



Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 270 (2025) 582–591

Procedia
Computer Science
www.elsevier.com/locate/procedia

29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025)

Multimodal Financial Sentiment for Stock Return Prediction

Petr Hajek^{a,*}, Josef Novotny^a, Michal Munk^{a,b}, Dasa Munkova^b

^aFaculty of Economics and Administration, University of Pardubice, Studentska 95, Pardubice, 53210, Czech Republic

^bDepartment of Computer Science, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74, Nitra, Slovakia

Abstract

This paper proposes a novel multimodal deep learning framework for stock return prediction that integrates heterogeneous data sources: technical indicators, market investor sentiment indices, and textual sentiment extracted from earnings conference call transcripts. The proposed model employs a hybrid architecture combining transformer encoder for the technical modality and neural networks for market and textual modalities. A modality-level attention mechanism is used in a late fusion setup to dynamically weight the contributions of each modality. We evaluate our model on a large-scale dataset comprising 24,821 samples from 497 S&P 500 companies over the period 2010–2022. The results show that our model outperforms traditional models (LSTM, BiLSTM, CNN-LSTM) and alternative fusion strategies, achieving a directional accuracy of 59.94% on the test set. Attention weight analysis confirms that all three modalities contribute meaningfully to prediction performance. These results demonstrate the overall effectiveness of the proposed framework in accurately predicting abnormal stock returns in a multimodal setting.

© 2025 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0>)

Peer-review under responsibility of the scientific committee of the KES International.

Keywords: Multimodal fusion; Financial sentiment; Stock market; Prediction; Earnings conference calls.

1. Introduction

The stock market is a complex, dynamic system influenced by a wide range of factors, including macroeconomic indicators, corporate performance, political events and, most importantly, investor sentiment [1]. Investor sentiment, broadly defined as the collective emotions and attitudes of market participants towards financial assets, plays a central role in shaping market trends and price movements [2]. Behavioral finance suggests that market participants do not always act rationally; instead, psychological biases and emotions can lead to asset mispricing, bubbles and crashes [1]. Understanding and quantifying investor sentiment is therefore critical to improving stock market forecasting models and investment strategies.

Recent advances in artificial intelligence (AI) and natural language processing (NLP) have enabled sentiment analysis to extract and interpret investor sentiment from diverse data sources such as news articles, social media platforms,

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.

E-mail address: author@institute.xxx

analyst reports, and financial statements [3]. Traditional financial theories, such as Fama's efficient market hypothesis, suggest that all available information is already reflected in stock prices [4]. However, empirical evidence increasingly challenges this view, showing that sentiment-driven fluctuations can create short-term inefficiencies that can be exploited [5]. In addition, sentiment analysis allows researchers and traders to quantify investor sentiment in real time, providing insights into market dynamics beyond fundamental and technical analysis. Studies have shown that positive sentiment correlates with bullish market behavior, while negative sentiment often precedes market downturns. This is consistent with theories such as prospect theory [6], which suggests that investors react asymmetrically to gains and losses, and further reinforces the importance of sentiment as a predictive variable in stock market modeling.

Financial sentiment analysis has evolved considerably over the past decade, with increasing reliance on machine learning and deep learning models to interpret and predict market trends [3]. Early sentiment analysis relied primarily on lexicon-based methods that mapped words to predefined sentiment scores. However, the advent of deep learning models such as transformers and recurrent neural networks has enabled more fine-grained interpretation of sentiment from textual data, allowing for greater accuracy in stock market forecasting [7]. Financial sentiment analysis has been successfully applied to the prediction of short-term stock price movements [8], volatility forecasting [9], and event-driven market reactions [10], demonstrating its practical value in investment decision making.

In recent years, multimodal financial sentiment analysis has gained prominence as researchers have recognized the limitations of text-based sentiment analysis alone [11, 12]. Financial sentiment is expressed through multiple channels, including textual reports, social media discussions, images, videos, and even the tone of voice on earnings calls [3]. By integrating multimodal data sources, such as combining textual sentiment with cues from financial charts or audio sentiment from CEO speeches, researchers can gain a more comprehensive view of market sentiment. The fusion of different modalities improves forecasting accuracy and provides deeper insights into investor behavior, making multimodal sentiment analysis a critical advancement in stock market forecasting.

