NLP and Finance: Can r/wallstreetbets Predict Stock Movements?

Final Project Report for CS406 NLP Course, January 2026

Sven Juhnke
University of Hamburg
9juhnke@informatik.uni-hamburg.de

Abstract

Financial sentiment analysis is typically applied to formal news sources and has limited transferability to informal social media discussions. This project investigates whether and how sentiment on *r/wallstreetbets (WSB)* can trigger significant market movements for a specific stock. To this end, multiple Transformer models were fine-tuned via domain adaptation on an annotated WSB dataset. The results show that an unadapted FinBERT largely fails to capture the sentiment, while a domain-adapted DeBERTa v3 achieves the best classification performance. Subsequent backtesting using thousands of posts over three years reveals via lead-lag analysis that sentiment on WSB is reactive for the selected large-cap stock. The project provides a robust template for testing whether WSB actually leads a chosen stock price or merely reacts to it. Jupyter Notebook available at https://github.com/9juhnke/CS406_NLP

1 Introduction

Sentiment analysis in the financial sector has traditionally focused on formal financial news and corporate reports. However, events such as the *GameStop short squeeze* have demonstrated that social media platforms are now capable of triggering significant market movements. In particular, the subreddit *r/wallstreetbets (WSB)* has established itself as a relevant driver of retail investor sentiment.

This project aims to develop an application that not only classifies retail investor sentiment but also quantitatively measures how the market reacts to these discussions. The subject of investigation is the Nvidia stock ($NVDA), which received exceptionally high attention during the investigation period from 2022 to 2025 due to the AI boom.

1.1 Problem Statement and Challenges

Transferring classic financial sentiment analysis approaches to social media involves specific challenges:

- *Linguistic Divergence:* Pre-trained financial models like FinBERT are primarily based on formal text sources such as Reuters news. In contrast, social media usage is often highly informal, ironic, and slang-heavy, especially in forums like WSB. Sentiment-bearing terms such as *YOLO*, *Tendies*, or *Diamond Hands* are difficult for these pre-trained models to interpret.
- *Data Availability:* Due to recently restrictive API access, stable real-time analysis is currently not feasible. However, a pure snapshot via live labeling would not provide a statistically robust basis for investigating market reactions, as temporal dependencies and lag effects would remain unconsidered.

1.2 Approach

To solve this, the project pursues a Domain Adaptation approach combined with historical backtesting. Instead of an isolated live analysis, an end-to-end pipeline with the following contributions is implemented:

- *Comparative Domain Adaptation:* Due to the described linguistic divergence, using FinBERT for WSB is not optimal. Therefore, it is adapted to the linguistic domain through subsequent fine-tuning. Since it is unclear whether FinBERT actually offers the best basis for this adaptation—or if prior domain adaptation to formal financial news might even be detrimental—the standard BERT Base model and the modern DeBERTa v3 are also adapted to the domain.
- *Backtesting instead of Snapshot Analysis:* The subsequent analysis is not a snapshot but is based on an extensive historical dataset spanning from March 2022 to March 2025.

- *Weighted Sentiment Aggregation:* To establish a relation between the daily stock price and the sentiments of individual posts, an aggregation metric is introduced to calculate a daily sentiment score. In addition to the sentiment from the pure text content (polarity), the social resonance of the contributions is also considered using community votes (score) and the number of comments.

- *Lead-Lag Analysis:* To quantitatively answer the initial question of how the market reacts to discussions, a time series analysis is conducted over a window of plus/minus several trading days to determine whether WSB sentiment leads the stock price or is reactive.

2 Theoretical Background and Related Work

Traditional sentiment analysis, such as on movie reviews, is hardly transferable to financial texts, as financial language is highly nuanced and context-dependent. The state-of-the-art approach in this field is FinBERT (Araci, 2019), a BERT-based model fine-tuned on a corpus of Reuters news and corporate filings. While FinBERT achieves excellent results in formal contexts, it is expected that performance drops significantly on informal channels, as the model did not see the specific jargon of private retail investors during pre-training.

Communication on platforms like Reddit, especially in the subreddit *r/wallstreetbets*, differs from editorial financial news. Users employ slang, irony, and emojis, which standard models often fail to detect or misinterpret.

