Predicting Market Trends Using Social Media Sentiment Analysis

Samuel Adebayo

Guru Raghavendra Reddy Batchu

University of the Cumberlands

Supervisor: Dr. Eve Thullen

June 26, 2025

Abstract

This study explores the predictive power of social media sentiment analysis on short-term financial market movements. Social media platforms, particularly Reddit, significantly influence market dynamics through the rapid dissemination of retail investor sentiment. Leveraging sentiment analysis, technical indicators, and advanced machine learning techniques such as Random Forest and LSTM networks, we assessed Reddit-derived sentiment's predictive capabilities. Our comprehensive analysis revealed weak correlations between social sentiment scores and stock returns, primarily due to high volatility and sentiment noise. Enhanced modeling approaches incorporating technical indicators showed minor improvements, emphasizing the market predictions' multifaceted nature. Practical implications suggest caution in solely relying on sentiment analysis, recommending a combined approach with traditional market indicators. The study underscores the necessity of integrating broader macroeconomic indicators and detailed sentiment metadata for more accurate predictive results.

Part 1: Predicting Market Trends Using Social Media Sentiment Analysis

Introduction

The influence of social media on market behavior is a growing area of interest in financial analytics. Platforms like Reddit have demonstrated significant power in shaping market activity, particularly among retail investors. This project investigates whether sentiment extracted from Reddit posts can be used to predict short-term market movements. Our objective is to evaluate the predictive strength of sentiment signals and understand how they may complement traditional financial indicators.

Literature Review

Bollen et al. (2011) first demonstrated that Twitter sentiment could have predictive power over stock market trends. FinBERT, a transformer-based sentiment model (Jiang et al., 2021), has shown promise in domain-specific sentiment analysis. Yet, multiple studies caution against relying solely on sentiment due to data noise and volatility (Sohangir et al., 2018; Xing et al., 2022). Our study builds on these findings but focuses on Reddit, which reflects unique retail investor behavior and has received less emphasis in prior studies.

Methodology

Data Collection and Preprocessing

Reddit data was collected using PRAW and Kaggle sources, focusing on popular subreddits like r/wallstreetbets. Stock data was sourced from Yahoo Finance. Data from 2019 to 2025 was used. Redundant and missing values were cleaned. Sentiment scores were generated using VADER and enhanced with FinBERT. Aggregation by ticker and date aligned the sentiment data with stock returns.

Feature Engineering

Lagged features (e.g., sentiment_lag_1) were created. Technical indicators such as RSI, MACD, VIX, MA_10, and EMA_10 were computed to improve predictive accuracy.

Modeling Techniques

Initial models included Linear Regression and Random Forest. A more advanced LSTM model was later implemented to capture sequential patterns.

Evaluation Metrics

Models were evaluated using R^2, MAE, RMSE, and classification accuracy for directional predictions.

Results

Correlation analysis showed weak links between sentiment and stock returns (e.g., sentiment_lag_1 correlation = 0.0099). RSI showed the strongest correlation at 0.2533.

Model Results

- Linear Regression: R^2 = -0.0049, MAE = 0.0367
- Random Forest: R^2 = -0.0277, MAE = 0.0372
- LSTM (regression): MAE = 0.0365
- LSTM (classification): Accuracy = 49.55%

Sentiment alone was not a strong predictor. The inclusion of technical indicators improved results only slightly.

Discussion

Reddit sentiment, while accessible and timely, represents mainly retail investor views. Market-moving actions often come from institutional investors, whose sentiment is not captured here. Moreover, high volatility and noise in Reddit data dilute its predictive strength.

Technical indicators consistently outperformed sentiment-based features. A hybrid approach that combines multiple signal types, including macroeconomic indicators, may yield better results.

***Limitations include*:**

Sentiment analysis based on Reddit predominantly captures retail investor sentiment, omitting significant institutional investor insights and external economic influences. Institutional investors substantially drive market movements through larger trades; thus, their exclusion likely contributes significantly to the weak predictive signals observed.

Implications suggest that sentiment analysis should serve as a supplementary input, not a standalone forecasting method.

Implications:

- Investors and analysts should leverage sentiment analysis cautiously, using it as a supplemental rather than primary analytical tool.
- Technical indicators like RSI should form the core of predictive modeling efforts due to their stronger correlations.