Despite advances in sentiment-based stock market forecasting, existing models suffer from several limitations. Many rely solely on textual sentiment extracted from financial news or social media [13, 14, 15], ignoring other important sentiment signals such as market-based investor sentiment indices and technical indicators. Moreover, traditional sentiment models often fail to account for interdependencies between different data sources [13, 16], leading to incomplete market representations and suboptimal predictions. To address these challenges, we propose a novel multimodal deep learning model that integrates FinBERT-based text sentiment modality with market-based investor sentiment modality, as well as technical analysis modality. By leveraging this comprehensive fusion of data sources, our model aims to provide a more robust and accurate stock market prediction framework that effectively captures stock market dynamics. More precisely, we propose a novel multimodal deep learning model that integrates multiple sources of financial sentiment for stock return prediction. Our key contributions are as follows:

- We develop a framework that combines FinBERT-based textual sentiment analysis with market investor sentiment indices (VIX, Twitter Happiness Index, Daily News Sentiment Index) and technical analysis indicators. This approach enables a more comprehensive representation of investor sentiment and its impact on stock returns.
- We introduce a novel cross-company attention mechanism that enhances predictive accuracy by capturing interactions between sentiment factors across different stocks.
- Our model leverages a multi-head attention mechanism to integrate multimodal data, ensuring that both short-term fluctuations and long-term trends in investor sentiment are accounted for. This technique addresses limitations in prior research related to data alignment and feature fusion.
- We validate our approach using a large-scale dataset covering 497 companies listed in the S&P 500 from 2010 to 2022. Empirical results demonstrate that our model achieves competitive predictive performance.

The rest of this paper is structured as follows. Section 2 reviews related literature on multimodal stock market prediction. Section 3 presents the research methodology, detailing the model architecture and multimodal fusion strategy. Section 4 describes the data sources, feature engineering. Section 5 presents experimental setup, reports the experimental results, and evaluates the performance of the proposed model against baseline models. Finally, Section 6 concludes the paper with key findings, limitations, and directions for future research.

2. Related Literature

Investor sentiment analysis has evolved through a variety of approaches, ranging from traditional surveys and expert opinion to automated computational techniques [3]. Early approaches relied primarily on investor sentiment indices derived from surveys, such as the American Association of Individual Investors (AAII) sentiment index [17]. With the rise of online discussions and financial news, sentiment analysis has increasingly incorporated NLP and machine learning techniques to extract sentiment from textual data sources. Sentiment lexicons, supervised learning models, and transformer-based deep learning models such as BERT and FinBERT have been widely used to assess investor sentiment with improved accuracy [18, 19, 20]. Furthermore, financial-specific sentiment models, such as FinBERT or FinLlama, have been developed to better capture the nuances of financial language [21]. Techniques such as sentiment-aware word embeddings and contextualized transformers have further improved financial sentiment classification [3].

Previous research has shown that social media sentiment, financial news sentiment, and investor discussion sentiment are predictive of short-term stock price movements [23] and market volatility [9]. However, existing models often face challenges in handling real-time sentiment fluctuations and interpreting the causal relationships between sentiment and market outcomes [24]. The integration of financial sentiment with macroeconomic indicators, fundamental financial analysis, and alternative data sources has been proposed to improve prediction capabilities [15, 25].

Multimodal approaches have emerged as a promising direction for improving stock market prediction by integrating multiple data sources, including text, numerical indicators, images and audio (see Table 1). Deep learning architectures such as attention-based transformers, convolutional neural networks (CNNs), and long short-term memory (LSTM) networks have been used to effectively process multimodal inputs. Overall, the use of multimodal learning enables more robust predictions by capturing interactions between different financial signals, thereby reducing the limitations of single-modality models.

Recent studies in multimodal stock market prediction have increasingly focused on fusing different data sources—such as news articles, social media sentiment, earnings call transcripts, and market technical indicators—to improve forecasting accuracy. For example, Wu et al. [16] used a CNN-LSTM model to integrate historical stock prices with leading market indicators such as options and futures, capturing complex dependencies in financial time series. Similarly, Chen et al. [22] proposed a CNN-BiLSTM-ECA model, leveraging Efficient Channel Attention (ECA) to enhance feature selection from financial time series data. Other studies, such as [27], introduced a two-channel attention fusion mechanism with CNN-LSTM, which extracts and integrates news sentiment with stock prices. These works demonstrate that incorporating multiple modalities can improve predictive performance, but each approach faces challenges in effective data fusion and model interpretability. For example, while CNN-based feature extraction enhances predictive accuracy [27, 16], it may struggle with long-range dependencies.

