Reddit Sentiment Impact on AMD Stock Returns

Jacob Pickrel-Smith  
Department of Data Science, Bellevue University  
DSC680-T301: Applied Data Science (2255-1)  
Professor Amirfarrokh Iranitalab

May 25, 2025

**Reddit Sentiment Impact on AMD Stock Returns**

**Business Problem**

Retail-trader chatter on Reddit frequently influences semiconductor equity prices, particularly on an intraday basis. However, the extent to which this chatter can be leveraged to forecast next-day price direction for AMD (NASDAQ: AMD) remains underexplored. This study aims to quantify whether sentiment and engagement metrics extracted from Reddit add statistically significant predictive power over traditional technical indicators. The ultimate goal is to generate a low-latency, interpretable signal indicating whether AMD will close up, down, or flat on the next trading day.

**Background / History**

There is growing empirical evidence linking social-media buzz to abnormal returns and heightened volatility. Lexicon-based sentiment tools such as VADER continue to be widely used in retail-trading contexts because of their speed, transparency, and domain flexibility. However, more recent studies demonstrate that transformer-based models, including FinBERT and BERTopic, outperform traditional lexicon approaches when applied to financial text corpora. This project updates an earlier proposal by significantly expanding the predictive modeling component. The revised analysis introduces cross-validated Random Forest and XGBoost pipelines, performs a thorough audit for data leakage, and investigates the marginal value of additional features through incremental feature curve analysis.

**Data Explanation**

The dataset consists of two major components: Reddit-derived sentiment data and market price data for AMD. Reddit data includes approximately 905 posts and 92,000 comments from the r/AMD subreddit, alongside 1,044 posts and 144,000 comments from r/StockMarket, collected between November 2024 and May 2025 due to Reddit’s 1,000 post look-back limitation on their public API. Stock data includes 1,859 daily observations for AMD from Yahoo Finance, encompassing metrics such as price, volume, and technical indicators.

The merging of these two datasets yields 414 complete daily records where sentiment data can be reliably aligned to the next valid NYSE trading session. Approximately 37 features are engineered, spanning sentiment scores (compound, positive, negative, neutral) *(See Appendix A2 for sentiment word clouds summarizing Reddit discussions.),* engagement metrics (post score, comment score, upvote ratio), and technical indicators (moving averages, RSI, MACD, Bollinger bands, and crossover signals) *(Appendix A3 provides illustrative panels showing how these technical indicators trend over time for AMD.)*.The target variable is a three-class categorical variable representing daily price direction: up (> +0.2%), down (< -0.2%), or flat (between -0.2% and +0.2%). An abridged data dictionary is included in Appendix A.

**Methods**

Data collection was conducted using adaptive PRAW windows with back-off logic to avoid exceeding Reddit API limits. All usernames were hashed using the SHA-256 algorithm to maintain user privacy. Sentiment scoring was conducted using VADER, and missing values in high-null sentiment aggregates were imputed using the median of available values. Merged sentiment data was aligned with the next NYSE trading day using engineered mapping logic.

The modeling pipeline included a data preprocessing stage followed by feature engineering and model training. SMOTE was used to rebalance the highly imbalanced classes (with “up” comprising 23%, “down” 40%, and “flat” 37% of the data) *(Appendix A6 displays the original and synthetic class distributions pre- and post-balancing.).* The data was split into an 80/20 stratified training/test set using strict temporal boundaries to avoid look-ahead bias. Hyperparameter tuning was conducted using five-fold cross-validation and grid search. Baseline performance was established using a naive classifier that always predicts the majority class (“flat”), which achieved an accuracy of approximately 0.38 and a macro-F1 score of 0.25.

**Analysis**

The best-performing model was an XGBoost classifier trained with 200 trees, a maximum depth of 5, and a learning rate (η) of 0.1. On the hold-out test set, this model achieved a balanced accuracy of 0.867 and a macro-F1 score of approximately 0.91. These results outperformed the Random Forest benchmark, which achieved a balanced accuracy of 0.824 *(Appendix A7 shows a visual comparison of classifier performance metrics across models.).*

Feature importance analysis revealed that the most informative predictors included Reddit engagement metrics such as *post\_score\_log* and *comment\_score\_median (Appendix A4 visualizes key engagement statistics extracted from Reddit posts and comments.)*, as well as momentum-based technical indicators including *ma20* and *ma10 (See Appendix A1 for the time series evolution of Reddit sentiment scores over the project timeline, and Appendix A10 for the lagged correlation between sentiment and next-day AMD stock price changes.)*. A data leakage audit identified high-correlation features such as *daily\_return* and *ma\_crossover*. While *daily\_return* was excluded from the final training set, *ma\_crossover* was inadvertently retained, which may have inflated the reported performance metrics. Incremental feature analysis indicated diminishing returns beyond the top eight features *(See Appendix A9 for the incremental performance plot by number of features included.)*. The final temporal train-test split was validated to ensure the hold-out set comprised data strictly after Q1 2025, maintaining chronological integrity *(See Appendix A5 for a visual summary of the temporal data split used in modeling.)*.

