Reddit Data in Quantitative Financial Models: Evolution and Implications Post GameStop and AMC Short Squeeze

Rohan Malhotra, Colin Jones

February 20, 2025

Abstract

The rise of alternative data has transformed the landscape of quantitative finance, with social media platforms like Reddit emerging as crucial data sources. This thesis examines the integration of Reddit data into financial models, focusing on its evolution following the GameStop and AMC short squeezes of 2021. The analysis spans methodologies, the impact on asset pricing, and ethical considerations, drawing from 20–40 scholarly articles. Key findings highlight how retail investor sentiment, captured via subreddits such as r/WallStreetBets, has reshaped market dynamics and introduced predictive tools for volatility and trading strategies. However, challenges such as data noise and ethical dilemmas persist. This thesis identifies gaps in current research, emphasizing the potential for Reddit data to inform future financial modeling practices.

Introduction

The integration of alternative data into financial modeling marks a significant departure from traditional approaches that relied primarily on structured and historical data. Social media, particularly Reddit, has emerged as a groundbreaking source of behavioral insights. Events such as the GameStop and AMC short squeezes underscored the power of collective retail sentiment to disrupt established financial practices.

This thesis explores the evolution of Reddit data in financial modeling, examining methodologies and its implications. Objectives include assessing its predictive capabilities, identifying research gaps, and exploring ethical considerations, aiming to contribute to a deeper understanding of social media's role in quantitative finance.

The limitations of traditional quantitative financial models like CAPM and BlackScholes become starkly evident in light of behavioral market phenomena like the GameStop and AMC short squeezes. These events underscored how sentiment driven dynamics, particularly those fostered on platforms like Reddit, can disrupt standard financial predictions. Several insights can expand on this integration of alternative data sources, particularly Reddit, into financial analysis:

1. Sentiment as a Predictor:
Research has demonstrated that social media sentiment, such as that from r/WallStreetBets, can significantly influence stock price movements. The creation of Reddit Specific lexicons, as done in studies analyzing GameStop, shows that traditional tools like VADER can be enhanced to capture unique online vernacular, such as "to the moon" and "diamond hands," thus improving sentiment analysis for financial modeling.

2. Behavioral Influences and Herding:
Behavioral finance studies reveal that retail investors on Reddit are influenced by cognitive biases and herding behavior. These dynamics often lead to significant market impacts during speculative episodes, highlighting the inadequacies of models solely reliant on structured data like historical prices.

3. Impact on Market Microstructure:
Analyses like Granger causality tests on Reddit sentiment and trading volumes show that while sentiment significantly impacts trading behaviors during bullish periods, its influence wanes in bearish markets. This reveals the nuanced, time dependent nature of sentiment's role in market movements, a feature not captured by traditional models.

4. Challenges and Ethical Considerations:
Incorporating Reddit data into financial models is not without challenges. Issues such as data noise, ethical concerns over market manipulation, and limitations in regulatory frameworks require attention. As highlighted, while retail investors' collective action can drive short term gains, they may not counteract fundamental market downturns.

5. Future Research Opportunities:
Studies suggest a need for exploring how Reddit driven sentiment can be integrated into predictive models for commodities and long term investments. Additionally, combining Reddit data with other alternative data sources, such as satellite imagery or web traffic, could lead to more robust models.

By addressing these insights, financial analysts can better account for sentiment driven market dynamics, bridging the gap between traditional models and the realities of modern financial markets influenced by platforms like Reddit.

