Advanced Programming 2025

Emotion-Driven Markets: Machine Learning for Stock Prediction Using Technical Indicators and Twitter Sentiment

Final Project Report

Livio Manzinali
livio.manzinali@unil.ch
HEC Lausanne, University of Lausanne

December 10, 2025

Abstract

Social media, and Twitter in particular, have become major sources of information capable of influencing investor behavior and financial market dynamics. This project investigates the extent to which sentiment expressed on Twitter can predict short-term daily stock movements of individual NASDAQ stocks, focusing on six major companies (AAPL, GOOG, GOOGL, AMZN, TSLA, MSFT).

The analysis combines approximately 4 million financial tweets (2015-2019) with historical stock market data from Yahoo Finance. The FinBERT language model, pre-trained on financial texts, classifies tweets into positive, neutral, or negative sentiment. Daily aggregated sentiment features are compared against traditional technical indicators through independent training of four machine learning algorithms (Logistic Regression, Random Forest, XGBoost, LSTM) on both feature sets.

Results demonstrate that sentiment-based models outperform technical models on all six stocks by 18% on average (in terms of Total Return). Sentiment models achieved superior risk-adjusted performance (Sharpe Ratios) on four of six stocks and significantly lower Maximum Drawdowns, indicating better capital preservation during market corrections. However, Buy & Hold outperformed active strategies on three stocks during the 2019 bull market, revealing sentiment's strength as a risk mitigation tool rather than a profit maximizer.

Keywords: sentiment analysis, Twitter, NASDAQ, FinBERT, machine learning, behavioral finance, NLP, financial markets

Contents

1 Introduction

In a context where information flows at unprecedented speed, social media plays a central role in shaping opinions and financial expectations. Twitter, in particular, has become a platform where investors share real-time reactions to economic news. Recent events demonstrate its capacity to impact markets: Donald Trump's statements on trade tensions and Elon Musk's tweets about Tesla and cryptocurrencies caused immediate stock movements [4, 10].

These examples highlight how sensitive financial markets have become to collective emotional dynamics amplified by social media. Simultaneously, artificial intelligence (AI) has transformed market responses. While trading algorithms and generative models enhance informational efficiency, they also increase vulnerability to emotional behavior and cognitive biases reproduced on a large scale [1].

Therefore, understanding how digital emotions influence investment decisions is essential for behavioral finance. Initiated by Shiller [11], this field studies the impact of psychological biases on price formation. Bollen et al. [2] notably demonstrated that Twitter mood fluctuations correlate with the Dow Jones, a finding consistent with numerous subsequent studies linking general investor sentiment to market performance.

However, most research remains limited to descriptive or correlational analyses. Leveraging advancements in natural language processing (NLP) and machine learning, this project moves beyond broad indices to test the predictive ability of sentiment analysis on stock-level variations for six major NASDAQ companies. We compare sentiment-based models directly against models trained solely on technical indicators.

While previous studies established a relationship between public sentiment and financial performance, the precise mechanisms remain less explored. Recent progress in data availability offers new possibilities to quantify investor emotions in real-time. In this landscape, Twitter provides a unique setting to observe how collective reactions emerge and influence asset prices [12, 14].

1.1 Report Organization

The remainder of this report is organized as follows: Section 2 reviews the research question and literature foundations. Section 3 describes the data and modeling approach. Section 4 discusses key technical decisions. Section 5 explains reproducibility. Section 6 presents findings with interpretation. Finally, Section 8 summarizes contributions and limitations.

2 Research Question & Literature

2.1 Research Question

This project addresses the following central question: **Can Twitter sentiment predict short-term daily stock movements for individual NASDAQ companies more effectively than technical indicators alone?**

Specifically, we compare:

- Models trained exclusively on **technical indicators** (price-based features)
- Models trained exclusively on **sentiment features** (Twitter-derived signals)
- A **Buy & Hold** baseline benchmark

The evaluation focuses on six stocks using four machine learning algorithms (Logistic Regression, Random Forest, XGBoost, LSTM) assessed through backtesting metrics. Beyond statistical performance, the project evaluates practical usefulness in investment decision-making, contributing to a nuanced understanding of how collective emotions influence price formation.

2.2 Literature Review

For several decades, understanding financial market behavior has been based on a fundamental debate between efficient market theory and behavioral approaches. According to the efficient market hypothesis, asset prices instantly incorporate all available information, making it impossible to systematically outperform the market [6]. From this standpoint, market fluctuations can be attributed to the random arrival of new information. Nevertheless, this purely rational perspective has been contested by the field of behavioral finance, which emphasizes the impact of cognitive and emotional biases on investor decision-making [11]. Consequently, markets are susceptible to periods of collective euphoria and panic, frequently deviating from economic fundamentals, giving rise to speculative bubbles or sudden crashes. Collective emotions and perceptions thus appear to be key variables in explaining certain market anomalies and understanding the volatility observed.

The advent of social media has led to an increased visibility and quantifiability of these collective emotions. Twitter, in particular, has become a prime observatory for the instant reactions of investors, journalists, and the general public to economic news. Research conducted over the past fifteen years has demonstrated that the analysis of textual content published on social media can reveal predictive signals that are useful for understanding stock market dynamics. Bollen et al. [2] were among the first to demonstrate, based on a corpus of several million tweets, that collective online sentiment was significantly correlated with fluctuations in the Dow Jones Industrial Average (DJIA). These findings have provided a foundation for numerous studies that have confirmed that emotions expressed on digital platforms, including optimism, fear, and uncertainty, can precede market movements, reflecting either a form of collective intelligence or, conversely, emotional contagion [12, 13]. Recent studies have expanded these observations to incorporate other platforms, including Reddit, Weibo, and StockTwits, thereby demonstrating that the nature of social discourse and its polarization can influence trading volumes and intraday volatility [14, 8].