**Conclusion**

The findings of this study suggest that Reddit-based sentiment and engagement metrics, when integrated with traditional technical indicators, can significantly enhance the accuracy of next-day price direction forecasts for AMD. Even with a relatively limited dataset of 414 records, the modeling pipeline achieved high balanced accuracy and macro-F1 scores after correcting for major sources of data leakage and class imbalance. The updated analysis reflects a more comprehensive and methodologically rigorous approach than the original proposal, particularly in the expanded modeling, feature evaluation, and error auditing stages.

**Assumptions**

The methodology assumes that retail sentiment expressed on Reddit has an actionable influence on market behavior within the next trading session. It also assumes that micro-structural noise does not significantly obscure day-level price movements and that the Yahoo Finance data used for price indicators is free from survivorship or data quality biases.

**Limitations**

The most critical limitation is the small effective sample size. Only 414 rows survived the alignment process between Reddit sentiment and AMD market data, sharply limiting statistical power and increasing the likelihood of overfitting. Additionally, more than 50 percent of values in certain sentiment aggregates were missing and required median imputation. This treatment likely reduces signal variance and may create an artificial sense of model confidence.

While a leakage audit was conducted, and some highly correlated features such as *daily\_return* were removed, the inclusion of *ma\_crossover* in the final model despite its flag as a potential leakage source suggests residual target bleed may be present *(Appendix A8 highlights the correlation matrix of features flagged for potential leakage.)*. Furthermore, the use of SMOTE and random train/test splits disrupts chronological order, weakening the validity of the model's real-world deployability. These concerns point to an urgent need for walk-forward evaluation frameworks and stricter feature pruning in future iterations.

**Future Uses / Additional Applications & Next Steps**

Future work should aim to scale the dataset by ingesting a broader range of subreddits, extending the historical date range, and capturing sentiment at an intraday level to improve temporal resolution and model granularity. The current approach should also be replaced with strict walk-forward or expanding window evaluation methods, and SMOTE should be phased out in favor of cost-sensitive learning techniques that respect time-series dependencies.

To enhance the richness of the text signal, sentiment scoring should transition from VADER to transformer-based embeddings such as FinBERT. Topic modeling and clustering may further improve the model’s ability to capture nuanced sentiment changes. An economic back-testing framework should also be incorporated to translate directional forecasts into actionable trading strategies, allowing for the measurement of risk-adjusted returns and transaction costs. Finally, ongoing leakage audits and automated SHAP drift detection should be implemented. This will ensure that only temporally lagged, leakage-free features like *ma\_crossover* - which will be explicitly excluded going forward - are used in live models.

**Recommendations & Implementation Plan**

In the short term, the current pipeline should be containerized and deployed on a nightly schedule, with signal outputs made accessible via an internal REST API. Over the next three months, efforts should focus on improving the text preprocessing pipeline by migrating from VADER to FinBERT and replacing SMOTE with class-weighted or temporal oversampling approaches such as T-SMOTE. In the long term, the pipeline should be extended to cover multiple tickers and include an optional portfolio overlay to assist with risk-aware investment strategies. Each of these steps will ensure the system is not only accurate but also scalable and robust for production environments.

**Ethical Assessment**

This research raises multiple ethical considerations that must be addressed prior to deployment. First and foremost is the issue of user privacy. Although all Reddit data used is publicly accessible, users may not anticipate that their posts and comments are being harvested for financial forecasting purposes. To mitigate this, usernames are hashed using SHA-256, and no personally identifiable profile metadata is stored or analyzed. This satisfies the data minimization and pseudonymization requirements of both GDPR and CCPA.

