.

e) **Figure 5 - Online Discussion Volume vs. Abnormal Trading Volume:** The dual-axis graph is the same as Figure 3, with the left axis representing the number of mentions per week and the right axis representing the weekly abnormal trading volume.

f) **Figure 6 - Discussion Volume vs. Trading Volume by Ticker:** The scatter plot is accompanied by a red linear regression trend line, with the X-axis representing the number of mentions and the Y-axis representing the total transaction volume.

6.2 Observations and Conclusion

a) **Sentiment vs. Returns:** Positive or negative discussions cannot effectively predict the stock price return the next day.

1) Figure 1 shows that the sentiment indicators in the heat map have almost no linear relationship with the return rate for the next week.
2) In Figure 2, the sentiment curve fluctuates sharply without a significant response when the stock price curve fluctuates.
3) In Figure 3, regardless of whether it is positive, neutral or negative emotions, the mean difference of the return rate for the next week is very small, and the distribution shape is almost the same.
4) In Figure 4, the sentiment mean of most tickers is significantly greater than 0. Because users are more inclined to discuss the stocks they are optimistic about, this makes relying solely on "positive emotions" as a buy signal unreliable.

b) **Discussion Volume vs. Trading Volume:** The higher the attention, the higher the turnover rate and trading activity tend to be.

1) Figure 1 discusses the heat map of the volume and abnormal trading volume, showing a weak to moderate positive correlation.
2) In Figure 5, it can be clearly seen that the peak of the discussion volume often coincides with the peak of the abnormal trading volume in time.
3) In Figure 6, the regression lines of most stocks show an obvious upward trend, especially for popular stocks such as TSLA and NVDA.

6.3 Impact of findings

a) **Trading Strategy:** Avoid timing based solely on emotions. Data indicates that there is no statistically significant relationship between sentiment on social platforms and future returns.

b) **Risk Management:** An exceptionally high volume of social media discussions can serve as a warning signal. When the discussion volume of a stock suddenly deviates from the benchmark, it indicates that the stock is in a liquidity event, and the risk control model should reduce the leverage ratio of this asset.

7. Future Work

a) **Cross-Platform Data Integration**: Expanding beyond Reddit, future research could incorporate investor discussions from other platforms such as Twitter (X), StockTwits, and financial news comment sections. Merging these data sources would provide a more comprehensive view of market sentiment and strengthen predictive modeling.

b) **Advanced NLP Models**: Replace the VADER sentiment analyzer with transformer-based models such as FinBERT or RoBERTa fine-tuned for financial text to capture nuanced investor tone.

c) **Alternative Sentiment Metrics:** Explore non-linear relationships and interaction effects between sentiment, discussion volume, and returns using machine learning methods such as random forests or gradient boosting.