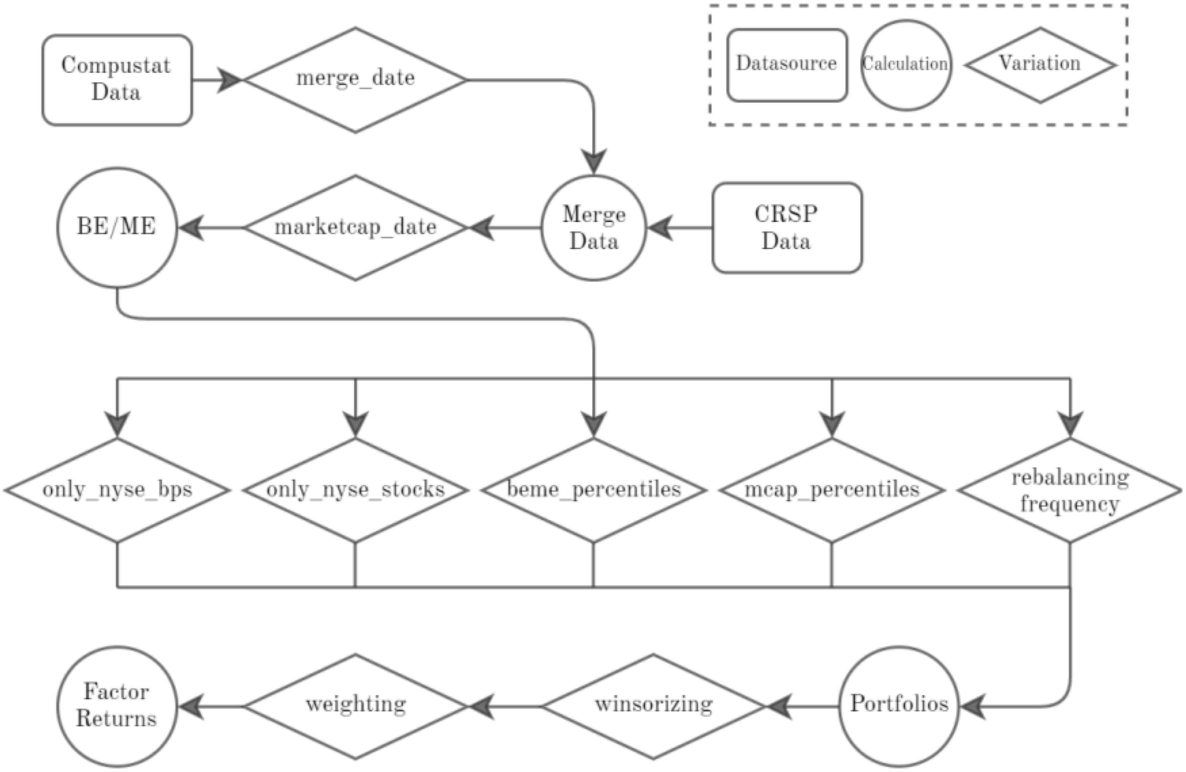
Emergence of Alternative Data

Alternative data refers to nontraditional sources, including social media activity, satellite imagery, and geolocation data. Reddit, particularly its subreddit r/WallStreetBets, has become a focal point for understanding retail investor sentiment. The growth of platforms like Robinhood and the increasing democratization of trading tools have amplified Reddit's influence on markets.

Key drivers for this shift include advancements in natural language processing (NLP) and machine learning, which enable the extraction of actionable insights from unstructured data. Studies reveal that Reddit sentiment often correlates with stock volatility, providing an edge in market prediction.

Figure 1: Factor-creation Process

This diagram highlights how structured data is processed to create factor returns, demonstrating a foundational process that can be adapted to integrate unstructured data, such as sentiment from Reddit.