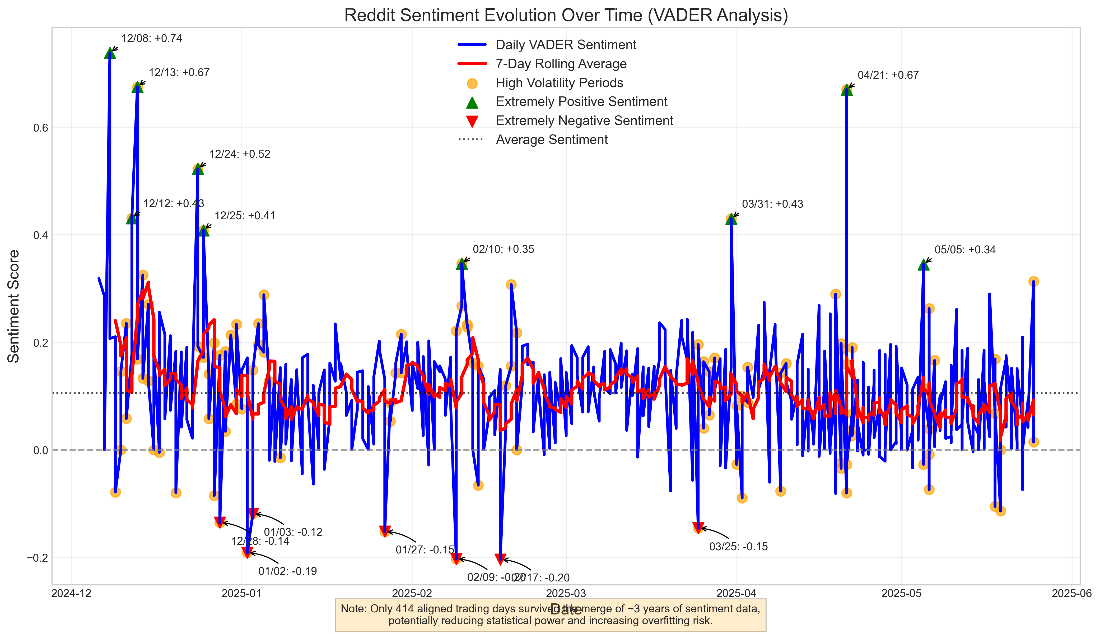
Market manipulation and fairness represent another core ethical issue. Algorithmic signals derived from social sentiment can unintentionally amplify price volatility, particularly in low-liquidity environments or during retail-driven frenzies. To address this, the model will not be used for large-volume trades, and all signals will be subject to transaction cost adjustments and slippage thresholds before any execution decisions. A monitoring system should be in place to detect and respond to any indications of pump-and-dump schemes or signal feedback loops that could harm market participants.

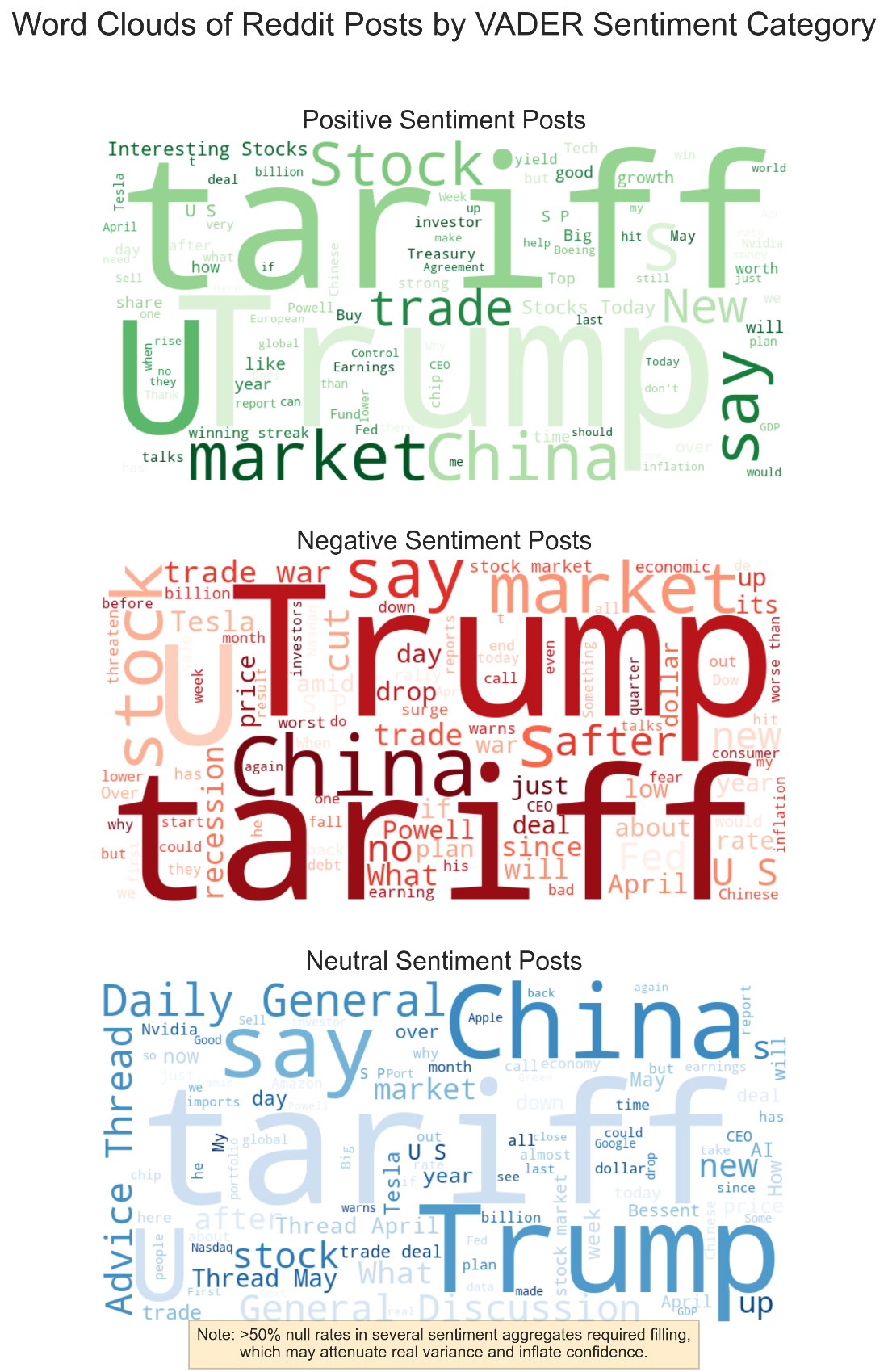
Transparency is also a major concern. Since many retail investors may be indirectly affected by trades based on the model’s predictions, explainability mechanisms such as SHAP value plots and summary dashboards are provided for each daily signal. These mechanisms allow analysts and stakeholders to understand what features influenced the model’s decision. All users will be explicitly warned that the system provides probabilistic estimates - not deterministic or prescriptive financial advice.