Recent studies such as [14, 19] have emphasized the alignment of earnings call data and the inclusion of additional indicators such as COVID-related metrics to better capture market dynamics, while Sawhney et al. [23] proposed a deep attentive learning model that fuses social media text with company correlation information. Li and Xu [20] pushed the research further by integrating generative adversarial networks (GANs) with transformer-based attention mechanisms to improve stock price prediction accuracy by selectively focusing on key market features. However, despite these advances, many models still face limitations such as high model complexity, and the difficulty of aligning asynchronous modalities [28]. Moreover, reliance on specific sentiment dictionaries or narrow data sources can hinder the generalizability of these approaches [29].

3. Research Methodology

The proposed multimodal deep learning model aims to improve stock return prediction by integrating multiple sources of investor sentiment and financial data. The model consists of four key components: (1) a transformer encoder block for processing technical indicators, (2) dense layers for market and text sentiment processing, (3) a multi-head attention mechanism for efficient multimodal fusion, and (4) cross-company integration using ticker embeddings. The framework of the proposed model is illustrated in Fig. 1.

The selection of these components is based on the need to capture both short-term fluctuations and long-term dependencies in financial markets. Sentiment-driven models have shown promise in capturing investor reactions, while

Table 1. Multimodal approaches in stock market prediction.

Ref.	Modalities	Model
[9]	Verbal and vocal cues	Multimodal deep regression (BiLSTM)
[30]	Stock prices, news sentiments	Multi-kernel learning + DNN
[14]	Earnings call text and audio	HTML
[23]	Social media text, company correlations	Deep attentive learning
[26]	Stock prices, news sentiment	Dep reinforcement learning
[27]	News sentiment, stock prices	Two-channel attention with CNN-LSTM
[16]	Historical stock prices, leading indicators (options, futures)	CNN-LSTM (SACLSTM)
[15]	Fundamental indicators, news sentiment	LSTM + transformer
[8]	News articles, stock prices, sentiment indexes	Trellis network + sentiment attention
[13]	Financial news sentiment, historical prices	FinBERT + LSTM
[19]	Market sentiment, stock prices, COVID indicators	LSTM + transformer
[20]	Stock prices, news sentiment, social media news	GANs + transformer-based attention
This study	Technical indicators, market sentiment, text sentiment	Cross-company attention mechanism

Notes: HTML - Hierarchical Transformer-based Multi-task Learning, SACLSTM - Stock Sequence Array Convolutional LSTM.

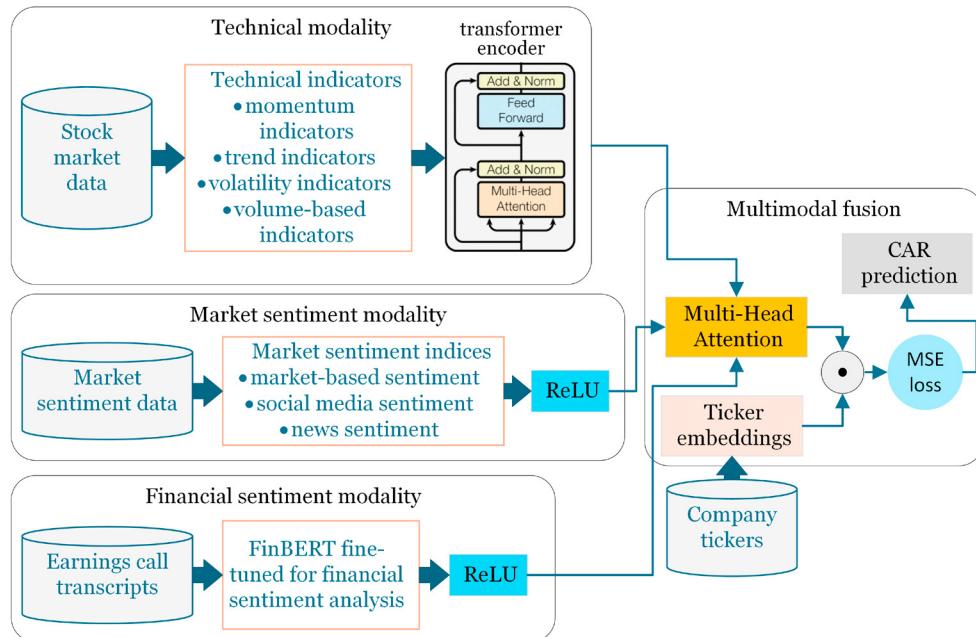


Fig. 1. Framework of the proposed multimodal stock return prediction model.