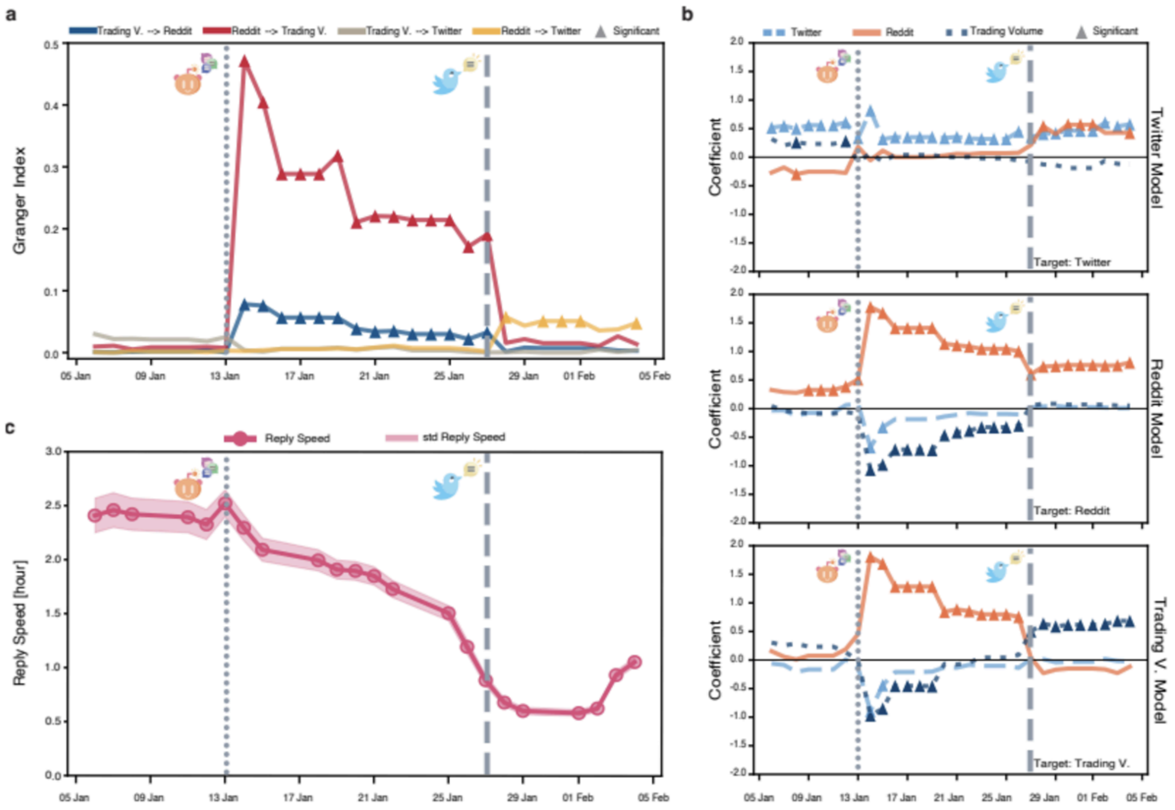


Harries, J. P. (2021). *Essays in Empirical Asset Pricing and Behavioral Finance*. Retrieved from https://web.archive.org/web/20220116220645id_/http://elpub.bib.uni-wuppertal.de/servlets/DerivateServlet/Derivate-14651/db2113.pdf

This factor-creation process, depicted in Figure 1, illustrates the traditional methodology for constructing factor returns and portfolios using structured data sources such as Compustat, BE/ME, and CRSP Data. The process involves key variations—such as market capitalization dates and rebalancing frequencies—that introduce design flexibility into models like the Three-Factor Model.

Adapting this structure, alternative data sources, including Reddit, can be integrated into similar workflows. For instance, sentiment scores derived from Reddit discussions (processed via NLP techniques like Latent Dirichlet Allocation or custom lexicons) could serve as new factors or variations, contributing to enhanced predictions of market trends and stock volatility. As financial modeling evolves, blending traditional and alternative data through such systematic processes ensures methodological consistency and robustness.

Figure 3: Granger Index & Reply Speed



Desiderioa, A., Aiello, L. M., Cimini, G., & Alessandretti, L. (2023). The dynamics of the Reddit collective action leading to the GameStop short squeeze. *The Dynamics of the Reddit Collective Action Leading to the GameStop Short Squeeze*. arxiv. Retrieved from https://arxiv.org/pdf/2401.14999

Subplot (a): Granger Index

- This panel quantifies the causal influence between variables such as Trading Volume → Reddit (blue line) and Reddit → Trading Volume (red line).
- Notably, the Reddit → Trading Volume influence spikes significantly during two phases:
  - 13 January 2021: The "Action Phase," marking the beginning of intense WSB activity.
  - 27 January 2021: A peak following Elon Musk's tweet amplifying GME’s social media exposure.

