Project Report: Leveraging Reddit Sentiment for Market Analysis (DS464)

Group Members:

Muhammad Afnan Shahid (2022333)
Muhammad Anns Khan (2022348)

1. Introduction

The objective of this project is to conduct a comprehensive analysis of social media discourse—specifically Reddit—to evaluate its effectiveness as a leading or lagging indicator of financial market movements. In recent years, Reddit has emerged as one of the most influential platforms for retail investors, fostering communities that actively discuss market trends, investment strategies, and speculative opportunities. The unprecedented impact of forums such as *r/wallstreetbets* on "meme stocks" and broader indices underscores the significance of crowd-driven sentiment in shaping liquidity and volatility. Consequently, understanding Reddit sentiment has become essential for assessing its potential predictive value in financial markets.
This study employs the S&P 500 (SPY) as a benchmark and adopts a structured approach built on three key dimensions:

1. **Sentiment Mapping:** This involves tracking sentiment fluctuations of posts on subreddits across 2024 and mapping it to market conditions to determine whether retail sentiment precedes or follows major price movements.
2. **Sentiment Comparison:** This parameter contrasts sentiment trends between speculative communities such as *r/wallstreetbets* and conservative forums like *r/dividends*. The comparison highlights how risk appetite influences sentiment volatility and its correlation with market performance.
3. **User Accuracy:** Beyond aggregate sentiment, this component identifies individual contributors who post often and whose expressed views consistently align with subsequent market returns, offering potential predictive insights.

Through this holistic framework, the project aims to provide a rigorous evaluation of Reddit's role within the financial ecosystem and its viability as a tool for market forecasting.

2. Methodology

The methodology for this project is structured into a rigorous three-stage pipeline: data acquisition, data cleaning, and multifaceted sentiment analysis. The technical implementation utilizes a suite of Python libraries, including Pandas and NumPy for data manipulation, PyTorch and Transformers for deep learning, yFinance for market data retrieval, and SciPy for statistical validation.

2.1 Data Acquisition and Storage

The data acquisition process utilizes the PullPush API to circumvent the historical limitations of the standard Reddit API. PullPush was designed specifically to enhance search over Reddit content and supports rich query parameters (time bounds, fields, sorting, aggregations), enabling efficient retrieval of large, date-bounded corpora for research workflows. In light of Reddit's Data API terms and rate-limit enforcement—particularly the free-tier limits and OAuth requirements—we selected PullPush for historical breadth and lower operational overhead during acquisition, while remaining cognizant of Reddit's policies on permitted access and usage. Benchmark market returns were sourced via yfinance, a Python library that provides programmatic access to Yahoo Finance's publicly available data endpoints.

The system is designed to fetch submissions from targeted subreddits, such as r/wallstreetbets and r/dividends, for the 2024 calendar year. To ensure data integrity, the scraper calculates UNIX timestamps for each ISO week and executes paginated requests. The logic filters for high-engagement posts by sorting the collected pool by upvotes, ensuring a high-quality buffer for analysis. Data is stored locally in structured JSON files, categorized by subreddit and week, containing key fields such as author, title, body, score (upvotes), and number of comments.

2.2 Data Cleaning and Preprocessing

We normalized text by lowercasing, punctuation and formatting removal ensuring uniformity across all posts and valid input for the sentiment analysis model. We also filtered posts flagged as [deleted] or [removed] markers widely recognized within Reddit's ecosystem to denote user deletion or moderator removal, respectively—so that the analysis reflects content accessible to readers at the time of discussion. Moreover, since finance discussions frequently reference instruments via tickers (e.g., $TSLA, $SPY), we extracted them using regular expressions that identify the dollar-prefixed token and uppercase tickers.

2.3 Sentiment Calculation and Statistical Analysis

The core analysis is driven by utilizing FinBERT (ProsusAI/finbert), a pre-trained BERT model fine-tuned specifically for financial language. We synchronized Reddit sentiment aggregates to trading days, merging on New York market calendars and handling weekends/holidays through forward-fill logic to ensure valid return comparisons for SPY

Sentiment is calculated as a scalar score derived from the difference between the probability of a text being "positive" and "negative" (). We assessed the relationship between aggregated Reddit sentiment and SPY returns using the Pearson correlation coefficient (lag -1, lag 0, lag +1). Pearson's *r* quantifies linear association on [-1, 1], with two-sided p-values testing the null of zero correlation; this choice aligns with our objective to detect linear co-movement across daily horizons

To account for social influence, sentiment is then weighted by engagement using the following formula:

$$\text{Weighted sentiment} = \text{sentiment} \times (1 + \log(\text{upvotes} + \text{comments}))$$

This ensures that highly visible community discussions have a greater impact on the aggregate metric.

