QuantEdge: Stock Prediction via Market and Sentiment Modeling from Social Media & News

Shiva Sai Vummaji, Tuan Minh Pham, Anish Nalluri, Chengyu Yang

CS 577 Final Project Proposal
Purdue University

October 27, 2025

1. Introduction

Stock prices are shaped by both quantitative fundamentals and qualitative sentiment. While traditional forecasting relies on historical prices and technical indicators, investor discussions on platforms such as Twitter (X), Reddit, and StockTwits often provide leading signals about market trends, volatility, and collective psychology. This project investigates how combining structured market features with sentiment information from social media can improve short-term stock movement prediction.

Research Question: Does retrieval-enhanced sentiment modeling combined with market data improve short-term stock prediction accuracy compared to traditional approaches, and which fusion strategy is most effective?

To address this, we propose a comparative framework testing three key hypotheses:

- Multi-modal models (market + sentiment) outperform single-modality baselines
- Retrieval-based filtering improves sentiment feature quality over raw text processing
- Fusion strategy choice (early vs. late) significantly impacts prediction accuracy

Our approach systematically evaluates traditional statistical models (ARIMA), classical machine learning (Random Forest, XGBoost), and deep sequence architectures (LSTM, GRU, Transformer) across multiple data configurations. We engineer market features from S&P 500 data and apply retrieval-enhanced sentiment analysis using Sentence-BERT + FAISS with FinBERT/FinGPT.

By rigorously comparing market-only, sentiment-only, and fused models, we aim to quantify the contribution of each modality and identify optimal integration strategies for short-term stock forecasting. This version:

2. Related Work

Prior research has explored the relationship between online sentiment and stock price movements. Bing et al. (2015) identified significant correlations between Twitter sentiment and market returns, while Ko and Chang (2021) applied LSTM-based architectures to financial news for price forecasting. More recent studies such as HiSA-SMFM (Bhattacharjee et al., 2022) and StockEmotions (Guo et al., 2023) combine historical price data with textual sentiment to improve predictive performance, demonstrating the effectiveness of multi-modal approaches.

Our work builds on this line of research by integrating both quantitative market modeling and sentiment-driven analysis into a unified prediction framework. We systematically compare classical machine learning models (e.g., XGBoost, Random Forest) with deep sequence models (LSTM, GRU) to evaluate their effectiveness in forecasting short-term stock movements. In parallel, we enhance sentiment modeling through a retrieval-based filtering step that identifies the most contextually relevant social media posts and news headlines before applying FinBERT or FinGPT for sentiment scoring. This retrieval mechanism refines sentiment features by reducing noise and redundancy, enabling a more reliable fusion of textual and numerical signals in the final predictive model.

3. Proposed Approach

Our research framework systematically compares multiple modeling approaches to evaluate the impact of multi-modal data integration on stock prediction accuracy. The experimental design tests three core hypotheses through five interconnected modules:

1. **Market Data and Feature Engineering:** We collect historical OHLCV data for S&P 500 tickers using `yfinance` and AlphaVantage APIs, handling stock splits and missing values. Technical indicators (SMA, EMA, RSI, MACD, Bollinger Bands), lagged returns, rolling volatility, and fundamental metrics (P/E ratio, EPS growth) form the quantitative feature set for baseline comparisons.
2. **Prediction Model Variants:** We implement multiple model classes to ensure robust comparisons: traditional statistical (ARIMA), classical ML (Linear Regression, Random Forest, XGBoost), and deep sequence models (LSTM, GRU, Transformer). Each model is evaluated on market-only data to establish performance baselines.
3. **Sentiment Data with Retrieval Enhancement:** We gather text data from financial news (Yahoo Finance, NewsAPI) and social media (Twitter, Reddit). To test retrieval impact, we compare two conditions: (1) raw text processing and (2) retrieval-enhanced features using Sentence-BERT + FAISS to filter relevant content. Both conditions use FinBERT/FinGPT for sentiment scoring, producing daily aggregates (mean sentiment, variance, frequency).
4. **Multi-Modal Fusion Strategies:** We experimentally compare two fusion approaches: early fusion (feature concatenation) and late fusion (ensemble averaging). This tests whether integration strategy affects how market and sentiment signals combine for improved prediction.
5. **Optional Deployment:** If time permits, we deploy the best-performing model via `FastAPI` to demonstrate real-time inference capability.

