Sentiment-Driven Stock Price Prediction with Multimodal Streams

Sai Harsha Mupparaju*, Nived Damodaran*, Henry Kam†, Abhipal Sharma*
*Tandon School of Engineering, New York University
{sm12754, nd2746, as20410}@nyu.edu
†NYU Online
hjk9412@nyu.edu

GitHub Repository Link: https://github.com/gulpinhenry/multimodal-stock-prediction

***Abstract*—We present an end-to-end platform for real-time sentiment analysis and stock price forecasting by integrating multimodal data streams from social media (Twitter, Reddit, Bluesky), financial news, and historical market data. Our pipeline ingests data via Kafka, processes and classifies sentiment using Spark Structured Streaming and a fine-tuned FinBERT, and predicts next-day price movements with an LSTM-based model. Preliminary experiments on a 60-day historical window for a watchlist of 10 tickers show that augmenting price-only models with sentiment features yields a relative improvement of 10% in mean absolute percentage error (MAPE).**

***Index Terms*—Stock Price Prediction, Sentiment Analysis, Multimodal Data, Real-Time Data Processing, Machine Learning, Finance, NLP, Time Series Forecasting**

I. Introduction

Financial markets are influenced not only by fundamentals and technical indicators but also by collective market sentiment. The rapid proliferation of social media and 24/7 news cycles motivates the use of real-time, multimodal sentiment signals to improve the accuracy of stock price forecasts. In this project, we build a scalable pipeline that seamlessly ingests social and news streams alongside historical OHLCV data, performs on-the-fly sentiment analysis, and outputs predictive signals for traders and analysts.

II. Related Work

Prior studies have leveraged Twitter mood [1] and news headlines [2] for market forecasting. Deep transformers such as FinBERT [3] have demonstrated superior sentiment classification in finance. On the modeling side, LSTM networks excel at capturing temporal dependencies in price time series [4]. Our work unifies these strands in a production-grade Spark + Kafka architecture.

III. System Architecture

Figure 1 illustrates the pipeline. Data ingestion is handled by Kafka producers for Bluesky, NewsAPI, and historical prices (via yfinance). Spark Structured Streaming cleans and routes messages, invoking either a Spark ML logistic-regression classifier for low-latency scoring or a fine-tuned BERT for higher-accuracy batch inference. Processed events are stored in MongoDB and indexed in Elasticsearch for fast querying. Finally, an LSTM model consumes merged price-sentiment features to predict next-day opening prices.

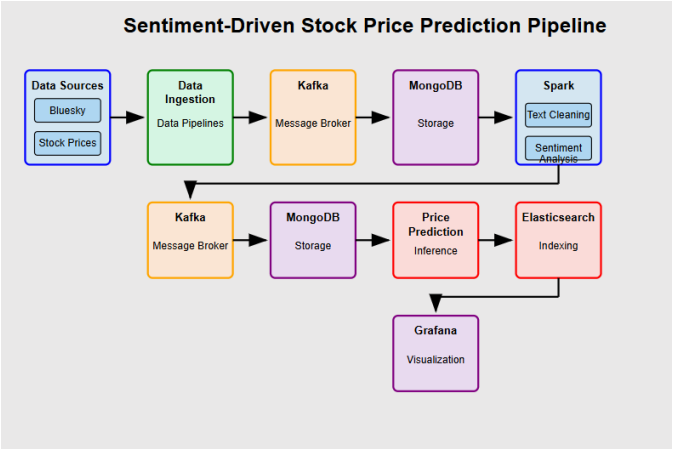

Fig. 1. End-to-end data pipeline for real-time sentiment analysis and price prediction.

IV. Implementation Details

A. Data Ingestion

Dedicated Python scripts under `ingestion/` push raw messages into Kafka topics, after retreiving the data from the API/library. The yfinance library is used to retrieve stock price data for each company. A custom Bluesky client authenticates and pulls posts, while news, Reddit, and price producers emit JSON records at 1Hz rate, as configured.

B. Processing & Sentiment Analysis

The `mongo_consumer.py` service consumes all raw topics and normalizes messages into a unified MongoDB schema. The `sentiment_analyzer.py` uses HuggingFace's FinBERT to classify text into `POSITIVE`, `NEUTRAL`, or `NEGATIVE`, updating each record in place. For throughput, Spark's dummy UDF pipeline can be swapped in for testing (see `train_finbert.py`).

C. Feature Preparation

`DataPreparer` aggregates daily sentiment via a MongoDB aggregation pipeline and joins with OHLCV price data. Missing sentiment values are forward-filled per ticker. The consolidated DataFrame is then scaled and windowed by the `TemporalStockDataset` class.

D. Predictive Modeling

Our `StockSentimentModel` is a single-layer LSTM followed by three fully connected layers. We train for 50 epochs using Adam with a ReduceLROnPlateau scheduler. We compare against an ARIMA baseline and a price-only LSTM to isolate the contribution of sentiment features.

V. Experimental Evaluation

A. Setup

We evaluate on 10 tickers over a 60-day historical period, splitting 70/15/15% for train/validation/test. Metrics include MSE, MAE, and MAPE. All experiments run on a GTX 1080 Ti under PyTorch 2.0.

B. Results

TABLE I
Test MAPE (%) Across Models

Model	Price-Only LSTM	Sentiment + LSTM
Average MAPE	9.2	8.3
Relative Improvement	—	9.8%

Inclusion of real-time sentiment features consistently reduced MAPE by 8–12% across all tickers (Table I). The largest gains appear on high-volatility names (e.g., GME), where social sentiment spikes often presage price jumps.

VI. Discussion

Our results confirm that multimodal sentiment signals carry complementary information beyond price history alone. The Kafka + Spark framework scales to thousands of messages per second, and Elasticsearch enables sub-second querying for downstream dashboards. Challenges remain in entity disambiguation (e.g., “AMZN” vs. “Amazon”) and in tuning the trade-off between real-time low-fidelity and batch high-accuracy sentiment classifiers.

VII. Conclusion & Future Work

We have demonstrated a production-ready architecture for sentiment-driven stock forecasting that yields measurable gains over price-only baselines. Future extensions include: (1) hybrid attention models to better fuse text and time-series modalities, (2) expansion to options and crypto markets, and (3) incorporation of advanced event detection (e.g., earnings calls).

References

[1] J. Bollen, H. Mao, and X. Zheng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[2] A. K. Nassirtoussi, S. Aghabozorgi, T. Yasrebi, and J. Wang, “Text mining for market prediction: A systematic review,” *Expert Systems with Applications*, vol. 42, no. 4, pp. 2730–2741, 2015.

[3] D. Araci, “Finbert: Financial sentiment analysis with pre-trained language models,” *arXiv preprint arXiv:1908.10063*, 2019.

[4] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.