market data and technical indicators provide essential quantitative financial features. By combining these modalities for cumulative abnormal return (CAR) prediction, our model aims to improve predictive accuracy and provide an effective decision-support tool for investors.

3.1. Transformer Encoder for Technical Features

Technical indicators are widely used to assess stock price trends, momentum, and volatility. Traditional models rely on recurrent networks, but transformers have demonstrated superior ability in capturing long-range dependencies in time series data [31]. In the proposed model, the transformer encoder processes technical indicators as follows:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (1)$$

where Q, K, V represent the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. The self-attention mechanism allows the model to weigh the importance of past technical indicator values adaptively. The output is then normalized:

$$\hat{H} = \text{LayerNorm}(H + \text{Dropout}(\text{Attention}(Q, K, V))), \quad (2)$$

where H is the input to the transformer layer.

3.2. Market and Text Sentiment Processing

Investor sentiment, reflected through market-based indicators and textual sentiment, influences stock prices [1]. Market sentiment is captured via VIX, THI (Twitter Happiness Index), and SSW (Shapiro, Sudhof and Wilson news sentiment index) [32], while textual sentiment from earnings conference calls is extracted using FinBERT, a financial-domain-specific transformer [33]. These features are processed through dense layers:

$$M_{\text{dense}} = \text{ReLU}(W_m M + b_m). \quad (3)$$

Similarly, text sentiment extracted using FinBERT is transformed via:

$$T_{\text{dense}} = \text{ReLU}(W_t T + b_t), \quad (4)$$

where M and T are the input feature vectors for market and text sentiment, respectively. The dense layers enable non-linear transformation, which enhances the extraction of predictive patterns from sentiment features.

3.3. Multimodal Fusion via Attention Mechanism

To effectively combine technical, market, and textual features, we employ a multi-head attention mechanism. Instead of simply concatenating features, this mechanism dynamically assigns importance weights to different modalities:

$$F_{\text{fusion}} = \text{MultiHeadAttention}(Q = T_{\text{tech}}, K = [M_{\text{dense}}, T_{\text{dense}}], V = [M_{\text{dense}}, T_{\text{dense}}]). \quad (5)$$

This fusion technique allows the model to learn which modality is more relevant for stock return prediction at any given time, ensuring that the model does not over-rely on a single data source.

3.4. Cross-Company Integration with Ticker Embeddings

Stock returns are influenced by company-specific factors such as financial health, sector performance, and competitive positioning. To incorporate these effects, we introduce ticker embeddings $E_{\text{ticker}} = \text{Embedding}(\text{ticker_id})$. These embeddings are learned representations that capture company-specific characteristics. The final fused representation is obtained through element-wise multiplication between ticker embeddings and multimodal features:

$$O_{\text{cross}} = F_{\text{fusion}} \odot \sigma(W_e E_{\text{ticker}}), \quad (6)$$

where σ is the sigmoid activation function. This mechanism enhances the model's ability to generalize across different stocks and improve prediction accuracy.

3.5. Final Prediction Layer

The final prediction layer performs regression on the fused multimodal features $\hat{Y} = W_f O_{\text{cross}} + b_f$, where \hat{Y} represents the predicted CAR. The use of dense layers in the final stage allows the model to learn complex non-linear relationships between input features and stock returns.

3.6. Training and Optimization

The model is trained using mean squared error (MSE) loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2, \quad (7)$$

where Y_i is the actual CAR and \hat{Y}_i is the predicted CAR. The Adam optimizer is used to enhance training efficiency, with early stopping and learning rate scheduling implemented to prevent overfitting.