4. Experimental Plan

To evaluate our hypotheses, we design controlled experiments comparing different data modalities and fusion mechanisms under identical market conditions. All models are trained and tested on chronologically split datasets to avoid look-ahead bias.

The experiments aim to measure both predictive accuracy and financial interpretability.

4.1 Data Sources

Our experimental framework integrates three complementary data streams to enable rigorous comparison across modeling approaches: quantitative market data, fundamental indicators, and unstructured sentiment data from both financial news and social media.

Market and Fundamental Data: We collect up to ten years of historical OHLCV data for S&P 500 constituents using `yfinance` and AlphaVantage APIs. Our preprocessing pipeline automatically handles stock splits, missing values, and delistings to ensure data consistency. Fundamental indicators including P/E ratios, EPS growth, and revenue metrics are sourced from AlphaVantage and SEC EDGAR API to incorporate company-specific financial context. For real-time validation, we implement live data streaming via AlphaVantage or Polygon API.

Financial News and Social Media Sentiment Data: To capture market sentiment, we gather textual data from two primary sources: financial news headlines from Yahoo Finance and NewsAPI, and social media discussions from Twitter (X) and Reddit communities (*r/stocks*, *r/wallstreetbets*) using `snscrape` and PRAW.

Integrated Dataset: All data modalities are synchronized along a unified temporal axis aligned with trading dates. This integrated dataset supports controlled experiments across different data configurations: market-only baselines, sentiment-only models, and multi-modal fusion approaches, enabling systematic evaluation of each component's contribution to prediction accuracy.

To evaluate our hypotheses, we design controlled experiments comparing different data modalities and fusion mechanisms under identical market conditions. All models are trained and tested on chronologically split datasets to avoid look-ahead bias.

The experiments aim to measure both predictive accuracy and financial interpretability.

4.2 Baseline and Comparative Design

Phase 1: Baseline Selection We evaluate multiple candidate models on market+fundamental data to identify the strongest baseline:

- **Traditional:** ARIMA, Linear Regression
- **Classical ML:** Random Forest, XGBoost, SVM
- **Deep Learning:** LSTM, GRU

The best-performing model becomes our official baseline for hypothesis testing.

Phase 2: Comparative Setups Using the selected baseline, we test:

- **Market-only baseline** — technical indicators (SMA, EMA, RSI, MACD) + fundamentals (P/E, EPS) + lagged returns
- **Sentiment-only model** — FinBERT/FinGPT sentiment features (mean, variance, frequency)
- **Multi-modal Fusion** — combined market + retrieval-enhanced sentiment features

4.3 Evaluation Setup

Data Split: Training (2016–2021), Validation (2022), Testing (2023–2024).

Metrics: RMSE, MAE for regression; Accuracy, F1-score for classification; Sharpe Ratio and Hit Rate for financial interpretation.

Statistical Tests: Paired t-tests and Diebold–Mariano tests are applied to evaluate whether fusion models outperform the baseline with statistical significance.

4.4 Ablation Studies

Retrieval Effectiveness: Compare FinBERT sentiment features with and without Sentence-BERT + FAISS retrieval filtering.

Fusion Strategy: Compare early vs. late fusion for model stability and predictive variance.

4.5 Expected Results and Validation

We expect fusion models, especially retrieval-enhanced late fusion, to outperform the baseline in both predictive metrics and stability. Performance improvements will be validated using significance tests, ensuring that observed gains are not due to randomness or overfitting.