To analyze sentiment across different subreddits, posts were first grouped by community and aggregated over defined time intervals. This grouping allowed us to capture the collective tone of discussions within each subreddit, reflecting distinct investment philosophies bullish versus bearish. For each group, individual sentiment scores derived from FinBERT were averaged to compute a mean sentiment value, ensuring that extreme opinions did not disproportionately influence results. Comparing these averages across subreddits provided insights into how risk appetite and community behavior correlate with market trends, revealing whether certain forums exhibit heightened optimism or caution during volatile periods.

At the contributor level, “User Alpha” was computed by correlating an individual poster’s weighted sentiment time series with next-day returns, surfacing “oracle” users (high positive *r*) and “inverse” indicators (consistently negative *r*). While exploratory, this design mirrors factor discovery workflows and relies on the same inferential backbone as the aggregate analysis

3. Data Analysis

Our analysis integrates two primary data sources: (i) Reddit submissions collected via the PullPush interface and (ii) market data for SPY retrieved programmatically from Yahoo Finance using yfinance. Reddit records were ingested as structured JSON to enable quick and efficient data storage. We favored the PullPush search endpoints (submission/comment) because they allow precise, date-bounded queries, field selection, sorting, and aggregations suitable for research workflows. We pulled around 125000 reddit posts from the year 2024 which proved to be a significantly large enough sample. After preprocessing and cleaning we were left with around 80000 posts.

3.1 Data Points (JSON Schema)

The raw data scraped includes:

- **ID:** Unique Reddit post identifier.
- **Title/Body:** The primary text analyzed by the FinBERT model.
- **Upvotes/Num_Comments:** Engagement metrics used for weighting.
- **Timestamp_UTC:** Converted into datetime objects to align with market trading days.
- **Author:** Tracked to perform the "User Accuracy" analysis.

These data points allowed us to perform the analysis we needed without affecting memory efficiency. The subreddit field is appended later while analyzing the raw Json file by extracting the subreddit name from the file name.

3.2 Subreddits Analyzed

This variety represents **risk heterogeneity**—from momentum-seeking speculation to conservative, fundamentals-driven approaches—and thus supports meaningful cross-community comparisons.

- **r/wallstreetbets:** High-volatility, sentiment-driven discussions focusing on short-term gains.
- **r/dividends:** Conservative, long-term investment discussions focused on yield and stability.
- **r/smallstreetbets:** A smaller, niche counterpart to *r/wallstreetbets*, often mirroring speculative strategies but with less reach.

- **r/investing:** Broad discussions on macroeconomic trends, portfolio strategies, and fundamental analysis.
- **r/stocks:** General equity market conversations, including earnings, valuations, and sector performance.
- **r/StockMarket:** A community centered on overall market conditions, indices, and major events.
- **r/StocksAndTrading:** Focused on active trading strategies, technical analysis, and short-term opportunities.
- **r/ValueInvesting:** Dedicated to fundamental valuation principles, emphasizing intrinsic value and long-term holding.