4. Data

The dataset used in this study consists of 24,821 instances from 497 publicly traded companies listed in the S&P 500 index between 2010 and 2022. The data were collected from multiple sources, including financial databases and earnings call repositories, to create a multimodal dataset for stock return prediction. The dataset integrates company stock prices, technical indicators, market investor sentiment, and textual sentiment analysis from earnings call transcripts. Daily stock price data for each of the 497 companies were collected using the `yfinance` Python library, which provides historical data from Yahoo Finance. To calculate the Cumulative Abnormal Return (CAR), we followed a standard event study methodology, using the S&P 500 index as the market benchmark. CAR was computed over a three-day event window surrounding each company's earnings conference call date as $CAR_i = \sum_{t=-1}^{+1} (R_{i,t} - R_{m,t})$, where $R_{i,t}$ is the daily return of stock i on day t , and $R_{m,t}$ is the return of the S&P 500 index on the same day. The earnings conference call dates were obtained from Seeking Alpha.

Three daily market investor sentiment indicators were incorporated into the dataset:

- VIX (Volatility Index)¹: Collected from the Chicago Board Options Exchange (CBOE), the VIX measures market expectations of near-term volatility and serves as a proxy for market-based investor sentiment.
- THI (Twitter Happiness Index)²: The THI measures social media investor sentiment by analyzing the overall mood of financial discussions on Twitter.
- SSW news sentiment index³: This index captures sentiment shifts in financial news articles, providing an additional signal for stock market movements.

Technical indicators were derived from the daily stock price data to capture market trends, momentum, volatility, and trading volume patterns. Each technical indicator was computed using a rolling window approach to ensure real-time applicability in stock return predictions. The purpose of the momentum technical indicators was to consider overbought/oversold signals (RSI), trend reversals (MACD), and speed of price movements (ROC). Similarly, trend indicators were used to incorporate short and long-term trends (EMA 50 and EMA 200) and strength of the trend (ADX), while volatility indicates price extremes and mean reversion (BB), event-driven volatility (ATR), and past volatility impact on CAR (Rolling Volatility). Finally, volume-based indicators measure accumulation/distribution trends (OBV) and institutional buying interest (VWAP). Details on the used technical indicators are given in Table 2.

The textual data component of our multimodal model is derived from earnings call transcripts. We collected transcripts from Seeking Alpha by identifying URLs corresponding to each company's earnings call. The transcripts were then processed to extract textual sentiment features. To measure the sentiment of earnings calls, we employed FinBERT, a domain-specific NLP model trained for financial sentiment analysis⁴. Each transcript was tokenized into sentences and passed through FinBERT to compute sentence-level sentiment scores. The sentiment polarity score was calculated as $S_{\text{sentiment}} = \frac{1}{N} \sum_{j=1}^N (P_{\text{positive},j} - P_{\text{negative},j})$, where $P_{\text{positive},j}$ and $P_{\text{negative},j}$ are the FinBERT-derived probabilities for positive and negative sentiment for sentence j , and N is the number of sentences in the transcript.

¹ https://www.cboe.com/tradable_products/vix/vix_historical_data/

² hedonometer.org

³ <https://www.frbsf.org/research-and-insights/data-and-indicators/daily-news-sentiment-index/>

⁴ <https://huggingface.co/ProsusAI/finbert>

Table 2. Financial Indicators, Their Categories, and Equations

Category	Indicator	Equation
Momentum	RSI (Relative Strength Index)	$RSI = 100 - \frac{100}{1+RS}$
	MACD (Moving Average Convergence Divergence)	$MACD = EMA_{12} - EMA_{26}$
	ROC (Rate of Change)	$ROC = \frac{Price_t - Price_{t-n}}{Price_{t-n}} \times 100\%$
Trend	EMA 50 & EMA 200	$EMA_t = \alpha \cdot Price_t + (1 - \alpha) \cdot EMA_{t-1}, \quad \alpha = \frac{2}{N+1}$
	ADX (Average Directional Index)	$ADX = EMA\left(\frac{ DI^+ - DI^- }{DI^+ + DI^-}\right)$
Volatility	BB (Bollinger Bands)	$Upper = MA + k\sigma, \quad Lower = MA - k\sigma$
	ATR (Average True Range)	$ATR = EMA(ATR)$
	Rolling Volatility (σ)	$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (r_i - \bar{r})^2}$
Volume	OBV (On-Balance Volume)	$OBV_t = OBV_{t-1} \pm Volume_t$
	VWAP (Volume Weighted Average Price)	$VWAP = \frac{\sum_t Price_t \times Volume_t}{\sum_t Volume_t}$

After collecting stock price data, technical indicators, market sentiment indices, and earnings call transcripts, we merged all features into a single dataset. Each instance in the dataset corresponds to an earnings conference call and includes: (1) the date of the earnings call, (2) the ticker symbol of the corresponding company, (3) a set of technical indicators derived from stock price movements, (4) market sentiment indicators (VIX, THI, SSW), and (5) FinBERT sentiment scores from the earnings call transcript.