A central challenge is the lack of high-quality, labeled datasets. For reliable analysis, targeted **Domain Adaptation** via fine-tuning of Transformer models on verified gold-standard datasets is required to learn the community's specific vocabulary and implicit meanings. A dataset identified to meet these requirements is presented in Section 3.

To evaluate the effectiveness of domain adaptation, a comparison of different model generations is necessary. BERT (Base) serves as a generic baseline with no specific financial knowledge. FinBERT acts as a financial domain-specific baseline without knowledge of WSB jargon. As an advancement, DeBERTa v3 is considered, which is characterized by a disentangled attention structure and improved masking (He et al., 2023). Other BERT-based models, such as RoBERTa, were tested experimentally but settled between BERT and DeBERTa and are not further considered here due to space constraints.

The goal of this project is not only sentiment classification but the measurement of market reaction. This is based on the principles of Behavioral Finance, which, unlike the Efficient Market Hypothesis (Fama, 1970), assumes that investor sentiment can influence short-term price anomalies and volatility (Shiller, 1981; Barberis et al., 1998).

The empirical validation of such hypotheses requires **backtesting** on historical data. Related works show that text and social media signals can correlate with returns, volume, and volatility, both for classic media (Tetlock, 2007) and for microblogging and Twitter data (Bollen et al., 2011; Sprenger et al.). An extensive dataset suitable for backtesting is also presented in Section 3.

3 Data and Preprocessing

A central challenge of this project was obtaining high-quality training and backtesting data, as official APIs like the Reddit API were unavailable at the time of processing due to restrictive access policies.

3.1 Training Data for Domain Adaptation (Gold Standard)

For model fine-tuning, a specialized, annotated dataset based on (Rahman et al., 2025) was used. The dataset is divided into:

- *WSB Full:* A comprehensive corpus with approximately 3,000 labeled samples.

- *WSB All-Agree (Gold Standard):* A subset of the full dataset with about 1,500 samples where multiple annotators agreed on the labels.

To measure model quality against particularly reliable labels, a correspondingly large part of the All-Agree dataset was reserved exclusively for validation and testing. The Full dataset was cleaned of these samples to prevent data leakage during training. The final split into training, validation, and test data was *stratified* to keep the class distribution consistent across all subsets.

3.2 Inference Data for Backtesting

Due to current restrictions of the Reddit API, the Kaggle dataset `WallStreetBets 2022` (Preda, 2025) was used. An exploratory analysis showed

that the dataset (contrary to its name) covers the period from **March 2022 to March 2025**. In addition to the text fields `title` and `body`, the dataset contains metadata such as `score` (upvotes) and `comms_num` (number of comments), which are later used to weight sentiment. Relevant posts regarding Nvidia were extracted using regex filters, resulting in approximately 19,000 samples.

Both training and inference data underwent an identical cleaning pipeline to reduce noise.

3.3 Financial Data

Nvidia stock price data was obtained via the `yfinance` interface for the same period. Daily adjusted close prices were used to correctly account for potential dividends and splits.

4 Methodology and System Architecture

The methodological approach aims to overcome the discrepancy between typically formal financial language models and the language used on social media platforms like WSB, as well as to establish a statistically robust correlation of daily aggregated sentiment signals with market data over a multi-year period.

4.1 System Overview and Architecture

The developed system follows a classic NLP pipeline, extended by specific modules for backtesting and time-series aggregation. The architecture is divided into two main components:

1. *Domain Adaptation:* Fine-tuning of Transformer models on a gold-standard dataset to capture specific WSB jargon.

2. *Backtesting with Lead-Lag Analysis:* Aggregation of discrete text classifications into time-series signals and correlation analysis with financial market data.

Figure 1 shows the overall architecture of the developed system.

4.2 Model Development and Domain Adaptation

The underlying hypothesis is that generic financial models like FinBERT, trained on formal news corpora, often misinterpret the semantic context of WSB-specific jargon. To verify this assumption, a comparative approach with four model configurations was chosen:

- *FinBERT (Base):* Use of the pre-trained model without adjustment as a baseline.
- *FinBERT (Adapted):* Fine-tuning on the WSB training dataset.
- *BERT Base (Adapted):* Baseline to isolate the effect of financial pre-training.
- *DeBERTa v3 (Adapted):* Modern state-of-the-art architecture with disentangled attention.