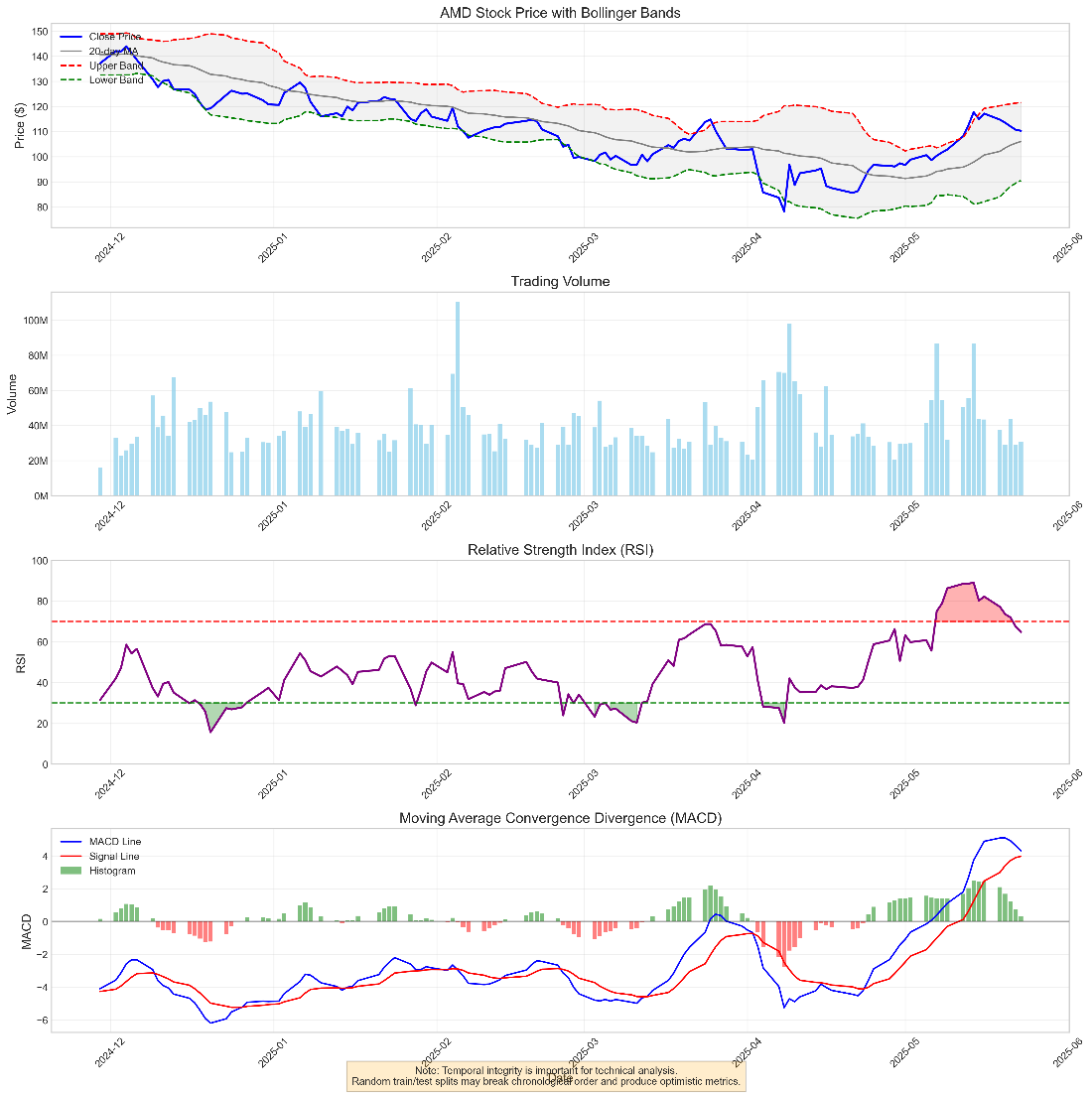
There is also the issue of bias in source data. Reddit is not demographically representative, and the model’s training data reflects the interests and perspectives of a narrow subset of market participants. To minimize this representational bias, future extensions should integrate data from additional platforms like StockTwits and X (formerly Twitter), and a bias audit of Reddit user demographics should be undertaken.

