Analyzing Stock Price Movement Based on Current Sentiment

CSCI 185 - Fall 2025
Jonah E and Dylan S
Github Repo

1. Introduction

This project investigates how investor sentiment on Reddit's r/wallstreetbets relates to short-term stock price movements for a small set of highly discussed stocks, as well as to the market as a whole through two index funds. Using a Kaggle dataset of Reddit posts, we cleaned and merged textual data with daily market returns for GME, AMC, NOK, SPY, and QQQ. We applied sentiment analysis, topic modeling with Latent Dirichlet Allocation (LDA), Association rule mining (Apriori), as well as other topics from class like regex in the preprocessing stage to uncover patterns that link reddit conversations to price changes in the stock market.

The concepts we used:

- Sentiment Analysis: We used a transformer-based model (Twitter RoBERTa) to assign a sentiment score as well as a discrete label (negative/positive/neutral) to each reddit post
- Topic Modeling (LDA): To group posts into latent topics
- Regex: To identify which tickers are being discussed in the dataset and how frequently
- Association Rule Mining (Apriori): To discover relationships between discretized sentiment and topic indicators and stock price movements
- Time Series Analysis: Correlations and Granger causality tests between sentiment and stock returns

2. Dataset

The Reddit data used in this study was obtained from a publicly available Kaggle dataset of posts from the subreddit r/WallStreetBets:
https://www.kaggle.com/datasets/gpreda/reddit-wallstreetsbets-posts

The dataset contains Reddit posts exclusively from 2021, a period during which retail investor activity surrounding stocks such as GameStop (GME), AMC Entertainment (AMC), and Nokia (NOK) was especially prominent. From this dataset, we used the post title and text

content, the post score (upvotes), and the post timestamp. The text data provided the basis for sentiment analysis, while the score was used to approximate post visibility and relative importance. Timestamps enabled aggregation of sentiment at the daily level.

Historical stock market data was collected using the open-source Python library yfinance, which sources data from Yahoo Finance. We retrieved daily closing prices for GME, AMC, and NOK, along with two broad market index exchange-traded funds: SPY (S&P 500) and QQQ (NASDAQ-100). From these price series, we computed daily returns for each security.

In addition to individual stock returns, two composite measures were constructed: a retail basket, defined as the equal-weighted average of GME, AMC, and NOK daily returns, and a market index composite, defined as the equal-weighted average of SPY and QQQ daily returns. These series allow for direct comparison between Reddit sentiment, retail-focused stocks, and the broader market.

3. Data Preprocessing

For the Reddit data, the post title and main text were first concatenated into a single text field to ensure that all available linguistic content contributed to sentiment estimation. The combined text was then cleaned by removing URLs, punctuation, emojis, and non-alphanumeric symbols, retaining only word characters and whitespace. Finally, posts containing fewer than five words after cleaning were removed to eliminate low-information entries and reduce noise.

For the stock market data, historical price data was filtered to retain only daily closing prices for each security. These prices were converted into daily returns using percent change to normalize price movements across assets. Two composite return series were then constructed: a retail basket representing the equal-weighted average returns of GME, AMC, and NOK, and a market index composite representing the equal-weighted average returns of SPY and QQQ.

To support time-series analysis, stock returns were aligned with daily Reddit sentiment by date. In addition, lagged ($t-1$ to $t-3$) and lead ($t+1$ to $t+3$) return features were generated for each stock and composite index. Rows containing missing values introduced by lag and lead construction were removed to ensure consistent temporal alignment across all variables.

4. Ticker Extraction

- In this section, we wanted to understand the scope of tickers mentioned within the WSB dataset. The dataset mentions that GME, AMC, NOK, are the most commonly used tickers in the reddit dataset. In this step we extract all of the commonly mentioned

tickers, to confirm this behavior and look into potentially other frequently mentioned tickers. The goal is to compare against the most common tickers, and the market as a whole, so it is important to understand the scope of stocks being mentioned in the dataset.