5. Experiments

To evaluate the performance of the proposed multimodal model for stock return prediction, we utilized a sliding window approach for testing (with the sliding window size of 15), ensuring that each prediction used past market conditions while maintaining temporal integrity. The dataset, consisting of 24,821 instances from 497 S&P 500 companies between 2010 and 2022, was split into training (70%) and testing (30%) subsets. The training phase involved optimizing hyperparameters through grid search, evaluating different settings of layer configurations, units, and dropout rates to enhance model generalization. The model was trained using the Adam optimizer with an initial learning rate of 5×10^{-4} , decaying over epochs based on validation performance.

Multiple configurations were tested, varying the number of transformer encoder layers for technical features, dense units for market and text sentiment branches, and attention heads for modality fusion. The final optimized settings were selected based on minimizing validation loss and maximizing predictive accuracy. The finalized hyperparameters are as follows: Technical features (2 transformer encoder layers, each with 64 units and 8 attention heads), Market sentiment (2 dense layers with 32 and 16 units, respectively, activated using ReLU), Textual sentiment (2 dense layers with 32 and 16 units, using ReLU activation), and Multimodal fusion (multi-head attention mechanism with 8 heads, followed by a 64-unit dense layer). Furthermore, Adam optimizer was used with $\beta_1 = 0.9$, and $\beta_2 = 0.999$ with learning rate decay, and dropout (0.2) regularization was applied in all major branches to prevent overfitting. To assess the performance of the model, we followed previous studies [11] and computed root mean square error (RMSE), mean absolute error (MAE), and directional accuracy (DA (%)).

The deep learning experiments were conducted using the Keras library with TensorFlow backend. The model was trained and tested on an NVIDIA Jetson AGX Xavier platform, featuring a 512-core Volta GPU and 32GB memory. This setup allowed efficient execution of transformer-based architectures and multimodal fusion mechanisms.

First, to evaluate the impact of different fusion mechanisms, we designed three variants of the proposed multimodal model using early fusion, late fusion, and hierarchical fusion. Each approach aims to integrate technical indicators, market sentiment, and text sentiment in the following ways. Early fusion strategy concatenates all modalities at the input level, feeding the merged vector into a unified neural network for processing. While simple, early fusion assumes equal contribution and scale compatibility across modalities. In the late fusion setup, each modality is processed independently through dedicated sub-networks. The outputs are fused at a later stage using an attention mechanism. This allows the model to learn modality-specific representations before integration. Finally, the hierarchical two-level approach first combines similar modalities (market and text sentiment) using attention, followed by fusion with the technical signal and ticker embedding at a higher level. Thus, it introduces a structured modality integration path.

The performance of each fusion strategy was evaluated on the test set using three metrics: RMSE, MAE, and DA (%), and the results are summarized in Table 3.

Table 3. Performance comparison of different fusion strategies.

Fusion strategy	RMSE	MAE	DA (%)
Late fusion	4.610	3.256	59.94
Hierarchical fusion	4.633	3.283	57.64
Early fusion	4.624	3.291	56.20

As shown in Table 3, the late fusion model achieved the best overall performance, particularly in terms of directional accuracy (59.94%), suggesting that allowing each modality to learn its representation independently before integration helps preserve their unique characteristics. Although the hierarchical fusion model introduces a more structured integration, its added complexity may not translate into performance gains. The early fusion model performed the worst in directional accuracy (56.20%), reinforcing the limitations of naive modality concatenation in financial multimodal learning.

To further evaluate the effectiveness of the proposed multimodal fusion model, we compared its performance with several deep learning baselines used in existing research, including LSTM, BiLSTM, and CNN-LSTM. Each baseline model processes the concatenated multimodal input through its respective architecture. LSTM is a basic unidirectional long short-term memory network applied to the concatenated input vector. BiLSTM represents a bidirectional version of LSTM that captures both forward and backward temporal dependencies. CNN-LSTM combines convolutional layers for local feature extraction with LSTM for sequential modeling.