Training was performed using the HuggingFace `Trainer` API using `CrossEntropyLoss`. To avoid overfitting, early stopping based on the F1 score was employed.

4.3 Sentiment Aggregation and Signal Generation

Since the sheer volume of posts alone is often not meaningful and individual posts have widely varying visibility, a multi-stage aggregation algorithm was implemented. This transforms the probabilistic outputs of the language model into a daily signal, taking community resonance into account.

To model the visibility of a post i on day t, a weight w_i is calculated. This is based additively on the log-transformed interaction metrics (upvotes u_i and comments c_i) to dampen the dominance of viral outliers:

$$w_i = \ln(1 + u_i) + \ln(1 + c_i)$$

Instead of hard classification (positive/negative), the softmax probabilities $P(\text{pos} \mid d_i)$ and $P(\text{neg} \mid d_i)$ output by the model are used directly, where d_i denotes a single post and D_t the set of all posts on day t. The Net Sentiment $S_{\text{net},t}$ for day t results from the weighted difference of the probability masses:

$$S_{\text{net},t} = \frac{\sum_{i \in D_t} w_i \cdot \big(P(\text{pos} \mid d_i) - P(\text{neg} \mid d_i)\big)}{\sum_{i \in D_t} w_i}$$

To avoid overweighting days with a small data basis, the signal is scaled by the effective number of posts (N_{eff}). Since posts are weighted very differently, N_{eff} is calculated as:

$$N_{\text{eff},t} = \frac{\sum_{i \in D_t} w_i^2}{\left(\sum_{i \in D_t} w_i\right)^2}$$

The final sentiment signal S_t for time series analysis is finally calculated by a logarithmic damping of the effective quantity:

$$S_t = S_{\text{net},t} \cdot \ln(1 + N_{\text{eff},t})$$

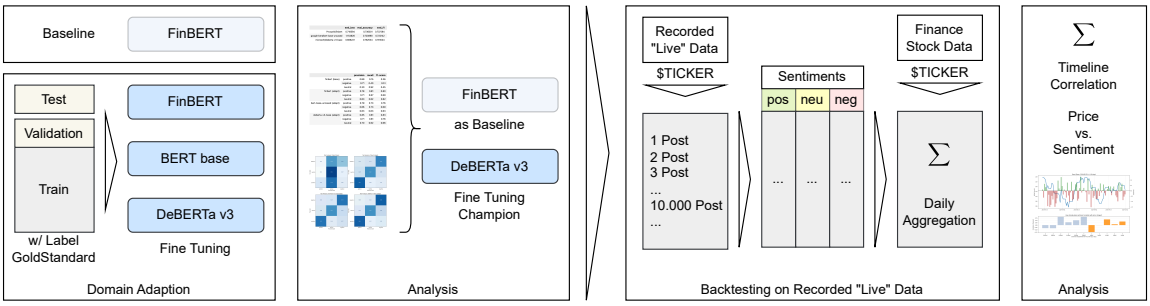

Figure 1: System overview and architecture of the NLP pipeline for Domain Adaptation and Backtesting with Correlation (Lead-Lag) Analysis.

	eval_loss	eval_accuracy	eval_f1
ProsusAI/finbert	0.718594	0.736301	0.737294
google-bert/bert-base-uncased	1.154626	0.720890	0.722142
microsoft/deberta-v3-base	0.668221	0.782534	0.781044

Figure 2: Validation metrics of the adapted models after convergence.

		precision	recall	f1-score
finbert (base)	positive	0.89	0.15	0.26
	negative	0.71	0.40	0.51
	neutral	0.30	0.92	0.45
finbert (adapt)	positive	0.78	0.81	0.80
	negative	0.71	0.67	0.69
	neutral	0.63	0.62	0.62
bert-base-uncased (adapt)	positive	0.79	0.73	0.76
	negative	0.66	0.73	0.69
	neutral	0.63	0.63	0.63
deberta-v3-base (adapt)	positive	0.85	0.81	0.83
	negative	0.71	0.81	0.76
	neutral	0.70	0.62	0.66

Figure 3: Test performance of model configurations. The gain in positive recall through domain adaptation is particularly evident.

4.4 Evaluation Strategy: Market Correlation

The calculated sentiment signal is compared with the daily closing price of the Nvidia stock. Evaluation is performed as a *Lead-Lag Correlation* by calculating the Pearson correlation coefficient for time shifts of -5 to $+5$ days to analyze potential causalities. Additionally, a detailed analysis of specific market phases is conducted for visual and statistical validation during high volatility.