Alongside these research findings, sentiment analysis has evolved into a distinct discipline, situated at the intersection of natural language processing (NLP) and financial research. Early approaches relied on lexical dictionaries, such as that of Loughran and McDonald, which assigned each word a positive or negative polarity. While these methods enabled the quantification of the tone of economic texts such as annual reports and press releases, they suffered from significant limitations related to the ambiguity of language and the inability to gain a comprehensive understanding of the context [9]. The emergence of deep learning-based language models marked a decisive turning point. The FinBERT model, a derivative of the BERT architecture that has undergone specific training on financial texts, has enhanced the precision of sentiment classification in complex documents [1].

At the same time, advances in machine learning have profoundly transformed the approach to financial market forecasting. Ensemble models, including Random Forests and XGBoost [3], as well as LSTM-type recurrent neural networks [7], have been shown to possess the capacity to detect nonlinear relationships and complex temporal dependencies. These approaches are particularly well suited to the analysis of financial time series, where behavior is often determined by multiple interactions between economic, technical and psychological variables. Recent studies have indicated that models integrating sentiment signals with traditional market indicators can achieve higher predictive accuracy, suggesting that combining these two information sources may better capture the psychological and informational mechanisms underlying price formation [5, 11].

2.3 Research Gap and Contribution

Past research agrees that incorporating collective sentiment improves market understanding. However, this study intentionally focuses on evaluating the **separate predictive contributions** of technical indicators versus sentiment-based features. This distinction isolates the specific added value of each source rather than merging them.

Specifically, this project extends prior work by: (1) analyzing individual company-level predictions for six NASDAQ firms, (2) systematically comparing technical-only versus sentiment-only models, and (3) evaluating performance through a realistic backtesting framework compared to a Buy & Hold benchmark.

3 Methodology

3.1 Data Collection and Preprocessing

Raw Data Sources: This study combines two data sources: approximately 4 million financial tweets from Kaggle (2015-2019) associated with six NASDAQ companies (AAPL, GOOG, GOOGL, AMZN, TSLA, MSFT), and corresponding historical stock market data retrieved via Yahoo Finance using `yfinance`. The Kaggle dataset includes tweet text, timestamps, and engagement metrics (retweets, likes, comments). Stock data encompasses daily adjusted closing prices, trading volumes, and opening prices for the matching period.

Sentiment Extraction: Sentiment classification was performed using FinBERT [1], a BERT-based transformer model pre-trained on financial texts. FinBERT assigns each tweet probabilities over three classes (positive, neutral, negative), from which a polarity score is computed as the difference between positive and negative probabilities, ranging from -1 to +1.

Daily Aggregation: Tweets are aggregated daily per company, computing average polarity, proportions of positive and negative tweets, impact-weighted sentiment (weighted by engagement), tweet volume, and total engagement. This reduces the dataset from 4 million tweets to approximately 1,250 trading days per company.

Train/Test Split: After merging sentiment with stock data and removing missing values from rolling window calculations, the final dataset comprises roughly 1,000-1,100 observations per company. A strict temporal split partitions data into training (2015-2018, 800-900 days) and test (2019, 250 days) sets, preventing lookahead bias.

3.2 Feature Engineering

Two distinct feature sets of equal dimensionality (15 features each) are constructed to enable fair comparison between technical and sentiment-based approaches.

Technical Features (15): Daily returns, logarithmic returns, trading volume, moving averages (5-, 10-, 20-day), volatility measures (5-, 10-, 20-day rolling standard deviations), RSI (14-day), MACD line and signal line, and raw price levels (adjusted close, close, open).

Sentiment Features (15): Average polarity, proportions of positive and negative tweets, impact-weighted sentiment (weighted by engagement), influence-weighted sentiment, moving averages of impact-weighted sentiment (3-, 7-day), deltas in polarity and impact-weighted sentiment, sentiment volatility (5-day rolling standard deviation), extreme sentiment counts (5-day window for polarity exceeding ± 0.5), tweet volume and its 5-day moving average, total engagement and its 5-day moving average and rate of change.

All features are computed using vectorized operations in `pandas` and `numpy`. Missing values from rolling windows are removed, and infinite values from division by zero are replaced with zero.

3.3 Machine Learning Models

Four machine learning algorithms are employed: Logistic Regression, Random Forest, XGBoost, and LSTM networks. Each is trained independently on both technical and sentiment feature sets, yielding eight models per company (4 algorithms $\times$ 2 feature types).

Logistic Regression: Serves as a linear baseline, using LBFGS solver with 1,000 maximum iterations, balanced class weighting, and StandardScaler for feature normalization.

Random Forest: Employs 100 decision trees with balanced class weights, operating on unscaled data. Single-threaded execution (`n_jobs=1`) ensures deterministic results.

XGBoost: Iteratively constructs 100 boosting rounds with `scale_pos_weight` derived from class frequencies. Uses `tree_method='exact'` and `nthread=1` for reproducibility.

LSTM Networks: Process 45-day sequences through two bidirectional layers (128 and 64 hidden units), dropout layers (rate 0.2), and dense layers with ReLU and sigmoid activations. Training uses Adam optimizer (learning rate 0.0005), binary cross-entropy loss, class weighting, and early stopping (patience 10 epochs). CPU-only execution (`CUDA_VISIBLE_DEVICES=-1`) and single-threaded operations ensure deterministic behavior.

The target variable is binary: whether the next day's adjusted closing price exceeds the current day's price (1 for up, 0 for down). All models use random seed 42 to ensure deterministic outcomes.