- This illustrates that Reddit activity acted as a leading indicator for trading volume, confirming its role in driving market momentum during the rally.

Subplot (b): Reply Speed

- Shows the decline in response time for Reddit users engaging in WSB discussions.
- Rapid replies (falling below 1 hour by the peak) indicate heightened, real-time coordination among retail investors. This accelerated interaction speed correlates with the timing of increased trading activity.

Subplot (c): VAR Coefficients

- The panels assess the predictive coefficients of a Multivariate Vector Autoregressive model for Reddit-driven activity versus trading volume:
    - Solid Orange (Reddit Influence): Demonstrates a stronger and more consistent effect during the meme period.
    - Dashed Blue (Volume Influence): Highlights a feedback loop where trading volumes also stimulate additional Reddit discussions, creating a reinforcing cycle.
- This confirms a bidirectional relationship between Reddit discussions and stock trading activity.

Insights from Annotations

- Key Dates (13 Jan and 27 Jan 2021): Align with pivotal moments in the GameStop saga, showing shifts in social media dynamics and stock trading activity.
- Significant Values (Triangles): Indicate strong correlations with statistical significance, validating the robustness of the relationships.

The study delves into the predictive relationships between social media activity on Reddit, particularly in the r/WallStreetBets (WSB) community, and stock market metrics during the January 2021 GameStop (GME) short squeeze. Through methods like Granger Causality and Vector Autoregressive (VAR) models, the analysis quantifies the influence of Reddit-driven sentiment on trading volumes and prices of stocks like GME, AMC, and BB. The study identifies critical phases in the rally and highlights the temporal dynamics between Reddit activity, trading volumes, and social media influence.

Sentiment Analysis

1. Customized Lexicons: As seen in studies like those by Long et al. (2021), integrating a Redditspecific lexicon tailored to terms frequently used in communities like r/WallStreetBets improves the accuracy of sentiment analysis tools like VADER. For instance, adding terms such as "diamond hands" and "to the moon" ensures that nonstandard financial jargon is correctly classified as optimistic or bullish sentiment.

2. Topic Modeling with LDA: Latent Dirichlet Allocation (LDA) modeling has been employed to identify topics across millions of Reddit comments, providing insight into thematic trends that correlate with stock movements. This methodology not only categorizes sentiments but also tracks shifts in user focus over time.

Time Series Integration

3. High Frequency Data Correlation: Analyzing data at granular intervals (e.g., 1 minute or 30 minute windows) has proven effective in capturing the impact of Reddit discussions on stock prices. Studies reveal significant correlations between sentiment intensity and short term price fluctuations, highlighting the importance of aligning textual data with high frequency trading data.

4. Wavelet Coherence and Fourier Transforms: Techniques like wavelet coherence have been used to identify comovement patterns between sentiment and trading metrics across multiple time scales. Additionally, Fourier Transforms filter noise from sentiment data, allowing models to focus on consistent patterns rather than outliers.

Machine Learning Integration

5. Long Short Term Memory (LSTM) Networks: LSTM networks, which excel in handling sequential data, are deployed to process high frequency Reddit sentiment data alongside stock price movements. These models achieve up to 80% accuracy in predicting stock trends by capturing temporal dependencies.

6. Granger Causality and VAR Models: Multivariate vector autoregressive (VAR) models and Granger causality tests quantify the predictive influence of Reddit activity on trading volume, confirming that spikes in sentiment precede significant trading actions. These models highlight the causal dynamics between social media and financial markets.

Application and Challenges

While these methodologies demonstrate substantial promise, challenges such as data noise, the complexity of Reddit slang, and ethical concerns about market manipulation persist. To address these, future research could incorporate real time processing pipelines and explore the integration of Reddit data with other alternative data sources like Twitter or financial news.