5 Result 1: Domain Adaptation

First, it was necessary to quantify whether and what performance gain could be achieved through domain adaptation. This chapter presents the empirical results.

5.1 Training and Convergence

The three fine-tuned models were trained under identical hyperparameters (Learning Rate: 2e-5, Batch Size: 16). As the table in Figure 2 shows, DeBERTa-v3 achieved the best values for F1-Score and Accuracy on the validation set, followed by FinBERT (Adapt).

5.2 Quantitative Analysis

Evaluation on the test set reveals serious differences between the off-the-shelf approach and the adapted models (see Figure 3).

The unadapted FinBERT model shows a massive bias towards the *Neutral* class and fails almost completely in detecting positive WSB posts. As the confusion matrix in Figure 4 shows, the model incorrectly classifies a large extent of actually positive posts as neutral. For trading signal generation, this model is unusable as buy signals are systematically ignored.

Through fine-tuning, the recall for the positive class in FinBERT (Adapt) was significantly increased—an improvement by approximately a factor of 5. The separation between positive and negative also increased significantly.

DeBERTa v3 provided the most robust result in terms of F1-Score and recall. The confusion matrix in Figure 4 also shows that DeBERTa has the lowest rate of critical confusions (positive $\leftrightarrow$ negative).

5.3 Qualitative Analysis

To understand *why* the baseline fails, a qualitative analysis was conducted on samples where FinBERT (Base) was wrong with high confidence (>90%). The examples in Figure 5 illustrate the semantic misunderstanding of the standard model. FinBERT (Base) incorrectly classifies the phrase

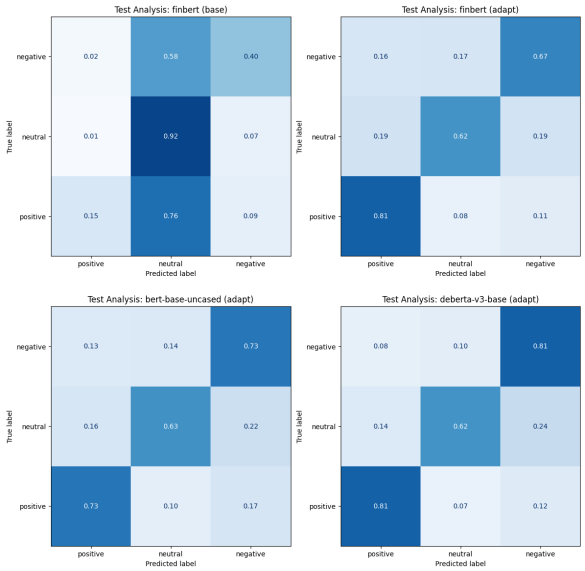

Figure 4: Confusion matrix of the four models on the gold-standard test set. Notable: FinBERT (Base) incorrectly classifies the majority of positive and negative samples as neutral.

	text	true	pred	conf
1174	EV maker Lordstown Down 60% from high with 77% short interest ripe for a short sqeeze pump	positive	negative	0.969016
59	Some simple math for y'all... SPY is down about 19% from its high. In the prior 20 years or so the market has had much larger declines. After the dotcom bust it declined about 42% and during the Housing bust it declined about 48%. A measly 19% drop off...	neutral	negative	0.959092
30	Looking at SPCE and CLOV price action, I think OPEN could be a multi-bagger next several weeks leading up to earnings. \~Doubled down on riding shares. No DD. Just a feeling.	positive	negative	0.955728
786	Get it on this dip. Dont get left behind! you'll be kickin yourselves. Gonna break $15 today! Has an earnings beat and gets downgraded!!!!! WHAT!? just another example of the street trying to kill THE REVOLUTION!!!!!	positive	negative	0.949151
1098	Retirement savings $214,000.00 on 14,000 shares of Lithium America! s con do DD and see just where this is going to go!	positive	neutral	0.937301

Figure 5: Examples of typical misclassifications where FinBERT (Base) was wrong with high confidence (>90%).

"Get it on this dip..." as negative, presumably reacting to the word "dip". However, in the WSB context, this signals buying opportunities. Expressions like *"ripe for a short squeeze pump"* are classified as negative (risk of loss) by the base model, while the community evaluates this as a maximally positive signal.