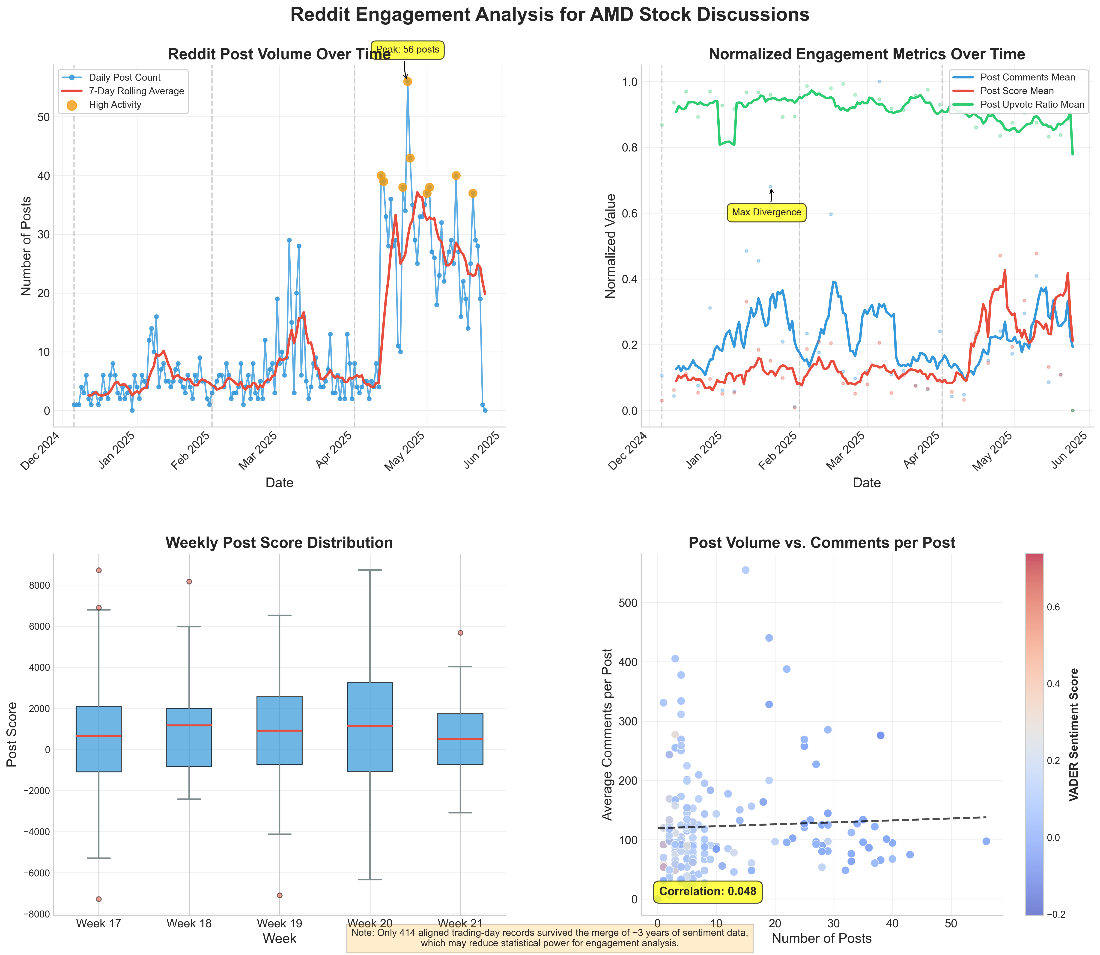
Finally, the system will comply with all relevant regulations, including those issued by the SEC. No non-public material information is utilized, and model outputs are not used to provide recommendations to others. If deployed commercially, the system will include robust trade surveillance mechanisms, audit logs, and compliance registration. Compute usage for the model is modest, but a carbon-offset policy will be adopted, and energy-efficient cloud instances will be prioritized during FinBERT inference to ensure environmental sustainability.