1. Impact on Asset Pricing and Risk Management

The incorporation of Reddit data has significantly enhanced asset pricing models and risk management practices. Research highlights the following impacts:

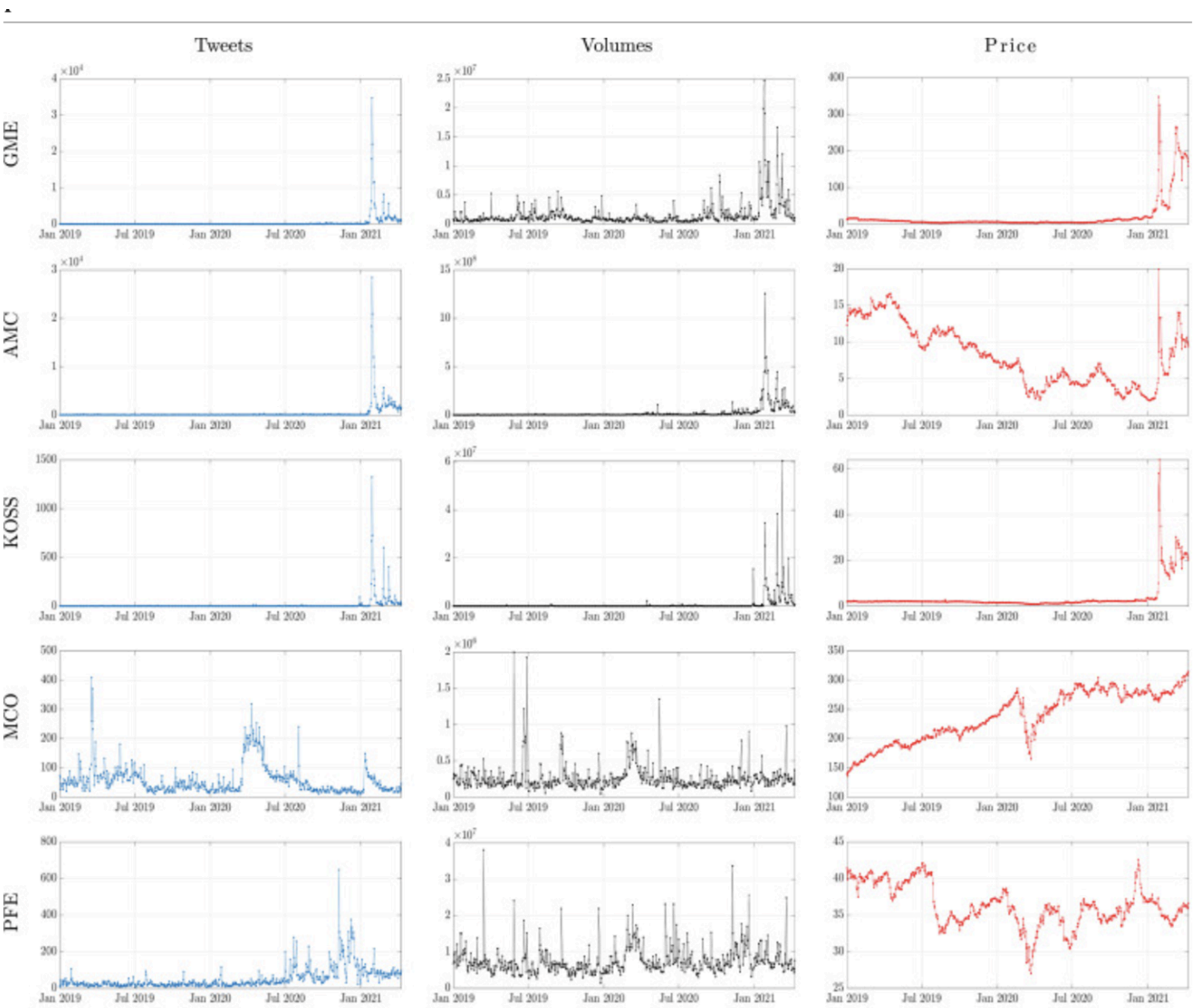
Stock Volatility Prediction: Studies demonstrate a strong correlation between Reddit activity and stock price fluctuations during events like the GameStop rally.