Interim Conclusion The results show that a direct application of FinBERT to social media without domain adaptation leads to systematic misinterpretations. For the subsequent backtesting (Section 6), DeBERTa v3 is considered, with FinBERT (base) serving as a baseline.

6 Result 2: Backtesting

After validating the models on static data (Section 5), this section focuses on recorded "live" data. The goal is to answer the central question: To what extent does the sentiment measured on r/wallstreetbets correlate with the actual price development of the Nvidia stock?

	Positive (%)	Negative (%)	Neutral (%)	Total (n)
finbert (base)	7.6	13.8	78.6	18891.0
finbert (adapt)	40.8	25.1	34.1	18891.0
bert-base-uncased (adapt)	31.8	31.1	37.1	18891.0
deberta-v3-base (adapt)	29.9	39.0	31.1	18891.0

Figure 6: Distribution of sentiment classes on the backtesting dataset (03/2022–03/2025).

6.1 Distribution of Sentiment Over Time

First, it was analyzed how the models classify the approximately 19,000 filtered posts. The results confirm the observations from the validation phase in Section 5 on unseen "live" data as well (see Figure 6).

While FinBERT (base) classifies almost 80% of discussions as neutral and thus irrelevant for signal generation, the adapted DeBERTa model shows a tendency towards negative sentiment, which may reflect the often cynical nature of WSB comments.

6.2 Global Correlation Analysis

To assess statistical significance, the time series of daily sentiment scores was correlated with Nvidia's percentage price changes (*returns*). Here, a lead-lag analysis was performed over a window of $t-5$ to $t+5$ trading days. Since comments naturally lag behind (main) posts, they were filtered out to reduce signal noise. The results in Figure 7 show clear differences in signal quality:

FinBERT Base (Baseline) shows significantly weaker sentiment signals and a rather weak correlation with a maximum at Lag 0 (same day) and Lag -1. Predictive power for future days (Lags > 0) is nonexistent.

DeBERTa v3 (Adapt) shows a significantly clearer signal. The correlation at Lag 0 rises higher than with FinBERT (base). A possible echo effect (Lag -1) is also depicted more clearly.

The fact that the correlation for the past (Lag < 0) is stronger than for the future (Lag > 0) means that WSB sentiment is primarily reactive—the community discusses price movements while or immediately after they occur. Predictive power cannot be derived from the volume of posts for the selected (large-cap) stock.

6.3 Event Study: Volatility in August 2024

To test model properties under stress conditions, an analysis was conducted for the time window