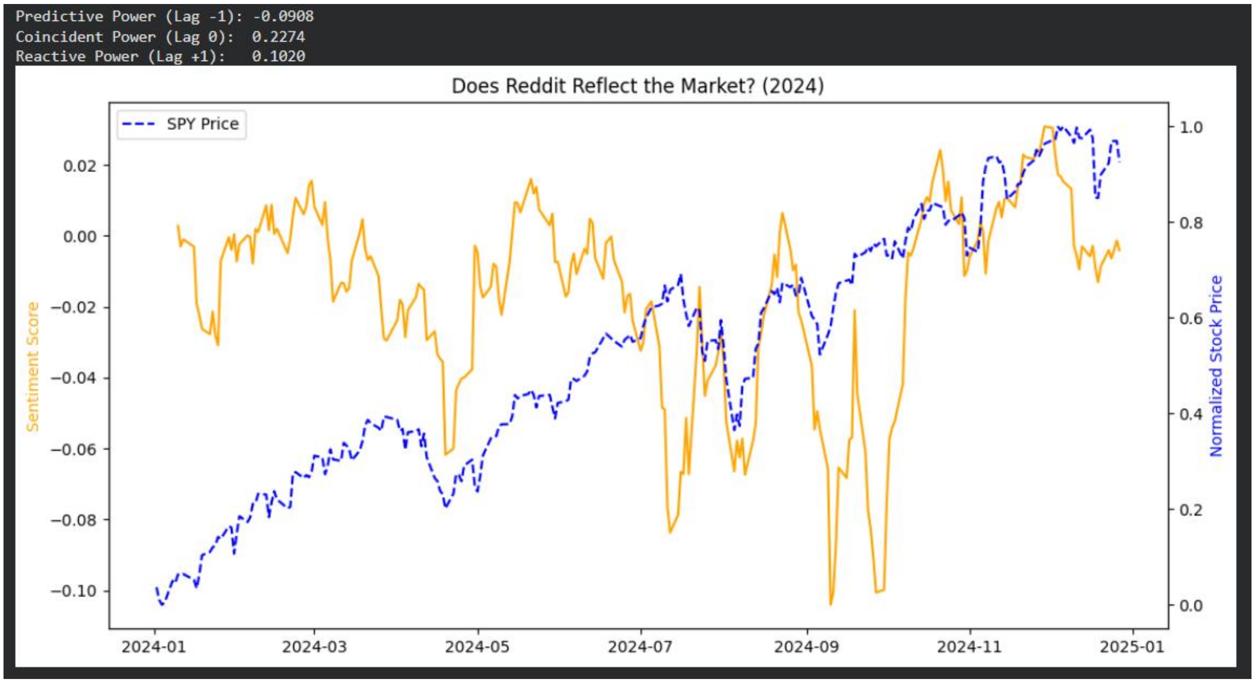
4. Results

The following section displays the results obtained on the analysis metrics described above.

4.1 Comparison Between reddit sentiment and market movements

The image below shows the comparison between post sentiment and normalized price of the S&P500 index. It also calculates the pearson correlation on three different parameters:

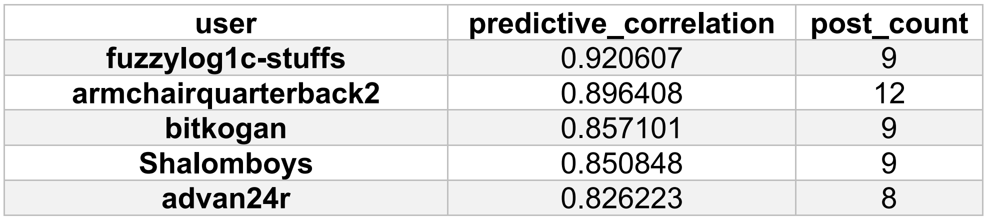
1. Predicitve: Reddit Sentiment (Day T-1) vs. Market Return (Day T)
2. Coincident: Reddit Sentiment (Day T) vs. Market Return (Day T)
3. Reactive: Market Return (Day T-1) vs. Reddit Sentiment (Day T)



```
Predictive Power (Lag -1): -0.0908
Coincident Power (Lag 0):  0.2274
Reactive Power (Lag +1):   0.1020
```

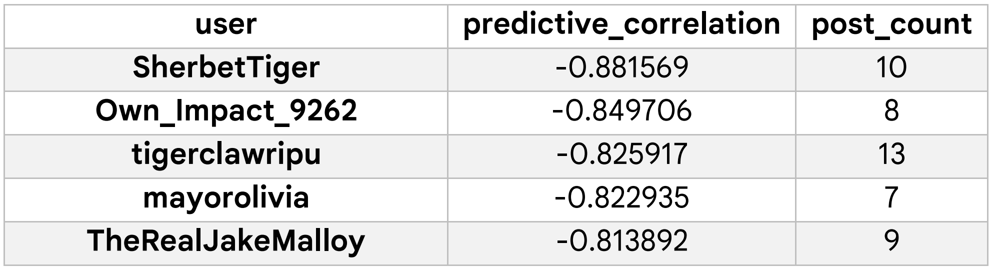
4.2 User Accuracy comparison

The first table lists the top “Oracle” users whose posts showed strong positive correlation with next-day market returns, suggesting predictive accuracy. Each entry includes username, correlation coefficient (up to 0.92), and post count, indicating consistent performance across multiple posts. The second table highlights “Inverse” users whose sentiment was negatively correlated with returns, meaning their predictions were consistently wrong. Correlations range from -0.81 to -0.88, with post counts between 7 and 13. Together, these tables illustrate that while overall crowd sentiment may lag, certain individuals exhibit strong directional signals—either highly accurate or reliably contrarian.