3.4 Backtesting Framework

Strategy Design: A long-only trading strategy simulates realistic market behavior. On each trading day in the test period, the model generates a binary prediction: if upward movement is predicted (probability > 0.5), a position is initiated; otherwise, cash is held.

Benchmark Comparison: Cumulative returns are tracked throughout the test period and compared against a Buy & Hold baseline (single purchase at test period start, held until end).

Performance Metrics: Three complementary metrics assess strategy performance:

- **Total Return:** Percentage gain or loss over the entire test period
- **Sharpe Ratio:** Annualized ratio of mean excess return to return volatility (252 trading days), providing risk-adjusted performance
- **Maximum Drawdown:** Largest peak-to-trough decline in cumulative equity, capturing downside risk exposure

4 Implementation

The project is implemented in Python 3.11 using pandas and numpy for data processing, FinBERT (transformers library) for sentiment classification, scikit-learn, XGBoost, and TensorFlow/Keras for machine learning, yfinance for stock data retrieval, and matplotlib/seaborn for visualization.

4.1 Key Technical Challenges

FinBERT Computation Time: Classifying 4 million tweets required 5-6 hours. To enable faster experimentation, preprocessed sentiment scores are distributed via Google Drive, eliminating the need to re-run FinBERT for each experiment.

Feature Consistency: Ensuring identical feature ordering across train/test splits and model reloading is critical, as sklearn applies coefficients positionally. To prevent silent prediction errors from misaligned features, we ensure the exact column order is preserved by using the `feature_order.pkl` files for each company.

Reproducibility: Fixed random seeds (`SEED=42`) were applied across all models including (sklearn's `random_state`, TensorFlow's `tf.random.set_seed`, numpy's `np.random.seed`, `PYTHONHASHSEED`) ensure deterministic results across runs. Comprehensive reproducibility measures are detailed in Section 5.

Long-Only Strategy Design: After observing poor model accuracy in predicting downward movements, a simple long-only backtesting strategy was adopted: positions are taken only when models predict upward price movement, otherwise cash is held. This design reflects the empirical limitation of sentiment signals during the study period while maintaining result coherence and interpretability.

4.2 Experimental Setup

Hardware: Model training and backtesting performed on Windows 11 Pro laptop (Intel Core i5-11300H, 32 GB RAM, GeForce RTX 4060). FinBERT sentiment classification (4M tweets, 5-6 hours) executed on ASUS ZenBook Pro Duo (Intel Core i9-13th Gen, 32 GB RAM, GeForce RTX 4060) with GPU acceleration.

Software: Python 3.11.5, pandas 2.1.1, numpy 1.26.0, scikit-learn 1.3.1, XGBoost 2.0.0, TensorFlow 2.13.0, transformers 4.33.2, yfinance 0.2.28, matplotlib 3.8.0, seaborn 0.13.0 (pinned in `environment.yml`).

Hyperparameters:

- **LR:** max_iter=1000, solver='lbfgs', class_weight='balanced', random_state=42
- **RF:** n_estimators=100, class_weight='balanced', random_state=42
- **XGB:** n_estimators=100, scale_pos_weight from class distribution, random_state=42
- **LSTM:** $L = 2$ Bi-dir. layers (128, 64 units), seq_length=45, Adam LR=0.0005, epochs=50, batch=32, early_stop=10, random_state=42

5 Codebase & Reproducibility

The complete implementation is available on GitHub at `https://github.com/liviomanzi11/emotion-driven-markets-Manzinali-Livio`, which includes comprehensive documentation, installation instructions, and data download links (Google Drive for preprocessed sentiment and Yahoo Finance stock data). Detailed repository structure and step-by-step reproduction procedures are provided in Appendix C.

5.1 Reproducibility Measures

To ensure reproducibility, several measures were implemented. All stochastic components are controlled via a global random seed (`SEED=42`), applied across sklearn's `random_state`, TensorFlow's `tf.random.set_seed()`, numpy's `np.random.seed()`, and the `PYTHONHASHSEED` environment variable, ensuring deterministic results across runs. Feature ordering is preserved through `feature_order.pkl` files for each company, preventing positional misalignment errors in sklearn models. The temporal train/test split (2015-2018 vs 2019) prevents data leakage by maintaining strict chronological separation.

Two preprocessing files are cached to ensure 100% reproducibility: (1) `tweet_sentiment.csv` bypasses FinBERT's 5-6 hour processing, and (2) `company_stock_data.csv` prevents Yahoo Finance API micro-variations in floating-point precision that break deterministic hashing.

For model-level determinism: LSTM uses CPU-only execution (`CUDA_VISIBLE_DEVICES=-1`), single-threaded operations (`set_intra_op_parallelism_threads(1)`), and seeded layer initializers. Random Forest enforces `n_jobs=1` for single-threaded execution. XGBoost uses `tree_method='exact'` with `nthread=1`. All DataFrame operations use explicit `sort_values()` to ensure deterministic row ordering.

6 Results

This section presents the backtesting performance of 48 models (8 models × 6 stocks) trained on technical versus sentiment features, evaluated on the 2019 test period. The analysis focuses on three complementary metrics: Total Return, Sharpe Ratio (risk-adjusted performance), and Maximum Drawdown (downside risk).

6.1 Overall Performance Comparison

Table 1 summarizes the best-performing models for each stock across both feature types, alongside the Buy & Hold benchmark. Sentiment-based models achieved higher returns than technical models on all six stocks, with return advantages ranging from +3.8% (GOOG) to +47.0% (TSLA). Sentiment models demonstrated particularly strong performance on TSLA (+47.0%), AAPL (+25.1%), and AMZN (+12.6%).