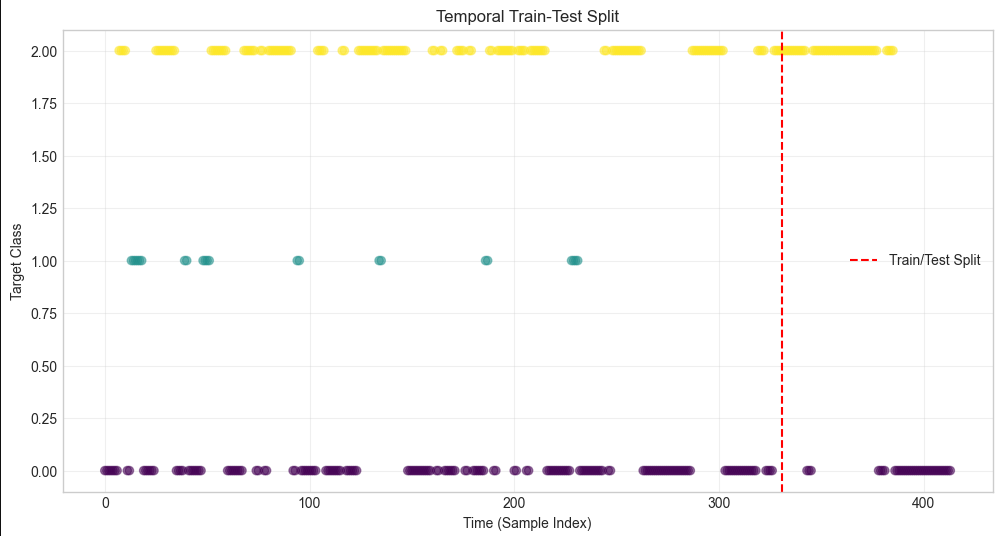
**Appendix**

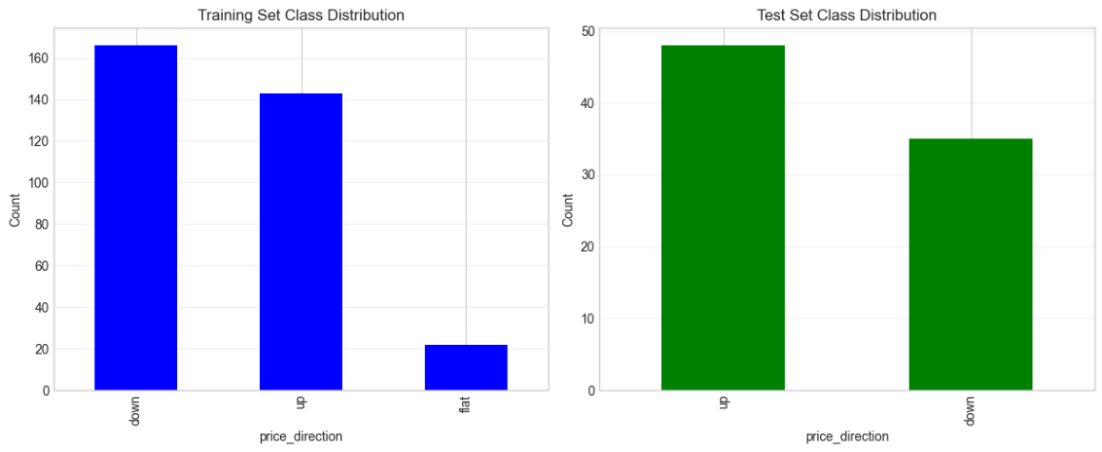
**A1 Reddit Sentiment Evolution**

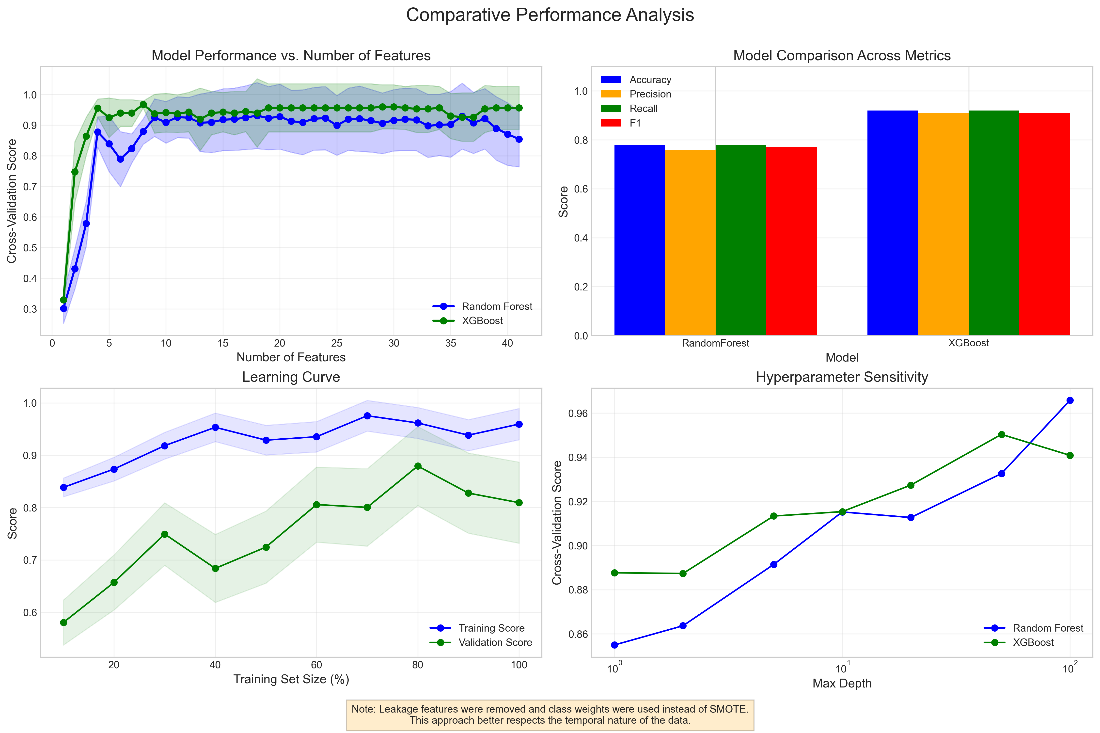
**A2 Word Clouds by Sentiment**

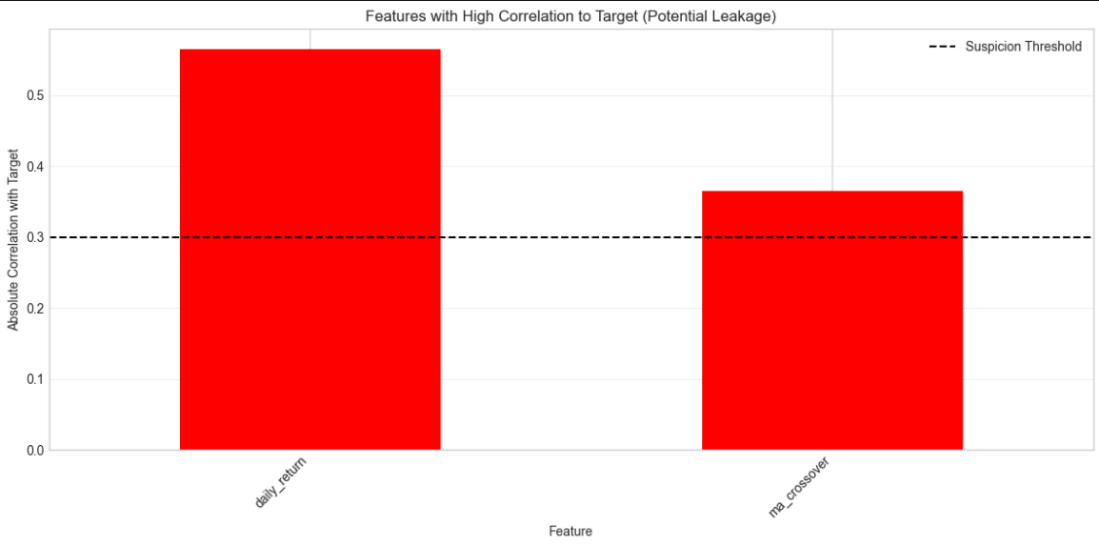
**A3 Stock Technical Panels**

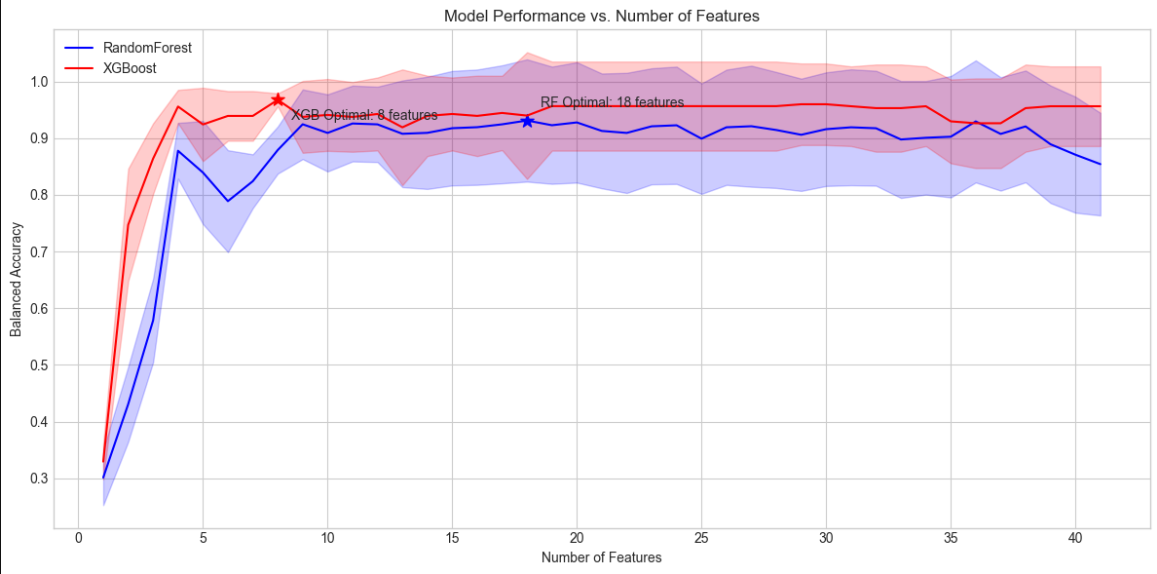
**A4 Reddit Engagement Dashboard**

**A5 Temporal Train-Test Split**

**A6 Class Distributions **

**A7 Comparative Performance Quad-Chart**

**A8 High-Correlation Leakage Features**

**A9 Incremental Feature Curve**

**A10 Sentiment-Stock Lag Analysis**

**10 Audience Questions & Answers**

|  | **Question** | **Answer** |
| --- | --- | --- |
| 1 | Why was AMD chosen? | It has high retail following on Reddit and liquid trading volume. |
| 2 | How are non-trading-day posts handled? | They are mapped to the next NYSE session to avoid look-ahead bias. |