user	predictive_correlation	post_count
fuzzylog1c-stuffs	0.920607	9
armchairquarterback2	0.896408	12
bitkogan	0.857101	9
Shalomboys	0.850848	9
advan24r	0.826223	8

Table 1. Users with highest correlation

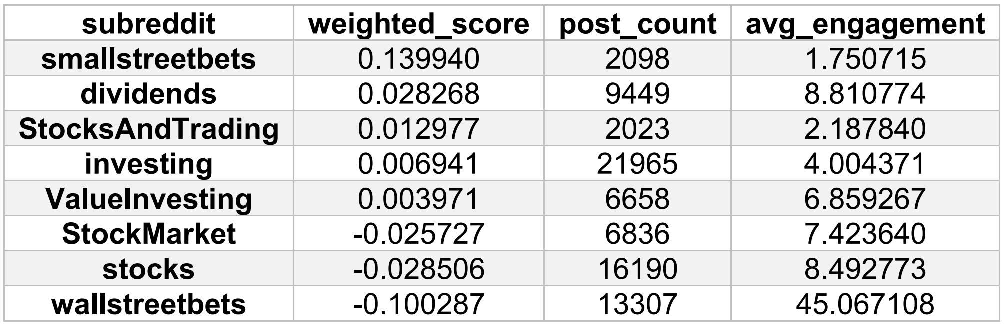


user	predictive_correlation	post_count
SherbetTiger	-0.881569	10
Own_Impact_9262	-0.849706	8
tigerclawripu	-0.825917	13
mayorolivia	-0.822935	7
TheRealJakeMalloy	-0.813892	9

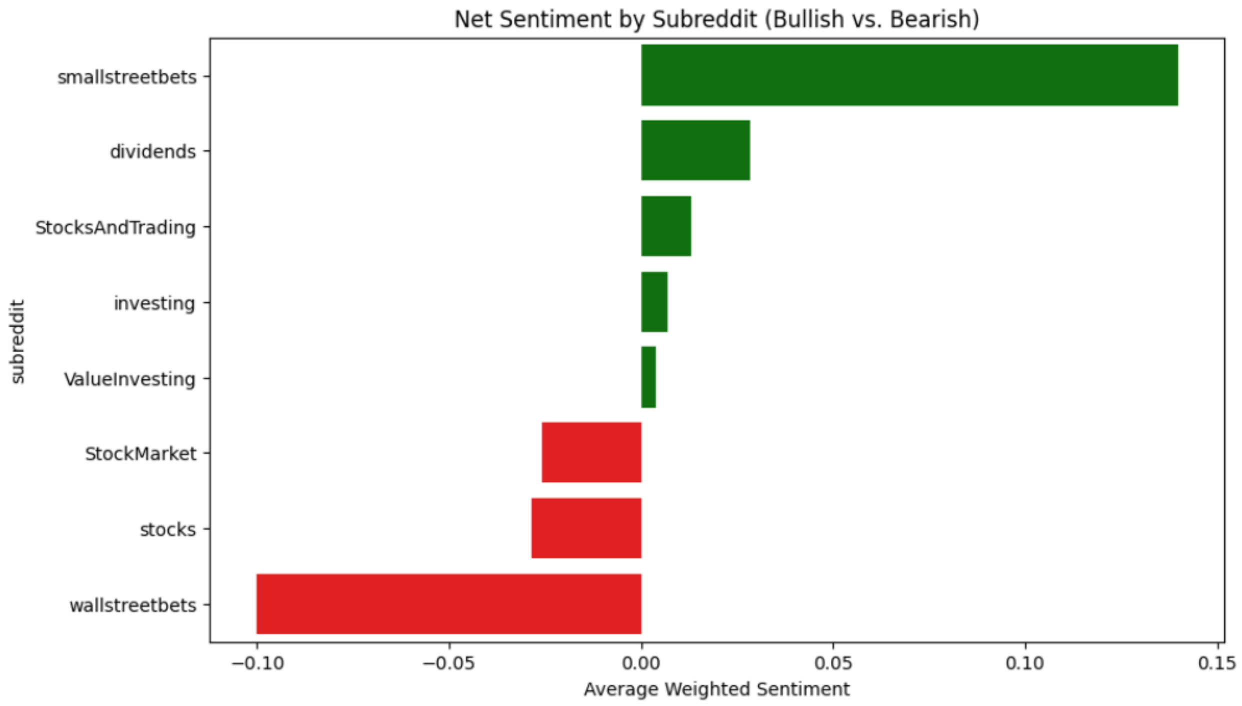
Table 2. Users with Lowest correlation

4.3 Subreddit sentiment comparison

This section includes the subreddit centric analysis. The table contains details on the weighted score and activity of each subreddit while the graph shows the average sentiment of the subreddit throughout the year.