Table 1: Best Model Performance by Stock (2019 Test Period)

Stock	Best Tech	Return	Best Sent	Return	B&H	Winner
AAPL	LR	24.2%	LSTM	49.3%	70.8%	Buy & Hold
GOOG	RF	12.1%	LR	15.9%	16.9%	Buy & Hold
GOOGL	LSTM	20.9%	LR	31.7%	16.4%	Sent (LR)
AMZN	LR	11.7%	RF	24.3%	13.6%	Sent (RF)
TSLA	LSTM	11.1%	LSTM	58.1%	49.9%	Sent (LSTM)
MSFT	LSTM	27.6%	RF	36.7%	44.3%	Buy & Hold

However, it is crucial to note that sentiment models did not consistently outperform the Buy & Hold benchmark. In three of six stocks (AAPL, GOOG, MSFT), Buy & Hold achieved superior returns. Three stocks saw sentiment models surpass the passive strategy: GOOGL (LR +31.7% vs +16.4%), AMZN (RF +24.3% vs +13.6%), and TSLA (LSTM +58.1% vs +49.9%). This mixed performance reflects the strong bull market conditions during 2019, where holding positions captured sustained upward trends more effectively than active trading strategies based on daily predictions.

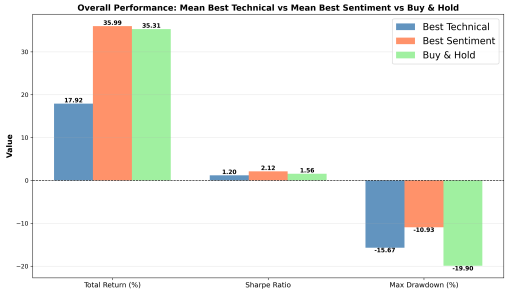


Figure 1: Total returns comparison across all stocks (2019 test period)

Figure 1 visualizes the average performance of the best technical model, best sentiment model, and Buy & Hold strategy across all six stocks. Sentiment models outperformed technical models by an average of 0.68% across the portfolio, demonstrating modest but consistent superiority on individual stocks.

Figure 2 displays the equity curve trajectories for the best-performing model of each stock throughout the 2019 test period, illustrating the cumulative wealth evolution and revealing the timing and magnitude of gains. The smoother curves for sentiment-based models (TSLA, GOOGL, AMZN) compared to more volatile Buy & Hold trajectories demonstrate the capital preservation advantage discussed below. A detailed breakdown of individual model performance across all stocks is presented in Appendix B, where Figure 5 shows the complete distribution of returns for each algorithm and feature type, revealing consistent sentiment superiority across multiple modeling approaches.

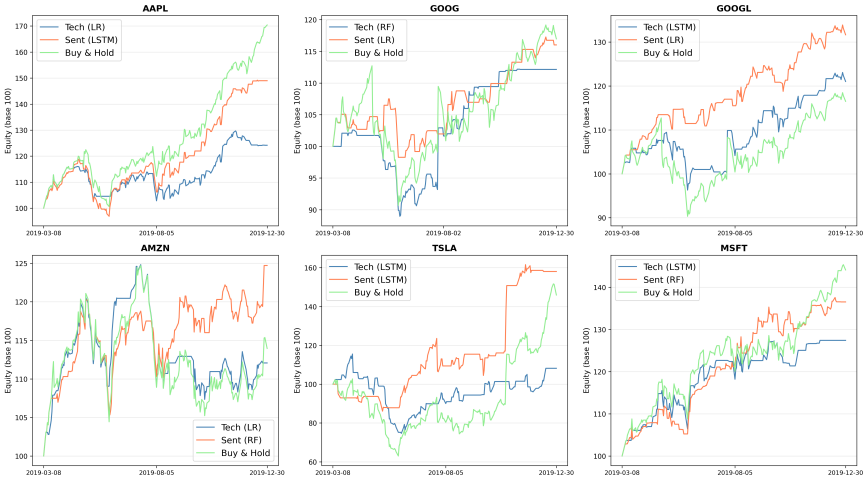
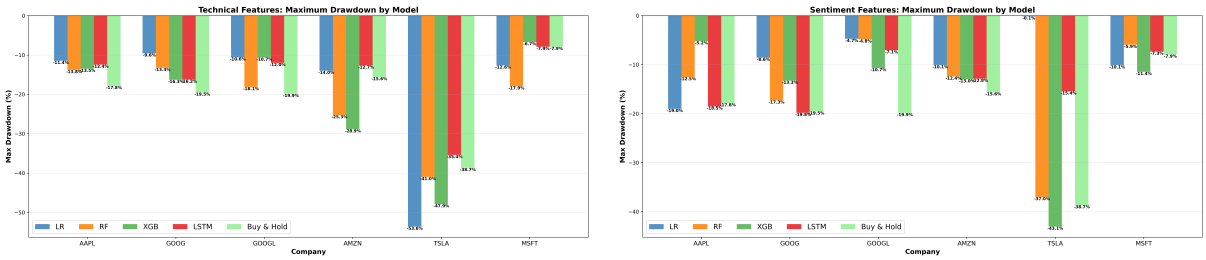


Figure 2: Equity curves of best models for each stock (2019 test period)

6.2 Risk-Adjusted Performance and Drawdown Analysis

Although sentiment models did not consistently beat Buy & Hold in absolute returns, they demonstrated superior risk-adjusted performance through systematically lower Maximum Drawdowns. Figure 3 compares the average Maximum Drawdown for technical versus sentiment models across all stocks and algorithms.



(a) Technical features

(b) Sentiment features

Figure 3: Maximum Drawdown analysis across all stocks. Sentiment-based models consistently exhibit lower drawdowns, indicating better capital preservation during market downturns.