As shown in Table 4, the proposed model achieves the best directional accuracy, indicating its superior ability to capture the correct direction of abnormal stock returns. While CNN-LSTM slightly outperforms our model in terms of RMSE, the higher directional accuracy of our model suggests that its attention-based late fusion architecture and ticker-aware integration provide meaningful improvements in trend prediction—an aspect more critical in financial forecasting applications.

Table 4. Comparison with existing multimodal deep learning models.

Model	RMSE	MAE	DA (%)
LSTM	4.661	3.300	53.03
BiLSTM	4.639	3.306	52.16
CNN-LSTM	4.609	3.276	55.91
Our model	4.610	3.256	59.94

To gain insights into the relative contribution of each modality in predicting CAR, we analyzed the average attention weights learned by the late fusion model over the entire test set. As shown in Figure 2, the market sentiment modality received the highest average attention weight, indicating that aggregated investor sentiment indices carry the most influential signals for CAR prediction in the short window surrounding earnings calls. The relatively strong contributions of all three modalities highlight the effectiveness of our attention-based late fusion mechanism in adaptively weighting each modality according to its relevance in the prediction context.

We further evaluated a prediction-based trading strategy, with daily buy/hold signals generated by applying a simple rule: buy when the predicted CAR is positive and sell otherwise. Our strategy was benchmarked against a passive buy-and-hold approach using several trading performance metrics. Results in Table 5 show that the model-driven strategy achieved an average yield of 1.9357% per signal versus 0.2959% for buy-and-hold. Furthermore, the strategy attained a Sharpe ratio of 0.4002, demonstrating that the predictive model can extract excess returns while maintaining a more favorable risk-return profile. Note that because the input features are aligned on a daily frequency, the model can be used at a daily minimum interval.

6. Conclusion

This study presents a novel multimodal deep learning framework for stock return prediction that integrates technical indicators, market investor sentiment, and textual sentiment derived from earnings conference call transcripts. By

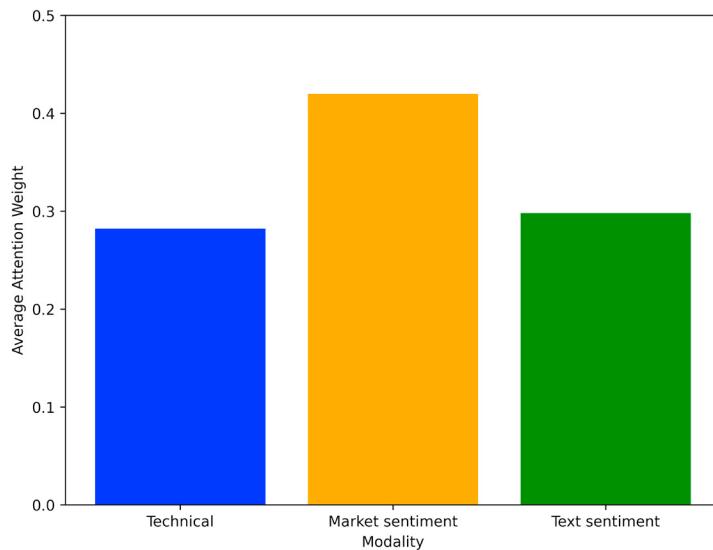


Fig. 2. Average modality attention weights over the testing data.

Table 5. Trading performance comparison for 2023-2024: prediction-based strategy vs. buy-and-hold.

Metric	Prediction-based strategy	Buy-and-hold
Average yield (%)	1.9357	0.2959
Sharpe ratio	0.4002	0.0569

combining transformer encoder for technical signals with neural network models for market and textual sentiment, and applying an attention-based late fusion mechanism, the model effectively captures diverse patterns from heterogeneous data modalities. Extensive experiments on a large dataset of 24,821 instances from 497 S&P 500 companies over the period 2010-2022 showed that our model outperforms several state-of-the-art multimodal fusion strategies. Attention analysis also revealed the relatively strong contributions of all three modalities, with market sentiment playing a slightly more prominent role in predicting short-term stock returns.