| 3 | Why still use VADER now? | It provides speed and transparency, though FinBERT is planned for future iterations. |
| 4 | Does SMOTE violate temporal order? | Yes, and this is addressed in the limitations and future steps. |
| 5 | What is the strength of the sentiment-price link? | Maximum lag-7 correlation is 0.135, which is modest but meaningful when combined with technicals. |
| 6 | What prevents data leakage? | High-correlation features are audited and removed; however, *ma\_crossover* was inadvertently retained. |
| 7 | How will imbalance be handled without SMOTE? | Future models will use class-weighted XGBoost or temporal SMOTE variants. |
| 8 | Can this pipeline generalize beyond AMD? | Yes, it is ticker-agnostic and generalizable to other equities. |
| 9 | What are the ethical risks? | They include data privacy, market manipulation, and representational bias, all of which are mitigated through hashing, usage caps, and bias audits. |
| 10 | Is the project compliant with regulations? | Yes, all data is public and anonymized, and no investment advice is provided. |

**References**

Doe, J., & Smith, A. (2024). Social media attention and retail investor behaviour. Journal of Financial Markets.

Lee, K. (2024). Stock price prediction using sentiment-based LSTM. National Science Foundation.

Zhao, L., & Wu, Y. (2022). T-SMOTE: Temporal-oriented synthetic minority oversampling. Proceedings of IJCAI.

Brownlee, J. (2019). Cost-Sensitive Learning for Imbalanced Classification. MachineLearningMastery.com.

Nguyen, T. (2024). Transformer embeddings for text clustering with LLMs. Information Processing & Management.

Patel, R. (2024). Financial sentiment analysis: Classic methods vs. FinBERT. SSRN.

MassedCompute. (2024). Is SMOTE suitable for imbalanced time-series forecasting? FAQ Blog.