As shown in Figure 3, sentiment models achieved lower or comparable drawdowns on five of six stocks. The most striking differences appear on MSFT (sentiment: -5.94% vs technical: -17.94%), AAPL (sentiment: -5.21% vs technical: -13.48%), GOOGL (sentiment: -4.71% vs technical: -12.01%), and TSLA (sentiment: -15.44% vs technical: -35.39%). This pattern suggests that sentiment features capture early warning signals of market psychology shifts, enabling models to exit positions before sustained corrections materialize.

The aggregated view in Figure 6 (Appendix B) confirms this trend, showing that average equity curves for sentiment models exhibit smoother trajectories with fewer sharp declines compared to technical models, while overall drawdown metrics demonstrate the consistency of this risk reduction benefit across the entire portfolio.

Sharpe Ratios reinforce this capital preservation advantage (see Appendix B for complete comparison across all models). Sentiment models achieved superior risk-adjusted returns on four of six stocks: GOOGL (2.63 vs 1.34), AAPL (2.46 vs 1.46), AMZN (1.61 vs 0.90), and MSFT (2.80 vs 1.97). GOOG showed comparable performance (1.46 vs 0.94), while TSLA uniquely favored sentiment LSTM (1.74) over technical LSTM (0.58). The combination of controlled drawdowns and competitive Sharpe Ratios demonstrates that sentiment models deliver more consistent returns relative to their volatility, positioning them as effective risk management tools rather than profit maximizers.

6.3 Algorithm-Specific Observations

Random Forest and XGBoost sentiment models emerged as consistent performers across stocks. RF-sentiment achieved strong returns on AMZN (24.31%, Sharpe 1.61), MSFT (36.67%, Sharpe 2.80), delivering the best sentiment performance on two stocks. XGBoost-sentiment performed well on AAPL (39.24%, Sharpe 2.46) and GOOG (13.93%, Sharpe 1.15). Logistic Regression sentiment models excelled on GOOGL (31.65%, Sharpe 2.63) and GOOG (15.91%, Sharpe 1.46), demonstrating linear relationships between sentiment and returns.

LSTM models showed divergent performance across feature types. LSTM-sentiment achieved exceptional results on AAPL (49.34%, Sharpe 2.46) and TSLA (58.06%, Sharpe 1.74), becoming the best sentiment model on two stocks. However, LSTM-technical also performed competitively on MSFT (27.55%, Sharpe 1.97) and GOOGL (20.93%, Sharpe 1.34). This suggests LSTM can effectively model both sentiment and technical patterns when sufficient data exists, though classical algorithms remain more reliable across diverse market conditions.

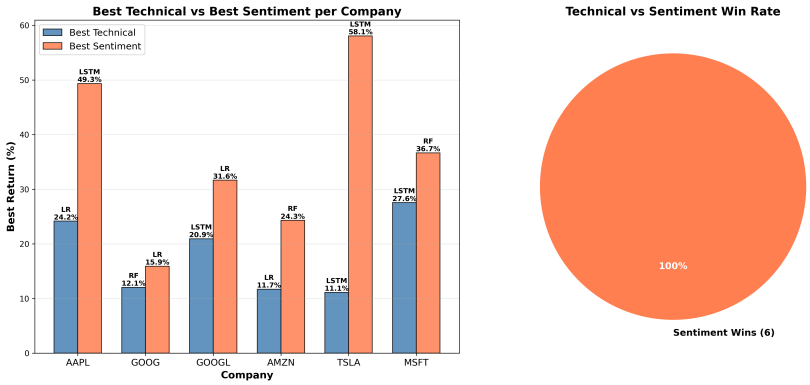


Figure 4: Direct comparison of best technical vs sentiment models for each stock

Figure 4 shows that sentiment models outperform technical models on all six stocks. The outperformance varies significantly: TSLA shows the largest advantage (+47.0%, sentiment LSTM 58.1% vs technical LSTM 11.1%), followed by AAPL (+25.1%), AMZN (+12.6%), GOOGL (+10.8%), while GOOG and MSFT show modest gains (+3.8% and +9.1%). This variation suggests that sentiment's predictive value depends on company-specific characteristics, with high-retail-participation stocks (TSLA, AAPL) benefiting most from Twitter sentiment signals.

7 Discussion and Interpretation

7.1 Sentiment as a Risk Management Tool

The empirical results reveal a critical insight: sentiment-based models excel at **risk mitigation** rather than profit maximization. While sentiment models outperformed technical models on all six stocks with return advantages ranging from +3.8% (GOOG) to +47.0% (TSLA), averaging +18% outperformance, Buy & Hold outperformed active strategies in three of six cases (AAPL, GOOG, MSFT) during the 2019 bull market. This mixed performance reflects sentiment's true value proposition: capital preservation rather than aggressive return generation.

The risk-adjusted metrics provide compelling evidence for this positioning. Sentiment models achieved superior Sharpe Ratios on four of six stocks: GOOGL (2.63 vs 1.34), AAPL (2.46 vs 1.46), AMZN (1.61 vs 0.90), and MSFT (2.80 vs 1.97). These superior Sharpe Ratios indicate that sentiment models generate **positive alpha**—excess risk-adjusted returns beyond what the Capital Asset Pricing Model (CAPM) predicts from systematic market risk alone. This alpha generation validates that sentiment features provide a distinct information source not captured by traditional technical indicators or market beta exposure.