- We defined two regex patterns:

```
# $GME, $AMC style
ticker_pattern_dollar = re.compile(r"\$[A-Z]{1,5}\b")

# Bare tickers like GME, AMC (2–5 capital letters)
ticker_pattern_plain = re.compile(r"\b[A-Z]{2,5}\b")
```

- For each raw_text:
    - Extracted matches from both patterns
    - Stripped leading $
    - Deduplicated within posts
- We also made a custom set of non-tickers that were frequently mentioned in all caps and weren't tickers, ex) "CEO" or "YOLO"
- We combined this with capitalized english stopwords as commenters in this forum frequently spoke in all caps.
- Here are the findings as well as frequencies:

GME	12635
AMC	4812
HOLD	2522
BB	2045
BUY	1673
NOK	1520
RH	1379
TLDR	1314
DR	922
TL	920
MOON	908
SELL	823
EDIT	757
PLTR	725
RKT	691
IPO	632
STOCK	620
FUCK	602
LIKE	571
LINE	569

    - While there are still non-tickers in here, it is still enough information to decide if we want to expand our analysis: We can see that GME (~12.6k), AMC (~4.8k), and NOK (~1.5k) are the most talked about which confirms what the dataset said. Other tickers are mentioned in the dataset like BB, PLTR, RKT, but we chose to just analyze the aforementioned three for our downstream tasks as we need enough observations per ticker for continuous daily comparison. The top three tickers (GME, AMC, NOK) are the only ones with a large enough sample size to support meaningful comparisons.

5. Data Analysis

In this section, we analyze Reddit sentiment, market relationships, discussion topics, and rule-based associations to examine how online behavior aligns with stock market dynamics.

5.1 Sentiment Analysis

- Each Reddit post was scored using the twitter-roberta-base-sentiment-latest model, a RoBERTa-based classifier trained on 124 million tweets and fine-tuned for three-class sentiment prediction (negative, neutral, positive).
- This model was selected because it is optimized for informal online language, making it well-suited for Reddit posts containing slang and profanity.
- Posts were tokenized and processed in batches of 32, with inputs truncated to a maximum of 128 tokens, and inference was run in evaluation mode using torch.no_grad().
- For each post, a continuous sentiment score was computed as the difference between the positive and negative class probabilities, producing values bounded between −1 and 1. Each post was also assigned a discrete sentiment label: positive, negative, or neutral.
- A sample of the resulting sentiment outputs is shown below to illustrate the range of scores and corresponding class labels:

	clean_text	sentiment_score	sentiment_label
0	it s not about the money it s about sending a ...	-0.236250	neutral
1	math professor scott steiner says the numbers ...	-0.775493	negative
2	exit the system the ceo of nasdaq pushed to ha...	-0.601432	negative
3	new sec filing for gme can someone less retard...	-0.424591	negative
4	not to distract from gme just thought our amc ...	-0.189812	neutral

5.2 Comparison with Market

- We computed the Pearson correlation matrix across all numeric variables, including daily sentiment, individual stock returns (GME, AMC, NOK), the retail basket, the market index, and all lagged and lead return features.
- The correlation column corresponding to daily_weighted_sentiment was extracted to measure linear relationships between sentiment and same-day, lagged, and future stock returns.
- To assess directional temporal dependence, we applied Granger causality tests using the statsmodels framework in both directions: sentiment → return and return → sentiment.
- Granger tests were conducted using lag orders of 1 to 3 trading days, with statistical significance evaluated using the SSR F-test.
- All causality tests were executed programmatically across GME, AMC, NOK, SPY, QQQ, the retail basket, and the market index to ensure consistent evaluation across retail and market-wide assets.

5.3 Topic Modeling

- We used Topic modeling in this project to understand ‘what’ people are talking about to supplement the ‘how’ that we understand from sentiment analysis.
- To vectorize the text we used sklearns’ CountVectorizer:
-

```python
vectorizer = CountVectorizer(
    max_df=0.8,         # drop words in >80% of docs
    min_df=30,          # require at least 30 docs
    stop_words=list[str](stop_words),
    token_pattern=r"(?u)\b[a-z]{3,}\b"  # only alphabetic words length>=3
)
```

- After multiple iterations, junk words were identified, and the number of topics was reduced to 8 from 10 to increase the interpretability and distinctness of the learned topics
- Here are the learned Topics: our interpretation, and the top 15 words in each
    - Topic 0: Broker/Platform

```
Topic #0: robinhood | trading | account | market | silver | gme | stocks | trade | buy | app | money | sec | use | brokers | fidelity
```

    - Topic 1: WallstreetBets / community chatter

```
Topic #1: wsb | post | rkt | good | don | apes | time | people | rocket | sub | going | ape | loss | think | day
```

    - Topic 2 & 3: Fundamental terms and company/earnings talk

```
Topic #2: company | market | companies | new | growth | year | years | business | revenue | industry | stock | think | tesla | world | going
Topic #3: company | million | stock | year | shares | share | market | revenue | price | billion | earnings | growth | cash | value | quarter
```

    - Topic 4: Short Squeeze narrative

```
Topic #4: short | shares | price | gme | stock | squeeze | market | shorts | people | hedge | options | buy | don | sell | money
```