subreddit	weighted_score	post_count	avg_engagement
smallstreetbets	0.139940	2098	1.750715
dividends	0.028268	9449	8.810774
StocksAndTrading	0.012977	2023	2.187840
investing	0.006941	21965	4.004371
ValueInvesting	0.003971	6658	6.859267
StockMarket	-0.025727	6836	7.423640
stocks	-0.028506	16190	8.492773
wallstreetbets	-0.100287	13307	45.067108



5. Conclusion

The results of our analysis reveal several important insights into the relationship between Reddit sentiment and financial market movements. First, the correlation study indicates that aggregated Reddit sentiment tends to function as a co-incident indicator rather than a lagging or predictive one. Sentiment spikes were observed primarily after significant market moves, suggesting that the general public react to news and price changes rather than anticipating them. This behavior aligns with the broader understanding of social media-driven trading, where information dissemination and emotional responses often follow major events rather than precede them.

The subreddit-level analysis further reinforces this interpretation. Communities such as r/wallstreetbets exhibited the most negative weighted sentiment score (-0.100287) despite having the highest average engagement (45.06) and a substantial post count (13,307). This combination suggests that speculative forums are highly reactive and emotionally charged, amplifying pessimism during downturns and optimism during rallies. Conversely, conservative communities like r/dividends and r/ValueInvesting maintained positive sentiment scores (0.028268 and 0.003971, respectively) with relatively stable engagement levels.

The identification of “Oracle” and “Inverse” users adds another layer of nuance. While crowd sentiment lacks predictive power, certain individuals demonstrated strong directional accuracy, with correlation coefficients as high as 0.92. These users may possess superior analytical skills, access to timely information, or disciplined strategies that enable them to anticipate market movements. Conversely, consistently wrong predictors—those with correlations near -0.88—

highlight the presence of contrarian signals that could be exploited for inverse trading strategies. This dichotomy suggests that while aggregate sentiment is noisy, micro-level patterns within user behavior can offer actionable insights.

6. Future Works

While the current study provides valuable insights into Reddit sentiment and its relationship with market movements, several avenues exist for extending this research.

6.1 Increasing the data collection scope: The scope of communities can be broadened significantly by including additional communities which would enhance coverage and allow for more robust sentiment analysis. Moreover, incorporating comment-level sentiment would provide a richer representation of community consensus, as replies often contain nuanced perspectives that differ from the original post. Beyond Reddit, integrating other social platforms such as Twitter—where financial discourse is highly active and often real-time—could provide complementary signals and improve predictive robustness. Twitter's cashtag-based conversations and influencer-driven narratives offer a distinct dynamic compared to Reddit's community-centric model.

6.2 Changing the time period: Our current framework operates primarily on daily sentiment aligned with trading days. Future iterations could aggregate sentiment over weekly or monthly horizons to capture long-term mood shifts and structural trends rather than short-term volatility. This approach would enable the detection of sustained optimism or pessimism within communities, offering insights into broader market cycles. Similarly, extending the analysis window from one year to multiple years would allow for the examination of sentiment behavior across different macroeconomic regimes, including bull and bear markets, interest rate cycles, and geopolitical events.

6.3 Specified Analysis: Reddit hosts numerous other forums dedicated to niche strategies, sector-specific discussions, and emerging markets. Creating a separate model for domain-specific or stock-specific data could enhance the accuracy of the analysis. Additionally, applying this methodology to ticker-specific sentiment rather than index-level aggregates could reveal stronger predictive relationships for individual equities.

Collectively, these enhancements would transform the current framework into a more comprehensive, multi-platform, and multi-horizon sentiment analysis system capable of delivering deeper insights into retail investor behavior and market dynamics.

7. Helpful Links:

Github Link: https://github.com/Afnan-sh/Reddit-Financial-Analyser/tree/main
PullPush API: https://pullpush.io/
FinBERT: https://huggingface.co/ProsusAI/finbert