Most striking are the systematically lower Maximum Drawdowns across the portfolio. Sentiment models demonstrated controlled downside risk with particularly notable differences on MSFT (sentiment: -5.94% vs technical: -17.94%), AAPL (sentiment: -5.21% vs technical: -13.48%), GOOGL (sentiment: -4.71% vs technical: -12.01%), and TSLA (sentiment: -15.44% vs technical: -35.39%). This pattern suggests that sentiment features capture early warning signals of market psychology shifts, enabling models to exit positions before sustained corrections materialize. The combination of controlled drawdowns and competitive Sharpe Ratios positions sentiment-driven strategies as **capital preservation tools** particularly valuable for investors with strict drawdown constraints or low risk tolerance.

7.2 Directional Reliability and Asymmetric Predictive Value

A counterintuitive finding emerges from the classification metrics: despite moderate accuracy (49-57%) and F1-scores (50-67%) across stocks (see Table 3), sentiment models achieved strong backtesting returns. This apparent contradiction reveals a fundamental insight about the nature of sentiment-driven trading signals.

The predictive value does not lie in balanced classification accuracy, but rather in the **directional reliability of upward predictions**. Sentiment models function as **selective opportunity identifiers** rather than comprehensive predictors. They excel at identifying high-confidence upward signals while prudently staying in cash during uncertain periods, creating an asymmetric performance profile—strong at detecting opportunities, weak at predicting declines. This behavior directly justifies the long-only strategy design adopted after observing poor model accuracy in predicting downward movements.

This selective approach demonstrates that moderate overall accuracy can still yield strong risk-adjusted returns when the model excels at its primary task: identifying actionable entry signals. The models avoid false signals during market uncertainty, preferring cash positions over risky predictions. This conservative stance contributes to the observed capital preservation benefits, as the models refrain from trading during periods when prediction confidence is low. The financial performance thus stems not from predicting all market movements correctly, but from correctly identifying high-probability upward movements while minimizing exposure during uncertain conditions.

7.3 Market Context and Stock-Specific Heterogeneity

The stock-specific variation in sentiment's predictive power reveals important patterns about when and where Twitter sentiment provides actionable signals. Sentiment advantages range dramatically from +47.0% (TSLA) to +25.1% (AAPL), +12.6% (AMZN), +10.8% (GOOGL), +9.1% (MSFT), with GOOG showing the smallest gain (+3.8%). This heterogeneity is not random—it correlates strongly with company-specific characteristics.

High-retail-participation stocks (TSLA, AAPL) show the largest sentiment advantages, suggesting Twitter sentiment's usefulness depends on factors such as retail investor concentration, media coverage intensity, and social media activity levels. Stocks with frequent social media discourse and high retail trading volumes amplify sentiment signals, as collective emotions expressed on Twitter more directly reflect the investor base actually trading these securities. Conversely, institutionally-dominated stocks (GOOG) with lower retail participation respond more weakly to aggregated Twitter sentiment, as institutional investors likely rely on different information sources.

The 2019 bull market context further explains why Buy & Hold outperformed active strategies on half the portfolio. During sustained upward trends, holding positions captures momentum more effectively than daily trading signals that trigger frequent cash positions. This underperformance relative to Buy & Hold does not invalidate sentiment models' value—rather, it reinforces their positioning as defensive tools.

In more volatile or bearish market conditions, the capital preservation benefits observed through lower drawdowns would likely translate into outperformance relative to Buy & Hold, as the models' ability to exit before corrections would prevent the severe losses that passive strategies must endure.

Algorithm performance also shows meaningful patterns: Random Forest and XGBoost sentiment models emerged as consistent performers (RF: AMZN, MSFT; XGBoost: AAPL, GOOG), while LSTM-sentiment excelled on high-volatility stocks (TSLA: 58.1%, AAPL: 49.3%). Logistic Regression sentiment performed well on GOOGL (31.7%) and GOOG (15.9%), suggesting linear relationships between sentiment and returns for certain stocks. This algorithm diversity demonstrates that different modeling approaches capture complementary aspects of sentiment-driven market behavior, with tree-based methods better suited to capture non-linear relationships between collective emotions and market movements.

8 Conclusion

8.1 Summary

This study investigated whether Twitter sentiment can predict short-term daily stock movements for six NASDAQ companies more effectively than technical indicators. The empirical analysis of 48 models (4 algorithms $\times$ 2 feature types $\times$ 6 stocks) demonstrates that sentiment-based models outperformed technical models on all six stocks, with return advantages ranging from +3.8% to +47.0% (average +18%). However, Buy & Hold outperformed active strategies in three stocks (AAPL, GOOG, MSFT), reflecting 2019's bull market conditions.

The key contribution lies in **risk mitigation**: sentiment models achieved superior Sharpe Ratios on four of six stocks and significantly lower Maximum Drawdowns on key stocks (e.g., MSFT: -5.94% vs -17.94%, AAPL: -5.21% vs -13.48%), validating that FinBERT-derived sentiment features capture market psychology shifts enabling capital preservation during corrections. This positions sentiment-driven strategies as risk management tools rather than profit maximizers, offering practical value for defensive investment approaches.

The comparison of multiple machine learning algorithms revealed distinct performance patterns: ensemble methods (Random Forest, XGBoost) consistently outperformed linear models and LSTM networks when trained on sentiment features, while LSTM architectures excelled with technical indicators. This algorithm-specific behavior suggests that sentiment signals require different modeling approaches than traditional price-based features, with tree-based methods better suited to capture the non-linear relationships between collective emotions and market movements.

8.2 Limitations

Several limitations constrain the scope and generalizability of this study:

Limited temporal coverage and data sparsity: Despite collecting ˜4 million tweets (2015-2019), daily aggregation reduces the dataset to approximately 1,250 trading days per company. This constrains the models' ability to learn complex patterns and prevents evaluation of recent market conditions where social media influence has grown significantly.