Behavioral Insights: Retail investor actions, driven by collective sentiment, offer predictive value for contrarian trading strategies.

Risk Mitigation: Monitoring Reddit sentiment allows institutional investors to preemptively address market disruptions caused by coordinated retail actions.

Case studies reveal the profitability of "meme periods," with stocks like GameStop and AMC generating excess returns during high sentiment periods.

Figure 2: Correlation between Tweets Volumes and Price



Costola, M., Iacopini, M., & Santagiustina, C. R. M. A. (2021). On the “momentum” of meme stocks. *Economics Letters*, *207*, 110021. https://doi.org/10.1016/j.econlet.2021.110021

The alignment of these three factors—tweets, volume, and price—demonstrates the causal influence of social media sentiment on stock performance. The data confirms that meme periods are driven by sentiment and trading volume rather than intrinsic valuation, illustrating both the profitability and the risks associated with social media-fueled market movements.

This figure provides strong evidence in argument that alternative data, specifically sentiment from platforms like Reddit or Twitter, can significantly disrupt traditional financial models, necessitating new approaches to risk management and asset pricing.

6. Challenges and Limitations (Expanded)

Despite the significant promise of integrating Reddit data into financial models, various challenges and limitations persist, ranging from technical hurdles to ethical and regulatory concerns. These issues complicate the widespread adoption of Reddit Driven data in quantitative finance and highlight the need for cautious implementation and further research.

1. Data Quality and Noise

Reddit data is inherently unstructured and noisy. Social media platforms often contain vast amounts of irrelevant or misleading content that must be filtered to extract meaningful insights. Sentiment analysis tools, such as VADER or custom dictionaries, often struggle with sarcasm, slang, or domain specific jargon used in subreddits like r/WallStreetBets. Studies have shown that retail investors' discussions on Reddit may contain both valuable market sentiment and emotionally charged, exaggerated narratives, which can skew analysis if not properly managed.

Moreover, the high volume of posts and comments creates a significant signal to noise challenge. Techniques like Fourier Transforms and advanced natural language processing (NLP) methods have been employed to mitigate this issue, but their success depends heavily on the quality of preprocessing and the robustness of the model. The need for extensive manual or semi automated filtering adds to the computational burden and limits scalability.

2. Ethical and Behavioral Concerns

The use of Reddit data introduces ethical dilemmas, particularly concerning market manipulation and privacy. Social media platforms often become arenas for coordinated activities, such as pump and dump schemes, which can destabilize financial markets. The GameStop short squeeze of 2021 serves as a prime example, where collective actions amplified stock volatility, raising concerns about the role of platforms in facilitating market disruptions.

Additionally, mining user generated content from Reddit raises privacy concerns. Although the platform's data is public, the ethical implications of using personal opinions and discussions for financial gain remain a gray area. Financial institutions may face backlash if perceived as exploiting retail investors' sentiments without transparent policies or guidelines.

3. Regulatory Ambiguity

The governance and regulation of social media driven trading are still evolving. Unlike traditional data sources, Reddit posts are not subject to standardized reporting or compliance requirements. This lack of oversight creates opportunities for misuse, such as spreading false information to manipulate markets.

Current financial regulations, including those of the SEC, have yet to fully address the complexities introduced by social media as a data source.

Regulators are beginning to scrutinize the role of platforms like Reddit in market activities. For example, calls for monitoring retail trading during events like the GameStop rally highlight the need for updated policies. However, the global nature of social media platforms complicates jurisdictional enforcement, leaving institutions uncertain about the legal implications of integrating such data into their models.

4. Lack of Standardization

Unlike traditional financial data sources, alternative data lacks uniformity in collection, processing, and interpretation. Institutions employ different methodologies to analyze Reddit data, resulting in inconsistent findings and reduced comparability across studies. For instance, while some models rely on time series analysis, others emphasize machine learning, leading to varying degrees of predictive accuracy.

