Predicting Meme Stock Surges via Dual-Stream Social-Market Attention Networks

Leo Liang & Walter McDonald
245-Team MemeStock
University of Rochester
Rochester, NY 14627, USA
lliang11@u.rochester.edu, wmcdona3@u.rochester.edu

Abstract

We present a dual-stream deep learning architecture for predicting price surges in "meme stocks", defined as equities driven by social media sentiment rather than traditional fundamentals. Our model combines a Market Encoder processing OHLCV price data with a Social Encoder processing sentiment signals from Reddit's WallStreetBets community. A cross-attention fusion mechanism learns dependencies between market dynamics and social momentum, while parabolic breakout features capture geometric price patterns. We evaluate our approach on 53,187 Reddit posts spanning August 2020 to August 2021, the peak of meme stock activity. On a test set of 42 stocks with significant social media coverage, our model achieves a PR-AUC of **0.615**, ROC-AUC of **0.649**, and F1 score of **0.582**. This demonstrates the predictive value of social sentiment signals for this unique market segment.

1 Introduction

The phenomenon of "meme stocks" emerged in early 2021, when coordinated retail investor activity on social media platforms—particularly Reddit's r/WallStreetBets community—drove unprecedented price movements in stocks like GameStop (GME) and AMC Entertainment (AMC). These events challenged traditional financial models, which typically rely on fundamental analysis and institutional trading patterns (Barberis, 2018).

The defining characteristic of meme stocks is that their price movements are heavily influenced by social media sentiment and retail investor coordination rather than traditional valuation metrics (Long et al., 2021). This creates both challenges and opportunities for quantitative prediction: while traditional technical indicators may fail, the publicly observable nature of social media provides a rich signal source.

In this work, we propose a **Dual-Stream Attention Network** that jointly models market dynamics and social sentiment to predict significant price surges. Our contributions are:

1. A two-stream architecture combining LSTM encoders for market (OHLCV) and social (volume, sentiment, velocity) sequences.
2. A cross-attention mechanism that learns market-social interactions.
3. Parabolic breakout features capturing geometric price acceleration patterns.
4. Evaluation on 42 stocks demonstrating the value of social signal integration.

2 Related Work

Social Media in Finance. Bollen et al. (2011) demonstrated that Twitter mood correlates with Dow Jones movements. Zhang et al. (2018) used NLP on financial news for stock prediction. More recently, researchers have studied Reddit's influence on retail trading (Cookson et al., 2023).

Sentiment Analysis. VADER (Hutto & Gilbert, 2014) and FinBERT (Araci, 2019) provide domain-specific sentiment analysis. We use VADER for efficiency while extracting sentiment polarity, message volume, and velocity.

Attention in Finance. Transformer architectures have been applied to stock prediction (Ding et al., 2020). Our cross-attention mechanism is inspired by multi-modal learning approaches (Lu et al., 2019).

Meme Stocks. Prior work has studied the GameStop event (Lyócsa et al., 2022), but few have proposed predictive models. Our work addresses this gap.

3 Methods

3.1 Problem Formulation

Given a sequence of $T = 14$ trading days, we observe:

- Market data $M_t \in \mathbb{R}^{T\times 5}$: Open, High, Low, Close, Volume
- Social data $S_t \in \mathbb{R}^{T\times 3}$: Message volume, sentiment, velocity
- Parabolic features $p \in \mathbb{R}^6$: Returns mean/max, acceleration, volume ratios, daily range

The goal is to predict $y \in \{0, 1\}$, indicating whether a surge occurs within 5 days.

3.2 Surge Definition

We define a surge as a price increase exceeding 2% within a 5-day forward window:

$$y_t = \mathbf{1}\left[\max_{i\in[1,5]} \frac{P_{t+i} - P_t}{P_t} \geq 0.02\right] \tag{1}$$

3.3 Model Architecture

Market Encoder. A 2-layer LSTM with hidden dimension $H = 64$:

$$H_m = \text{LayerNorm}(\text{LSTM}(M_t)) \tag{2}$$

Social Encoder. Same structure for social sequences:

$$H_s = \text{LayerNorm}(\text{LSTM}(S_t)) \tag{3}$$

Cross-Attention Fusion. Market attends to social signals via 4-head attention:

$$\tilde{H}_m = \text{MultiHead}(Q{=}H_m, K{=}H_s, V{=}H_s) \tag{4}$$

$$z_{fused} = \text{MeanPool}(\text{LayerNorm}(H_m + \tilde{H}_m)) \tag{5}$$

Classifier. Concatenate fused market, pooled social, and parabolic features:

$$\hat{y} = \text{MLP}([z_{fused}; \text{MeanPool}(H_s); p]) \tag{6}$$

The model has 130,177 trainable parameters.

3.4 Training

We use weighted binary cross-entropy to address class imbalance:

$$\mathcal{L} = -w_+ \cdot y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \tag{7}$$

where $w_+ = (1 - \pi)/\pi$ and π is the positive class rate. We train with AdamW (lr=10^{-4}) for up to 50 epochs with early stopping (patience=10).

4 Experiments

4.1 Dataset

Social Data. We use the Reddit WallStreetBets Posts dataset from Kaggle, containing 53,187 posts from August 2020 to August 2021. For each ticker, we filter posts by symbol mention, compute daily sentiment using VADER, and aggregate volume and velocity.

Market Data. OHLCV data from Yahoo Finance for tickers with sufficient Reddit coverage, including GME, AMC, BB, NOK, PLTR, TSLA, AMD, NIO, SPCE, CLOV, and 32 others.