    - Topic 5: Meme/YOLO “buy and hold”

```
Topic #5: gme | buy | hold | amc | fucking | don | money | sell | holding | fuck | going | moon | bought | stock | shares
```

    - Topic 6: General Market/technical discussion

```
Topic #6: market | price | stock | earnings | week | time | stocks | year | inflation | chart | prices | day | going | high | look
```

    - Topic 7: daily/weekly discussion and social media

```
Topic #7: best | daily | weekly | discussion | thread | wall | street | robinhood | yolo | trading | follow | wsb | people | make | twitter
```

5.4 Association Rule Mining

- Building the Association Rule mining Dataset:
    - We took the dataset that combined Sentiment scores, Topics weights, and the daily returns from the 5 stocks as well as the 2 aggregated return bags for this task, into daily averaged items.
    - We computed daily average topics weights, and normalized timestamps.
    - For Apriori we needed to transform our continuous numerical data into boolean items
    - Sentiment
        - We used 33rd and 66th percentiles to define: sentiment_low, sentiment_medium, and sentiment_high
    - Same-day price movements for GME, AMC, NOK, SPY, QQQ, retail_basket, market_index:

    - X_up = (return > 0), X_down = (return < 0)
- Topics:
    - For each topic column topic_k, we computed its median and defined:
        - Topic_k_high = (topic_k > median)
- Apriori Implementation:

```
# Build transaction matrix and run Apriori

item_cols = [c for c in df_arm.columns if df_arm[c].dtype == bool]
transactions = df_arm[item_cols].astype(int)
print("Transactions shape:", transactions.shape)

# Mine frequent itemsets
freq_itemsets = apriori(transactions, min_support=0.10, use_colnames=True)
print(f"Found {len(freq_itemsets)} frequent itemsets (support >= 0.10)")

# Generate association rules
rules = association_rules(freq_itemsets, metric="confidence", min_threshold=0.6)
print(f"Generated {len(rules)} rules (confidence >= 0.6)")

rules.head()
```

    - Min_support = .10
    - Min_confidence = .60
- We then filtered antecedent and consequent variables
    - Antecedents were filtered to contain only the sentiment and topic features for each ‘day’ item
    - Consequents were filtered to contain at least 1 ‘movement_item’ meaning that a market up or down finding had to be included in the consequent for the rule to be mentioned

6. Final Results

- 10 strongest Pearson correlations between daily Reddit sentiment and stock returns:

```
AMC_lag3                0.289172
retail_basket_lag3      0.243836
retail_basket_lag1      0.221257
AMC_lag1                0.203723
NOK_lag1                0.193182
NOK_lead2               0.184068
NOK_lag3                0.175956
GME_lag1                0.171632
GME_lag3                0.158831
NOK                     0.139365
```

- Granger causality test p-values for bidirectional relationships between Reddit sentiment and stock returns ($p < 0.05$ is considered significant):

Predictor → Target	Lag 1 p-value	Lag 2 p-value	Lag 3 p-value
Sentiment → GME	0.7677	0.8465	0.7097
GME → Sentiment	**0.0295**	0.1252	0.1688
Sentiment → AMC	0.5897	0.6329	0.7538
AMC → Sentiment	0.0516	0.0997	**0.0063**
Sentiment → NOK	0.1393	0.1225	0.2701
NOK → Sentiment	**0.0305**	0.1365	0.1605
Sentiment → SPY	0.8750	0.9749	0.9378
SPY → Sentiment	0.4360	0.6948	0.5906
Sentiment → QQQ	0.3786	0.6489	0.7037
QQQ → Sentiment	0.6332	0.7721	0.6710
Sentiment → Retail Basket	0.7825	0.7078	0.6871
Retail Basket → Sentiment	**0.0131**	**0.0423**	**0.0216**
Sentiment → Market Index	0.5514	0.8034	0.8059
Market Index → Sentiment	0.5349	0.7385	0.6232

- Top 10 Apriorio Rules sorted by lift:

	antecedents	consequents	support	confidence	lift
10422	(sentiment_low, topic_0_high, topic_1_high)	(QQQ_down, topic_4_high, market_index_down)	0.103448	0.631579	3.052632
5755	(sentiment_low, topic_4_high, topic_1_high)	(topic_0_high, SPY_down)	0.103448	0.631579	2.930526
5809	(sentiment_low, topic_0_high, topic_1_high)	(QQQ_down, topic_4_high)	0.103448	0.631579	2.817814
10427	(sentiment_low, topic_4_high, topic_1_high)	(topic_0_high, QQQ_down, market_index_down)	0.103448	0.631579	2.817814
5821	(sentiment_low, topic_0_high, topic_1_high)	(market_index_down, topic_4_high)	0.103448	0.631579	2.817814
5753	(sentiment_low, topic_0_high, topic_1_high)	(topic_4_high, SPY_down)	0.103448	0.631579	2.817814
5823	(sentiment_low, topic_4_high, topic_1_high)	(topic_0_high, market_index_down)	0.103448	0.631579	2.616541
5812	(sentiment_low, topic_4_high, topic_1_high)	(topic_0_high, QQQ_down)	0.103448	0.631579	2.616541
5662	(sentiment_low, topic_0_high, topic_1_high)	(topic_4_high, NOK_down)	0.103448	0.631579	2.526316
5664	(sentiment_low, topic_4_high, topic_1_high)	(topic_0_high, NOK_down)	0.103448	0.631579	2.526316

7. Discussion

This section synthesizes the correlation, causality, and association-rule findings to interpret how Reddit sentiment relates to market behavior.

Correlation and Granger Causality:

Across both the Pearson correlation analysis and Granger causality tests, a consistent pattern emerges in which market movements precede changes in Reddit sentiment rather than the reverse. The strongest correlations occur between sentiment and lagged returns, particularly for AMC_lag3 ($r = 0.289$), retail_basket_lag3 ($r = 0.244$), and retail_basket_lag1 ($r = 0.221$). Similar lagged relationships are observed for GME, AMC, and NOK, indicating that Reddit sentiment is most strongly associated with prior stock price behavior, consistent with reactive investor discussion following major price moves.

The Granger causality results further reinforce this interpretation. For multiple retail-focused assets, including AMC, GME, NOK, and the retail basket, price returns Granger-cause sentiment, while sentiment does not Granger-cause returns at any tested lag. In contrast, no statistically significant causal relationships are observed between sentiment and broader market indices. Together, these findings indicate that WallStreetBets sentiment primarily functions as a response signal to observed market volatility, rather than a reliable leading indicator of near-term price movements.

The presence of limited lead correlation for NOK suggests that isolated predictive relationships may occasionally emerge, but these effects are comparatively weak and inconsistent. Overall, the results support the conclusion that online retail investor sentiment during this period is predominantly market-reactive in structure.

Apriori:

Looking at the top 10 rules, a clear pattern is that low daily sentiment combined with topic 0 (broker/platform) and topic 4 (short squeeze narrative), often together with Topic 1 (WSB community talk), is strongly associated with broad market down days.

For example, rules such as (sentiment_low, topic_0_high, topic_1_high) → (QQQ_down, topic_4_high, market_index_down) or (sentiment_low, topic_4_high, topic_1_high) → (topic_0_high, SPY_down) all have confidence around ~0.63 and lift around ~3, meaning that on days with this combination of sentiment and topics, market-wide down moves are roughly three times more likely than on an average day. Overall these rules support the interpretation that panic

and upset brokers + talks about short squeeze on WSB correlates with risk off days, rather than with market rallies.

Note: These rules are just descriptive and they don't establish causality. Because our support is relatively low, and our sample size is small, these patterns are just exploratory evidence of how certain conversation themes could co-occur with price moves. They would not be reliable trading signals.

8. Conclusion

This project examined the relationship between Reddit sentiment on r/wallstreetbets and short-term stock market behavior using sentiment analysis, topic modeling, association rule mining, and time series methods. By integrating social media text data with daily stock returns, we analyzed both contemporaneous and directional relationships between online discussion and market movements. The results show that Reddit sentiment is more strongly associated with prior price movements than with future returns, supporting a predominantly reactive structure. Topic modeling and association rules further revealed that specific discussion themes tend to co-occur with broader market conditions. Overall, this study highlights the value of combining natural language processing and financial time series analysis to better understand the interaction between online investor behavior and market dynamics.

9. Limitations

Our analysis has several important limitations that should be mentioned:

- The daily sample is not very large, many of the statistical and Apriori patterns are noisy meaning that they may not generalize to periods outside of the time that this dataset had recorded (2021).
- The r/wallstreetbets has a very specific subset of traders that use a lot of memes, sarcasm, and slang within their speech that could confuse sentiment models and limit how representative the sentiment is in broader markets.
- In our Apriori implementations, making the continuous variables into binary items, gets rid of information and makes the rules sensitive to threshold choices.
- Meaningful text content is really only focused on just three main stocks (which could be considered meme stocks), as a result, our findings are really about a narrow set of stocks during a specific time period, so generalizing to the entire market in our comparison was ambitious.
- Association rules and Granger tests reveal correlation, they do not reveal causation. This means that none of our results should be interpreted as trading signals.