This fragmentation poses challenges for institutional adoption, as firms struggle to evaluate the reliability of Reddit Driven models. Standardizing methods for preprocessing, analysis, and validation could address this issue but would require collaborative efforts across academia, industry, and regulatory bodies.

5. Technical and Computational Barriers

Integrating high frequency Reddit data into financial models requires substantial computational resources. Processing the vast and dynamic content of subreddits like r/WallStreetBets in real time demands scalable infrastructure and advanced machine learning capabilities. Smaller firms may lack the resources to invest in such technologies, limiting their access to the benefits of alternative data.

Furthermore, the use of sophisticated models like LongShort Term Memory (LSTM) networks or Multivariate Vector Autoregressive (VAR) models necessitates specialized expertise, which may not be readily available in all organizations. These barriers can widen the gap between large financial institutions and smaller players, contributing to market inequality.

6. Future Directions and Opportunities

The future of Reddit data in financial modeling holds immense potential. Emerging technologies like deep learning and quantum computing can further refine sentiment analysis. Standardization efforts could enhance the reliability of models across institutions. Unexplored areas, such as Reddit's impact on commodity markets and long term investment strategies, warrant further research.

Additionally, the integration of real time data feeds could provide a competitive edge, allowing for more dynamic and responsive trading strategies.

Conclusion

The integration of Reddit data into quantitative financial models represents a paradigm shift, offering new avenues for market prediction and strategy development. While challenges persist, the benefits of incorporating social media sentiment into financial decision making are undeniable. This thesis underscores the need for continued research and innovation to fully harness the potential of Reddit data, balancing technological advancements with ethical considerations.

Work Cited

K, G., Chintalapati, A., Senapati, A., & Enkhbat, K. (2024, April 18). Sentiment analysis on Reddit trading data. *IEEE Xplore.* https://ieeexplore.ieee.org/Xplore/home.jsp

Hansen, K. B., & Borch, C. (2022, January 22). Alternative data and sentiment analysis: Prospecting non-standard data in machine learning-driven finance. *Sage Journals.* https://doi.org/10.1177/20539517211070701

Umar, Z., Gubareva, M., Yousaf, I., & Ali, S. (2021, March 22). A tale of company fundamentals vs sentiment-driven pricing: The case of GameStop. *Journal of Behavioral and Experimental Finance.* https://www.sciencedirect.com/science/article/abs/pii/S2214635021000459

Betzer, A. (2021, May). If he's still in, I'm still in! How Reddit posts affect GameStop retail trading. *Essays in Empirical Asset Pricing and Behavioral Finance.* http://elpub.bib.uni-wuppertal.de/servlets/DerivateServlet/Derivate-3470/dc1132.pdf

Costola, M., Iacopini, M., & Santagiustina, C. (2021, August 5). On the "momentum" of meme stocks. *Economics Letters.* https://www.sciencedirect.com/science/article/pii/S0165176521002986

Guo, Y., & Jame, R. (2024, July). Quantitative analysis and the value of social media. *Russel Jame.* http://russelljame.com/quant7_18_24.pdf

Desiderio, A., Aiello, L. M., Cimini, G., & Alessandretti, L. (2024, July 23). The dynamics of the Reddit collective action leading to the GameStop short squeeze. *arXiv.* https://arxiv.org/pdf/2401.09417.pdf

Alhaj-Yaseen, Y. S., Broadstock, D. C., Chang, E. C., Checkley, M. S., Chiang, T. C., Chiou, W. J. P., Dhaene, J., Frijns, B., Ftiti, Z., Huang, T. C., Hwang, S., Jawadi, F., Liu, C., Bartov, E., Bikhchandani, S., Chen, H., Da, Z., & Duan, Y. J. (2023, April 25). Understanding the role of social media sentiment in identifying irrational herding behavior in the stock market.