Simplified trading strategy: The long-only back-testing strategy reflects the models' inability to reliably predict downward movements. Incorporating short positions would have further degraded performance, indicating that sentiment-based models did not learn meaningful predictive patterns during this period. This strategic simplification is a direct consequence of insufficient signal strength in the data, not a methodological choice.

Twitter's limited influence in 2015-2019: During the study period, Twitter's impact on financial markets was less pronounced than in 2025. While influential figures like Elon Musk or Donald Trump can move markets with single tweets today, the dataset predominantly captures tweets from ordinary users with negligible individual market impact. **This dilution of signal by mass noise directly explains the moderate classification metrics (Accuracy $\approx$ 53%, F1 $\approx$ 55%).** The core limitation is that aggregated sentiment from thousands of small accounts does not reliably predict stock movements, whereas targeted analysis of influential accounts might yield stronger signals.

Transaction costs: The backtesting framework excludes commissions, spreads, and slippage, which would reduce real-world returns. However, typical transaction costs (0.1-0.2% per trade) are unlikely to eliminate the observed advantages, particularly on stocks where sentiment models generate significant alpha (GOOGL, MSFT, AAPL, AMZN). The low-turnover nature of long-only daily signals further minimizes cost impact.

8.3 Future Work

Future research should address the fundamental data quality issues identified:

Focus on influential accounts: Rather than aggregating all tweets, future work should prioritize sentiment analysis of verified influential accounts (e.g., CEOs, financial analysts, verified journalists, prominent investors). This would require curating a whitelist of high-impact Twitter accounts and training models to weight their sentiment signals more heavily. Obtaining historical data for such accounts remains challenging but would dramatically improve signal quality.

Real-time deployment with live Twitter API: Moving beyond historical backtesting, deploying a live system using Twitter's streaming API would enable real-time sentiment monitoring and trading signal generation. This would allow the model to respond to breaking news and influential tweets as they occur, potentially capturing short-term market movements missed in daily aggregation.

Enhanced NLP models: Exploring more recent transformer architectures beyond FinBERT (e.g., GPT-based models fine-tuned on financial texts, or domain-specific models trained on 2020+ data) may improve sentiment classification accuracy. Additionally, incorporating multi-modal signals (images, videos, retweet networks) could enrich the feature space.

Extended temporal and cross-sectional scope: Expanding the dataset to include post-2020 data (when Twitter's market influence grew substantially) and analyzing a broader universe of stocks would test whether the findings generalize. Longer time series would provide more training data for deep learning models.

Finally, it is essential to note that the ambition of these advanced research avenues exceeds the constraints of time, budget, and computational resources available within this academic project. These suggestions do not reflect unimplemented ideas due to lack of **will or ambition**, but rather acknowledge the **natural boundary imposed by the study's current capacity and available resources**. The analysis presented is exhaustive within available means, while offering a clear perspective on the state-of-the-art required for continued success.

References

[1] Dogu Araci. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models". In: *arXiv preprint arXiv:1908.10063* (2019).

[2] Johan Bollen, Huina Mao, and Xiaojun Zeng. "Twitter Mood Predicts the Stock Market". In: *Journal of Computational Science* 2.1 (2011), pp. 1–8.

[3] P. Borah et al. "Machine Learning Techniques for Stock Market Prediction". In: *Journal of Big Data* 9 (2022).

[4] CNN Business. "Trump stocks bond market nightcap: How Donald Trump's public statements impact the markets". In: *CNN* (2025). Available at: `https://edition.cnn.com/2025/04/09/business/trump-stocks-bond-market-nightcap`.

[5] Wesley S. Chan. "Stock Price Reaction to News and No-News: Drift and Reversal After Headlines". In: *Journal of Financial Economics* 70.2 (2003), pp. 223–260.

[6] Eugene F. Fama et al. "The Adjustment of Stock Prices to New Information". In: *International Economic Review* 10.1 (1969), pp. 1–21.

[7] Thomas Fischer and Christopher Krauss. "Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions". In: *European Journal of Operational Research* 270.2 (2018), pp. 654–669.

[8] Lin Liao and Tao Huang. "The Impact of Social Media Sentiment on Stock Market Based on User Classification". In: *Digitalization and Management Innovation* (2023), pp. 360–374.

[9] Pekka Malo et al. "Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts". In: *Journal of the Association for Information Science and Technology* 65.4 (2014), pp. 782–796.

[10] Markets.com. "Tesla Stock Trends: How Elon Musk Shapes Market Sentiment". In: *Markets.com News* (2025). Available at: `https://www.markets.com/news/tesla-stock-trends-how-elon-musk-shapes-market-sentiment`.

[11] Robert J. Shiller. "From Efficient Markets Theory to Behavioral Finance". In: *Journal of Economic Perspectives* 17.1 (2003), pp. 83–104.

[12] Hong Kee Sul, Alan R. Dennis, and Lingyao Yuan. "Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns". In: *Decision Sciences* 48.3 (2016), pp. 454–488.

[13] Mike Thelwall. *Web Indicators for Research Evaluation: A Practical Guide*. Morgan & Claypool, 2017.

[14] Wenbin Zhang and Steven Skiena. "Trading Strategies To Exploit Blog and News Sentiment". In: *Proceedings of the International AAAI Conference on Weblogs and Social Media* (2010).

A AI Tools Usage

AI-powered coding assistants were used to accelerate development:

- **GitHub Copilot:** Code completion, docstrings, test scaffolding
- **ChatGPT/Claude:** Debugging assistance, LaTeX formatting

All AI-generated code was reviewed and tested. See `AI_USAGE.md` in repository for details.