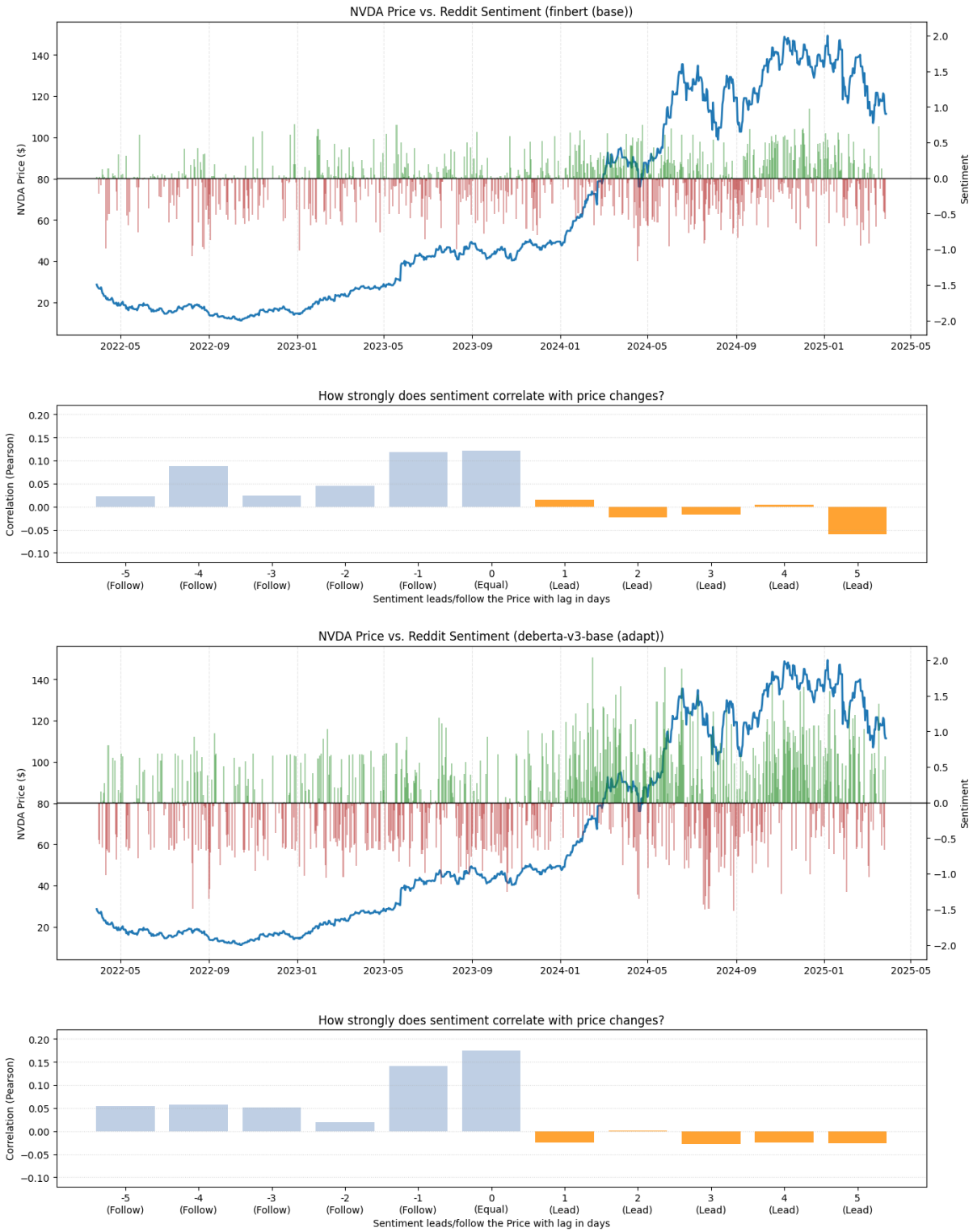

Figure 7: Global sentiment time series and lead-lag correlations for FinBERT Base (top) and DeBERTa v3 Adapt (bottom) over the backtesting period 03/2022–03/2025.

around August 22, 2024 (+/- 40 days).

Visual analysis (Figure 8) shows high precision of the adapted model. The price crash in early August (from approx. $129 to <$100) is accompanied by massive negative sentiment bars. With the price recovery starting around August 10, massive nervousness is evident; sentiment follows almost every price movement in both directions with very high signal strength (Lag 0). Immediately after the quarterly figures on August 28, sentiment flips back