4.6 Timeline

Weeks	Milestones
Week 1 (Oct 25 – Nov 1)	Collect and preprocess S&P 500 OHLCV and fundamental data (yfinance, AlphaVantage). Gather initial sentiment samples (Twitter, Reddit, NewsAPI). Clean and align datasets.
Week 2 (Nov 2 – Nov 8)	Engineer technical indicators (SMA, EMA, RSI, MACD, Bollinger Bands). **Phase 1:** Train and evaluate some baseline candidates (ARIMA, Linear Regression, Random Forest, XGBoost, LSTM, GRU) on market-only data. Identify the strongest baseline model.
Week 3 (Nov 9 – Nov 15)	Collect and preprocess full sentiment data. Embed text with Sentence-BERT, build FAISS index, apply FinBERT/FinGPT for sentiment scoring. Aggregate daily sentiment features (mean, variance, frequency).
Week 4 (Nov 16 – Nov 22)	Using the selected baseline, train and compare: (1) Market-only baseline, (2) Sentiment-only models, (3) Multi-modal fusion (early + late). Tune hyperparameters for fusion models.
Week 5 (Nov 23 – Nov 27)	Evaluate all models on chronological test split. Compute RMSE, MAE, F1, and Sharpe ratio. Perform statistical significance tests (t-tests, Diebold-Mariano). Conduct ablation studies (retrieval impact, fusion strategies).
Week 6 (Nov 28 – Dec 1)	Write final report and prepare presentation slides. Analyze why certain approaches worked better. (Optional) Deploy FastAPI demo for real-time prediction.

5. Expected Outcomes

We expect our findings to demonstrate how retrieval-enhanced sentiment features and multimodal fusion strategies can facilitate short-term stock prediction.

Specifically, the expected results are as follows:

(1) Models using Sentence-BERT + FAISS filtering are expected to outperform models using unfiltered FinBERT/FinGPT sentiment data, demonstrating that retrieval-based selection can improve the quality of text features by removing irrelevant or noisy inputs.

(2) The fused model is expected to significantly outperform the selected market baseline in terms of RMSE, F1 score, and Sharpe ratio, confirming the predictive value of sentiment data.

(3) This project will provide a comparative benchmark for retrieval-enhanced multimodal modeling in financial forecasting and provide empirical evidence on when and how sentiment data can add value to quantitative models

6. Conclusion

This project explores quantitative and textual drivers of stock behavior through a unified prediction framework. By combining engineered technical indicators with refined sentiment features, we evaluate how structured market data and filtered social media signals jointly contribute to short-term forecasting accuracy. The integration of retrieval-enhanced sentiment with traditional market modeling provides insights into multi-modal financial prediction. While our framework integrates diverse data sources and models, it is limited by data availability and noise in social media sentiment, as well as the short project timeline that constrains hyperparameter tuning and large-scale model training. Future extensions could explore larger data sources, and causal interpretation of sentiment signals to strengthen predictions.

References

Y. Yang, Y. Huang, and E. Lin. "FinBERT: A Pretrained Language Model for Financial Communications." *arXiv preprint arXiv:2006.08097*, 2020.

R. Bhattacharjee et al. "HiSA-SMFM: Historical and Sentiment Analysis based Stock Market Forecasting Model." *arXiv:2203.08143*, 2022.

H. Guo et al. "StockEmotions: Discover Investor Emotions for Financial Sentiment Analysis and Multivariate Time Series." *arXiv:2301.09279*, 2023.

X. Bing et al. "Predicting Stock Market Trends via Twitter Sentiment Analysis." *IEEE Int. Conf. on Data Mining*, 2015.

J. Ko and C. Chang. "LSTM-based Sentiment Analysis for Stock Price Forecast." *Applied Soft Computing*, 2021.

S. Mehtab, J. Sen, and A. Dutta. "Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models." *arXiv preprint arXiv:2009.10819*, 2020.

Y. Yang, Z. Chen, H. Wu, H. Yin, and L. Chen. "FinGPT: Open-Source LLMs for Financial Applications." *arXiv preprint arXiv:2407.16150*, 2024.