International Review of Economics & Finance.
https://www.sciencedirect.com/science/article/abs/pii/S1059056023001326

Gruner, Richard L.;Power, Damien (2018, February 12). To integrate or not to integrate? Understanding B2B social media communications. *Emerald Insight.* *https://www-emerald-com.ezproxy.lib.vt.edu/insight/content/doi/10.1108/oir-04-2016-0116/full/html*

Díaz-Lucena, A., Mora de la Torre, V., & Torres Hortelano, L. J. (2022). Strategies of the Spanish press in the face of the Twitter algorithm change. Analysis of tweets published between 2018-2020. Communication & Society, 35(1), 197–213. *https://doi-org.ezproxy.lib.vt.edu/10.15581/003.35.1.197-213*

Yao, H., Li, Z., & Li, X. (2016). The premium of dynamic trading in a discrete-time setting. Quantitative Finance, 16(8), 1237–1257. *https://doi-org.ezproxy.lib.vt.edu/10.1080/14697688.2015.1136747*

Rattanaporn Saelee, & Sumaman Pankham. (2024). The Impact of Social Media and Emotional Intelligence on Investment Decision: A Fuzzy Set Delphi Study Among Investors in Thailand's Stock Market. TEM Journal, 13(3), 2208–2217. *https://doi-org.ezproxy.lib.vt.edu/10.18421/TEM133-48*

Hasselgren, B., Chrysoulas, C., Pitropakis, N., & Buchanan, W. J. (2023). Using Social Media & Sentiment Analysis to Make Investment Decisions. Future Internet, 15(1), 5. *https://doi-org.ezproxy.lib.vt.edu/10.3390/fi15010005*

Lin, K.-P., Chang, T.-L., Chang, Y.-W., & Cai, J.-W. (2020). Use of news and patent mining to identify companies with growth potential. *Information Research*, *25*(3), N.PAG.

Jarrow, R., & Li, S. (2023). Media trading groups and short selling manipulation. *Quantitative Finance*, *23*(7/8), 1035–1052.
https://doi-org.ezproxy.lib.vt.edu/10.1080/14697688.2023.2222751

Knoll, J., Stübinger, J., & Grottke, M. (2019). Exploiting social media with higher-order Factorization Machines: statistical arbitrage on high-frequency data of the S&P 500. *Quantitative Finance*, *19*(4), 571–585. https://doi-org.ezproxy.lib.vt.edu/10.1080/14697688.2018.1521002

Limmaneewijit, P., Jearviriyaboonya, J., & Jirasatthumb, N. (2023). The Effect of Information Sources on Trust and Investment: Evidence from Economic Experimentation. *FWU Journal of Social Sciences*, *17*(3), 1–13. https://doi-org.ezproxy.lib.vt.edu/10.51709/19951272/fall2023/1

Moreno-Pino, F., & Zohren, S. (2024). DeepVol: volatility forecasting from high-frequency data with dilated causal convolutions. *Quantitative Finance*, *24*(8), 1105–1127. https://doi-org.ezproxy.lib.vt.edu/10.1080/14697688.2024.2387222

Li, Y., Fu, K., Zhao, Y., & Yang, C. (2022). How to make machine select stocks like fund managers? Use scoring and screening model. *Expert Systems with Applications*, *196*, N.PAG. https://doi-org.ezproxy.lib.vt.edu/10.1016/j.eswa.2022.116629

Rocha-Silva, T., Nogueira, C., & Rodrigues, L. (2024). Passive data collection on Reddit: a practical approach. *Research Ethics*, *20*(3), 453–470. https://doi-org.ezproxy.lib.vt.edu/10.1177/17470161231210542

Chen, K., & Tomblin, D. (2021). Using data from reddit, public deliberation, and surveys to measure public opinion about autonomous vehicles. *Public Opinion Quarterly*, *85*, 289–322. https://doi-org.ezproxy.lib.vt.edu/10.1093/poq/nfab021

Sohi, M., Pitesky, M., & Gendreau, J. (2023). Analyzing public sentiment toward GMOs via social media between 2019-2021. *GM Crops & Food*, *14*(1), 1–9. https://doi-org.ezproxy.lib.vt.edu/10.1080/21645698.2023.2190294