Financial News Sentiment and Stock Movement Prediction Using Transformer Models and Machine Learning

Devoux Deysel

2025

Contents

1 Title Page

2 Abstract

This report presents an integrated financial prediction system that combines natural language processing, sentiment analysis, market data engineering, and machine learning models to predict short-term stock price movements. Using the FNSPID financial news dataset and daily OHLCV stock data for AAPL, MSFT, AMZN, and GOOG, the study evaluates whether transformer-based sentiment scoring improves directional forecasting. FinBERT, a domain-specific BERT model optimized for financial language, is applied to news summaries and full articles to derive sentiment labels and confidence scores. These signals are aggregated and merged with technical indicators to build a hybrid predictive feature set. The system includes an interactive Streamlit dashboard and an LLM assistant for exploratory interpretation. Results demonstrate that sentiment-enhanced modeling is feasible and provides meaningful predictive insights, although limitations remain regarding sparsity, noise, and feature interactions.

1. Introduction

Financial markets respond rapidly to new information, much of which originates from unstructured news text. The ability to quantify the sentiment contained within financial news and incorporate it into predictive models has become increasingly relevant. Traditional statistical methods often fail to capture the contextual nuance of financial language.

Transformer-based language models, such as BERT, allow for high-quality semantic extraction from financial documents. FinBERT provides financial-domain sentiment classification and is well suited for modeling investor reaction. This project integrates FinBERT sentiment with technical indicators to predict short-term stock movements using machine learning.

2. Background and Related Work

Sentiment analysis in finance has shown predictive power across multiple studies. Tetlock (2007) demonstrated that negative media tone predicts downward market movements, while optimistic tone correlates with market rallies. The development of transformer-based models revolutionized NLP by allowing contextualized text representations.

FinBERT (Araci, 2019) adapts BERT specifically to financial text, improving sentiment classification accuracy. Machine learning models such as logistic regression, random forests, and gradient boosting (Friedman, 2001) serve as robust baselines for directional financial prediction.

3. Data Sources

3.1 3.1 Financial News Dataset (FNSPID)

The dataset contains more than 13,000 articles, including: - Titles - Publishers and authors - Multiple summaries (Textrank, LexRank, Luhn) - Raw text - Stock symbol labels - Timestamps

Only articles referencing AAPL, MSFT, AMZN, and GOOG were retained.

3.2 3.2 Stock Price Data

Daily OHLCV data for the four tickers was sourced and aligned with the news dataset.

3.3 3.3 Alignment Window

All timestamps were converted to UTC and aggregated to daily granularity.
The joint modeling window spans **1 September 2023 – 15 December 2023**.

6 4. Data Preprocessing

Preprocessing steps included: - Inspecting structure and missing fields
- Removing invalid or irrelevant entries
- Filtering news to supported tickers
- Standardizing time zones and date formats
- Cleaning text fields

When present, *Textrank_summary* was preferred as FinBERT input to reduce token lengths.

Ticker symbols were standardized, ensuring compatibility between datasets.

7 5. Sentiment Analysis Using FinBERT

FinBERT (ProsusAI, HuggingFace) was applied using a batch inference pipeline.

7.0.1 Steps:

1. Batch size: 32 records

2. Truncation to 512 tokens

3. Extraction of:
    - Sentiment label (Positive, Negative, Neutral)
    - Confidence score

The resulting dataset (*news_with_sentiment.csv*) included sentiment fields for each article.

8 6. Feature Engineering

8.1 6.1 Sentiment-Based Features

Daily aggregated metrics included: - Mean sentiment
- Positive and negative article counts
- Weighted sentiment (sentiment $\times$ confidence)

8.2 6.2 Technical Indicators

Extracted from OHLCV data: - Daily returns
- Rolling volatility
- SMA and EMA
- RSI and MACD

8.3 6.3 Hybrid Meta-Features

Sentiment and technical indicators were combined to produce a unified modeling dataset.

9 7. Predictive Modeling

The target variable was binary: - **1 = UP** (next-day close > current close)
- **0 = DOWN**

Models used: - Logistic Regression
- Random Forest
- XGBoost
- A simple ensemble

Performance metrics included directional accuracy and predicted probability of upward movement.

10 8. Explainability Using SHAP

SHAP (SHapley Additive exPlanations) provides a rigorous method for decomposing model predictions into additive feature contributions (Lundberg & Lee, 2017).

10.0.1 Findings:

- Rolling volatility consistently exhibited strong predictive influence.
- Mean sentiment and positive article counts contributed to upward predictions, particularly on news-heavy days.
- Technical indicators interacted with sentiment features, showing nonlinear dynamics.
- SHAP reinforced that sentiment-enhanced predictors were useful but not dominant—consistent with financial theory.

SHAP interpretability validated that sentiment signals are meaningful components of the hybrid model.

11 9. System Implementation

11.1 9.1 Insights Dashboard

Displays: - Predicted direction
- Probability of upward movement
- Price context
- Rolling volatility
- Sentiment indicators

11.2 9.2 Market Reaction Simulator

Allows exploration of hypothetical: - Sentiment shocks
- Volatility increases
- Price perturbations

11.3 9.3 News Browser

Shows all articles for a given day, along with FinBERT sentiment.

11.4 9.4 LLM Assistant

Provides: - Model explanations
- Daily insights
- User-driven interpretation using the merged dataset

12 10. Results and Discussion

Key findings: - Hybrid features (sentiment + technical indicators) outperformed technical-only models.
- Sentiment influenced predictions most during periods of high news density.
- Volatility remained one of the dominant predictors.
- The dashboard improved interpretability and exploratory analysis.

Overall, the sentiment-driven pipeline produced meaningful improvements and actionable insights.

13 11. Limitations

- Sparse news days weaken sentiment aggregation reliability.

- Automated summaries may omit relevant context.
- Short study window limits generalizability.
- Predictive relationships in financial markets are non-stationary.

14 12. Conclusion

This project demonstrates that transformer-based sentiment analysis, when combined with technical market features, enhances predictive modeling of short-term stock movements. FinBERT provided accurate domain-specific sentiment classification, and SHAP analysis verified meaningful feature contributions. The integrated pipeline—spanning preprocessing, sentiment extraction, modeling, and interactive visualization—offers a practical and interpretable framework for financial analysis.

Future work includes expanding datasets, integrating macroeconomic sentiment, using more advanced NLP architectures, and developing automated retraining pipelines for production use.

15 References

Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with BERT.*
Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine.* Annals of Statistics.
Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions.* NeurIPS.
Tetlock, P. C. (2007). *Giving content to investor sentiment: The role of media in the stock market.* Journal of Finance.
ProsusAI. *FinBERT documentation.* HuggingFace.