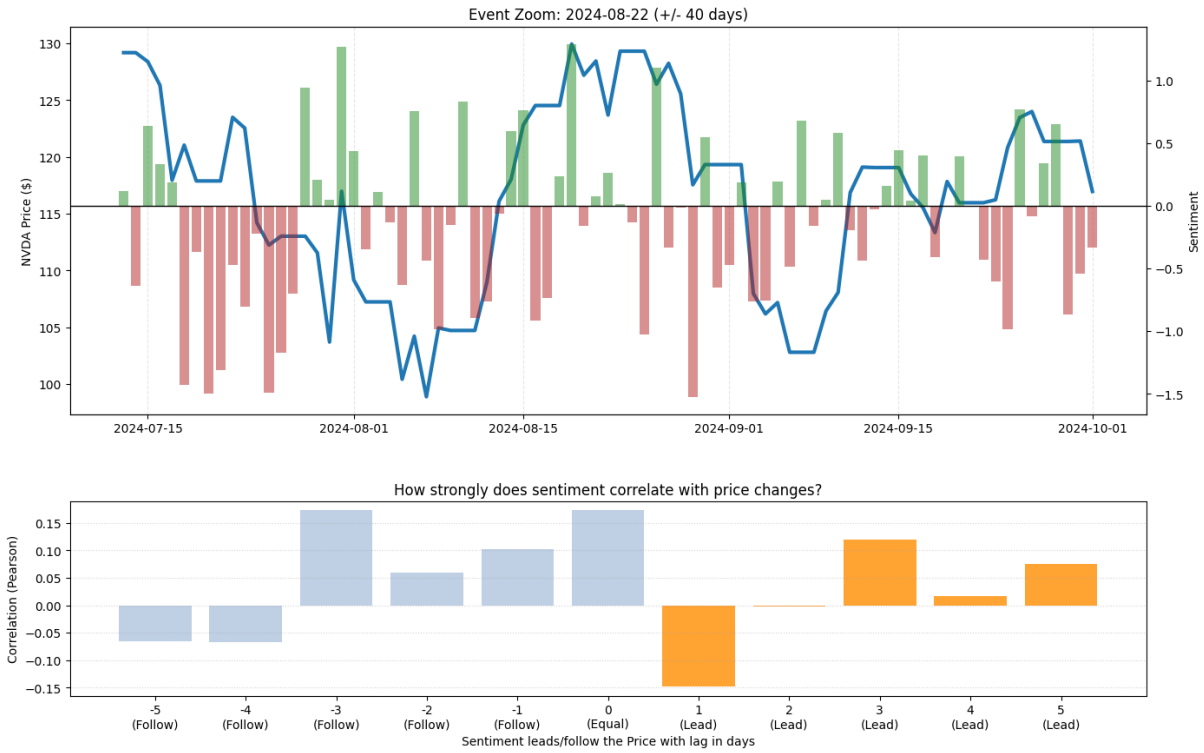

Figure 8: Zoom-in around earnings in August 2024 (DeBERTa v3 Adapt).

into the highly negative parallel to the price slide.

Particularly notable is the negative correlation at Lag 1 (Lead) in this phase. This implies that extremely positive sentiment on day t was frequently followed by a price loss on day $t + 1$.

7 Conclusion

The project aimed to quantify the predictive power of social media discussions on stock prices through a technically sound NLP pipeline. The results can be summarized in the following central theses, concerning both technical modeling and financial interpretation:

Domain Adaptation is Necessary The experiments clearly prove that generic financial language models are unsuitable for analyzing modern retail investors. The massive performance jump from FinBERT Base to DeBERTa v3 (Adapt) shows that the semantic gap between Reuters news and Reddit slang cannot be bridged without explicit fine-tuning.

Verification of Buy Signals The lead-lag analysis (Section 6.2) refutes, for the Nvidia example, the popular assumption that Reddit forums proactively drive the prices of large-cap companies. The stronger correlation at negative lags (t<0) means that the mass of commenters acts *reactively*: rising prices generate euphoria, falling prices panic. The forum lacks the capital power to function as a price driver for a large-cap like Nvidia. A valuable insight for algorithmic trading systems is provided by the event study in Section 6.3: The observed negative correlation at Lead-Lag 1 during peak phases indicates saturation behavior. Extremely positive sentiment often marks a local top, followed by immediate profit-taking.

Summary The project demonstrates that modern Transformer architectures like DeBERTa v3 are capable of extracting highly informal sentiment from financial forums through targeted fine-tuning. Technically, an end-to-end pipeline was realized that transforms raw, noisy web data into quantifiable time signals.

Focusing on Nvidia serves merely as an example. For stocks with small market capitalization and high short interest, the causal influence of sentiment could be significantly higher. The developed system serves in this context as a template for analyzing stocks (or other assets) of interest to r/wallstreetbets.

Limitations

The analysis was limited to text data. However, on r/wallstreetbets, fundamental information (e.g., portfolio screenshots, memes) is often communicated visually. A pure text model overlooks this significant part of the opinion spectrum.

References

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

Nicholas Barberis, Andrei Shleifer, and Robert Vishny. 1998. A model of investor sentiment. *Journal of Financial Economics*, 49(3):307–343.

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

Eugene F. Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Gabriel Preda. 2025. Wallstreetbets 2022.

A M Muntasir Rahman, Ajim Uddin, and Guiling "Grace" Wang. 2025. Evaluating financial sentiment analysis with annotators instruction assisted prompting: Enhancing contextual interpretation and stock prediction accuracy. *Preprint*, arXiv:2505.07871.

Robert J. Shiller. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review*, 71(3):421–436.

Timm O. Sprenger, Andranik Tumasjan, Philipp G. Sandner, and Isabell M. Welpe. Tweets and trades: the information content of stock microblogs. *European Financial Management*, 20(5):926–957.

Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.