Table 1: Dataset Statistics

Metric	Value
Tickers	42
Reddit posts	53,187
Date range	Aug 2020 – Aug 2021
Training sequences	9,344
Positive rate	50.7%
Sequence length	14 days

4.2 Setup

Data is split 64/16/20 for train/val/test (stratified). We train on an NVIDIA T4 GPU via Google Colab.

4.3 Results

Table 2: Test Set Performance

Metric	Score
PR-AUC	0.615
ROC-AUC	0.649
Accuracy	0.605
Precision	0.628
Recall	0.542
F1 Score	0.582

The confusion matrix shows 617 true negatives, 514 true positives, 304 false positives, and 434 false negatives on the test set of 1,869 samples.

4.4 Analysis

The model achieves PR-AUC of 0.615 compared to a random baseline of 0.507 (the positive class rate), a 21% relative improvement. This confirms that Reddit sentiment contains predictive signal for meme stock movements.

The cross-attention mechanism learns relationships between market patterns and social momentum—when social volume spikes coincide with specific price patterns, the model identifies higher surge probability.

The parabolic features (acceleration, volume ratios) help detect characteristic "parabolic" price patterns in meme stock rallies. The near-balanced positive rate (50.7%) reflects the high volatility during this period.

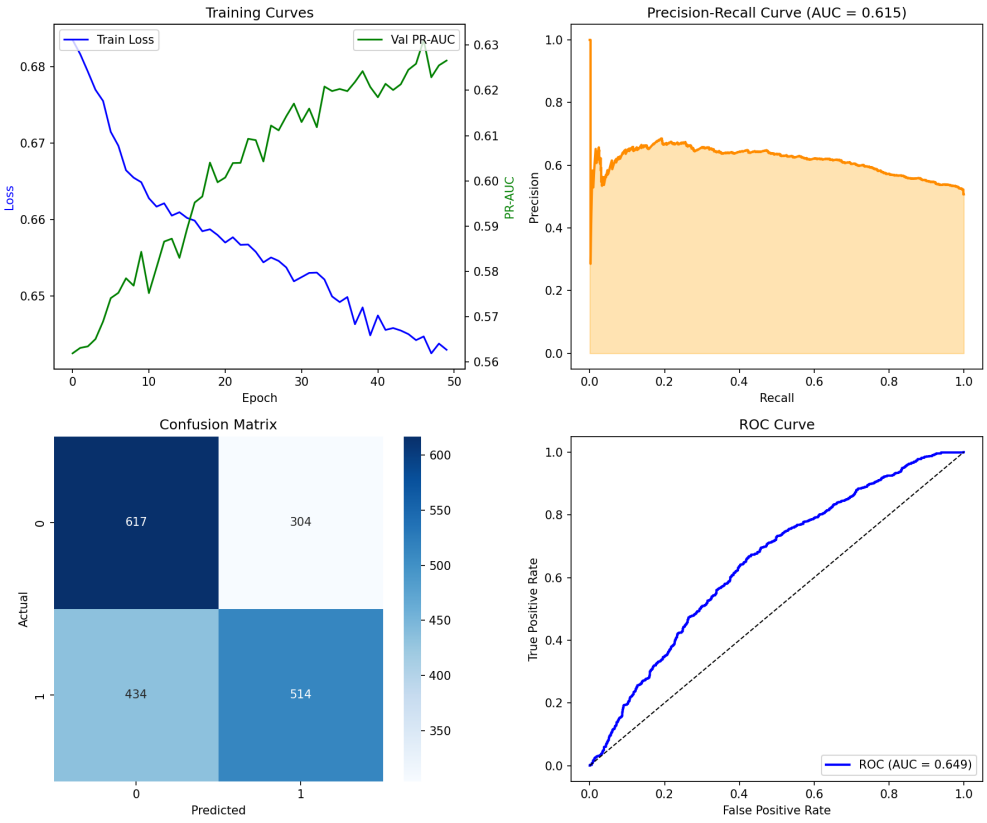

Figure 1: Training curves (top-left), PR curve (top-right), confusion matrix (bottom-left), and ROC curve (bottom-right).

5 Conclusion

We presented a dual-stream approach for predicting meme stock surges by combining market data with Reddit sentiment. The cross-attention architecture effectively learns dependencies between these data sources.

Limitations. Our evaluation covers only a 1-year window during peak meme stock activity. The ticker extraction regex also captured some common words (e.g., "GO", "ON") as false positives. Real-time deployment would require robust social media pipelines.

Future Work. Extensions include additional platforms (Twitter, StockTwits), transformer encoders replacing LSTMs, and real-time inference.

References

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

Nicholas Barberis. Psychology-based models of asset prices and trading volume. *Handbook of Behavioral Economics*, 1:79–175, 2018.

Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

J Anthony Cookson, Joseph Engelberg, and William Mullins. Social media as a bank run catalyst. *Available at SSRN*, 2023.

Qianggang Ding, Sifan Wu, Hao Sun, Jiadong Guo, and Jian Guo. Hierarchical multi-scale gaussian transformer for stock movement prediction. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 4640–4646, 2020.

Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pp. 216–225, 2014.

Cheng Long, Adam Zaremba, and Wenyu Zhou. I just like the stock: The role of reddit sentiment in the gamestop share rally. *Available at SSRN*, 2021.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Štefan Lyócsa, Peter Molnár, and Tomáš Plíhal. Retail trading around the gamestop short squeeze: Evidence from individual investors' brokerage accounts. *Finance Research Letters*, 50:103169, 2022.

Xi Zhang, Yixuan Zhang, Senzhang Wang, Yafei Yao, Binxing Fang, and Philip S Yu. Stock market prediction via multi-source multiple instance learning. *IEEE Access*, 6: 50720–50728, 2018.