Conclusion

Sentiment analysis can complement but not replace traditional financial indicators for market prediction. Reddit-based sentiment shows minimal correlation with returns and performs poorly in isolation. The study recommends that future models integrate broader economic signals and event-focused sentiment analysis.

References

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*.

Jiang, M., Cui, H., & Deng, S. (2021). Using FinBERT for financial sentiment analysis. *Journal of Financial Data Science, 3*(2), 75-93.

Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*

Xing, F. Z., Cambria, E., & Welsch, R. E. (2022). Natural language-based financial forecasting: A survey. *Artificial Intelligence Review, 55*(1), 61-95. https://doi.org/10.1007/s10462-021-10019-x

Part 2: Publication Statement

Purpose of Publication

This statement confirms our intention to submit our capstone project for educational and research-based publication on the Thullen Research Lab website. We believe our project offers value in several ways:

- It showcases the application of sentiment analysis and machine learning in financial prediction, contributing to our academic and professional portfolio.
- It adds to the Lab's repository of real-world case studies for instructional and reference purposes.
- It provides a foundation for further research, enabling future students and researchers to explore sentiment-based modeling in conjunction with financial market data.

IP and Authorship Acknowledgment

We fully acknowledge that:

- The intellectual property (IP) of this work remains under the authorship of Samuel Adebayo and Guru Raghavendra Reddy Batchu.
- All terms outlined in the Thullen Research Lab IP Policy will be adhered to.
- Any reuse, extension, or derivative of this project - whether by ourselves or others - must credit the original authorship appropriately.

Contribution to the Research Community

Our study offers evidence-based insights into the practical limitations of Reddit-based sentiment analysis in predicting stock market returns. It identifies gaps where future research can be directed, such as:

- Integrating macroeconomic indicators and institutional sentiment data,

- Exploring event-driven sentiment targeting,
- Applying more advanced architectures like BERT-based or Transformer models.

We believe this work can serve as a foundational reference for future academic inquiries into market prediction models and the intersection of social media and finance.

Next Steps

Although we are not currently planning to extend this project beyond the course ourselves, we recognize several meaningful avenues for future development. These opportunities could be explored by the Open Research Team or other research students who may wish to build upon our work:

- Participation in the Open Research Team for second-phase enhancements,
- Submission to academic conferences such as **KDD 2025** or **EMNLP 2025**,
- Presentation in a project workshop or being featured on the Lab's LinkedIn page.

We hope our project serves as a strong foundation for further research and collaboration within the Thullen Research Lab community.

Publication Confirmation

We, Samuel Adebayo and Guru Raghavendra Reddy Batchu, confirm our consent to publish our capstone project titled *"Predicting Market Trends Using Social Media Sentiment Analysis"* through the Thullen Research Lab for educational and academic reference. We understand this is a non-commercial publication and that our IP rights will be retained under the Lab's guidelines.

Authors:

Samuel Adebayo

Guru Raghavendra Reddy Batchu

June 26, 2025

Part 3: Further Development Proposal

Although we are not currently pursuing additional development of this project, we have outlined a roadmap of potential enhancements should the project be extended in a second phase by ourselves or other research students:

1. Broadening Sentiment Sources

Incorporating sentiment data from additional platforms like **Twitter**, **StockTwits**, and **financial news comment sections** could provide a richer and more diverse signal for sentiment-based forecasting.

2. Macroeconomic Indicator Integration

Introducing macroeconomic indicators such as **unemployment rates**, **inflation metrics**, or **consumer confidence indices** would offer more context to market behavior, potentially improving predictive performance.

3. Hybrid Modeling Approach

Developing models that integrate **sentiment analysis**, **technical indicators**, and **macroeconomic data** could create a more robust forecasting framework by leveraging different data dimensions.

4. Advanced Model Architectures

Exploring deep learning architectures such as **BERT**, **FinBERT**, or **Transformer-based models** could help extract more complex relationships between textual data and market outcomes.

5. Event-Driven Sentiment Analysis

Focusing sentiment analysis on specific financial events (e.g., **earnings announcements**, **regulatory changes**, **economic policy decisions**) may reduce noise and reveal more direct market sentiment responses.

6. Enhanced Feature Engineering

Utilizing **metadata** such as Reddit post **frequency**, **upvotes**, **comment volume**, or **thread length** could provide insights into sentiment intensity and the breadth of market attention.

These ideas are offered for future development and research, and we welcome the possibility of others expanding this work under the guidance of the Thullen Research Lab.