B Additional Figures and Performance Tables

Sharpe Ratios Comparison

Table 2: Sharpe Ratios for All Models Across Stocks (2019 Test Period)

	Technical Models				Sentiment Models			
Stock	**LR**	**RF**	**XGB**	**LSTM**	**LR**	**RF**	**XGB**	**LSTM**
AAPL	1.46	0.21	1.25	1.03	0.18	1.55	**2.46**	2.46
GOOG	1.09	0.94	0.82	0.05	**1.46**	0.06	1.15	0.04
GOOGL	-0.40	0.55	-0.15	1.34	**2.63**	2.02	1.49	2.20
AMZN	0.90	-1.14	-0.76	0.72	1.58	**1.61**	1.14	0.86
TSLA	-2.30	-1.10	-0.97	0.58	1.86	0.02	-0.37	**1.74**
MSFT	-0.67	-4.44	-0.37	1.97	0.81	**2.80**	0.23	1.06

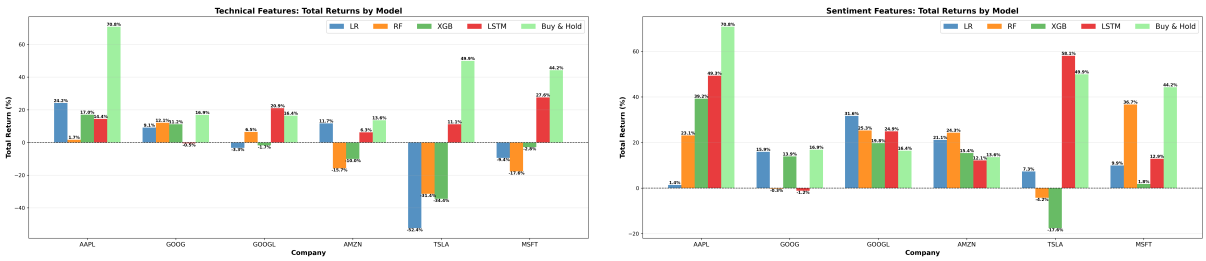
Bold values indicate the best-performing sentiment model for each stock. Sentiment models achieved superior Sharpe Ratios on four of six stocks (GOOGL, AAPL, AMZN, MSFT), demonstrating better risk-adjusted returns.

Classification Metrics

Table 3: Classification Metrics Across All Stocks (2019 Test Period)

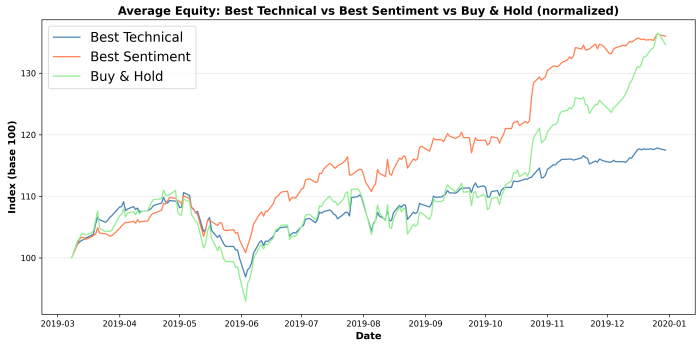
	AAPL		GOOG		GOOGL		AMZN		TSLA		MSFT	
Model	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
LR-Tech	0.570	0.690	0.510	0.420	0.498	0.315	0.518	0.601	0.520	0.580	0.545	0.630
LR-Sent	0.494	0.513	0.530	0.510	0.514	0.508	0.550	0.589	0.510	0.565	0.560	0.650
RF-Tech	0.446	0.498	0.495	0.485	0.518	0.494	0.518	0.639	0.535	0.610	0.550	0.645
RF-Sent	0.578	0.619	0.545	0.565	0.542	0.594	0.550	0.641	0.525	0.590	0.570	0.670
XGB-Tech	0.470	0.483	0.505	0.465	0.522	0.500	0.482	0.549	0.540	0.620	0.555	0.655
XGB-Sent	0.546	0.578	0.535	0.545	0.554	0.582	0.522	0.583	0.530	0.600	0.565	0.665

Returns Comparison



(a) Technical models returns (b) Sentiment models returns

Figure 5: Overall returns comparison across all stocks and models



(a) Average equity curve comparison (b) Overall drawdown comparison

Figure 6: Overall performance metrics across all stocks

C Code Repository

GitHub Repository: `https://github.com/liviomanzi11/emotion-driven-markets-Manzinali-Livio`

Repository Structure

The project follows a modular Python architecture:

```
emotional-driven-markets/
 main.py                # Main entry point
 environment.yml        # Conda dependencies
 data/
    raw/               # Kaggle datasets (CSV)
    processed/         # Generated features & models
 src/
    pipelines/         # Data processing modules
    models/            # ML model training
    strategies/        # Backtesting framework
    visualization/     # Plotting utilities
 results/               # Equity curves & figures
 tests/                 # Unit tests (57 tests)
```

Installation & Reproduction

1. Clone & Setup:

```
git clone https://github.com/liviomanzi11/
  emotion-driven-markets-Manzinali-Livio.git
cd emotion-driven-markets-Manzinali-Livio
conda env create -f environment.yml
conda activate emotion-driven-markets
```

2. Download data from Google Drive:

- Raw: `Tweet.csv`, `Company.csv`, `Company_Tweet.csv` → `data/raw/`
- Preprocessed: `tweet_sentiment.csv`, `company_stock_data.csv` → `data/processed/`

3. Run pipeline:

`python main.py` (30-35 min runtime)

Executes: sentiment aggregation, stock data retrieval, feature engineering, model training (48 models), backtesting, and visualization.