Despite these promising results, the study has several limitations that suggest avenues for future research. First, the model focuses primarily on structured sentiment and price-based data, leaving out other potentially informative modalities such as audio tones or visual charts. Second, while the attention mechanism provides some interpretability, further enhancements such as hierarchical or explainable attention levels could improve transparency. Finally, extending the analysis to international markets, high-frequency intraday data or sector-specific models could further validate and generalize the effectiveness of the proposed approach. Future research could also explore reinforcement learning-based trading strategies based on multimodal predictive signals.

Acknowledgements

This research is supported by the Czech Sciences Foundation [grant number 25-15405S] and the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences [grant number VEGA-1/0734/24].

References

- [1] Baker, M., and Wurgler, J. (2007). "Investor sentiment in the stock market." *Journal of Economic Perspectives* **21** (2): 129–151.

- [2] Hajek, P., and Novotny, J. (2024). "Beyond sentiment in stock price prediction: Integrating news sentiment and investor attention with temporal fusion transformer." In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Cham: Springer, pp. 30–43.
- [3] Du, K., Xing, F., Mao, R., and Cambria, E. (2024). "Financial sentiment analysis: Techniques and applications." *ACM Computing Surveys* **56** (9): 1–42.
- [4] Aggarwal, D. (2022). "Defining and measuring market sentiments: A review of the literature." *Qualitative Research in Financial Markets* **14** (2): 270–288.
- [5] Hsu, Y. T., Hua, M., Liu, H., and Wang, Q. (2024). "News sentiment and investment efficiency: Evidence from China." *European Financial Management* **30** (3): 1587–1617.
- [6] Reichenbach, F., and Walther, M. (2024). "Attention allocation of investors on social media: the role of prospect theory." *Journal of Behavioral Finance*: 1–18, doi: 10.1080/15427560.2024.2309145.
- [7] Kirtac, K., and Germano, G. (2024). "Sentiment trading with large language models." *Finance Research Letters* **62**: 105227.
- [8] Liu, W. J., Ge, Y. B., and Gu, Y. C. (2024). "News-driven stock market index prediction based on trellis network and sentiment attention mechanism." *Expert Systems with Applications* **250**: 123966.
- [9] Qin, Y., and Yang, Y. (2019). "What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL, pp. 390–401.
- [10] Li, Q., Tan, J., Wang, J., and Chen, H. (2020). "A multimodal event-driven LSTM model for stock prediction using online news." *IEEE Transactions on Knowledge and Data Engineering* **33** (10): 3323–3337.
- [11] Yuan, H., Tang, Y., Xu, W., and Lau, R. Y. K. (2021). "Exploring the influence of multimodal social media data on stock performance: an empirical perspective and analysis." *Internet Research* **31** (3): 871–891.
- [12] Das, R., and Singh, T. D. (2023). "Multimodal sentiment analysis: a survey of methods, trends, and challenges." *ACM Computing Surveys* **55** (13s): 1–38.
- [13] Gu, J. W., Zhong, Y. H., Li, S. Z., Wei, C. S., Dong, L. T., Wang, Z. Y., and Yan, C. (2024). "Predicting stock prices with finbert-lstm: Integrating news sentiment analysis." In *Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing*, pp. 67–72.
- [14] Li, J., Yang, L., Smyth, B., and Dong, R. (2020). "Maec: A multimodal aligned earnings conference call dataset for financial risk prediction." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3063–3070.
- [15] Wang, J., Hu, Y., Jiang, T. X., Tan, J., and Li, Q. (2023). "Essential tensor learning for multimodal information-driven stock movement prediction." *Knowledge-Based Systems* **262**: 110262.
- [16] Wu, J. M. T., Li, Z., Herencsar, N., Vo, B., and Lin, J. C. W. (2023). "A graph-based CNN-LSTM stock price prediction algorithm with leading indicators." *Multimedia Systems* **29** (3): 1751–1770.
- [17] Bu, Q. (2023). "Are all the sentiment measures the same?" *Journal of Behavioral Finance* **24** (2): 161–170.
- [18] Hajek, P., and Henriques, R. (2024). "Predicting M&A targets using news sentiment and topic detection." *Technological Forecasting and Social Change* **201**: 123270.
- [19] Wang, H., Xie, Z., Chiu, D. K., and Ho, K. K. (2025). "Multimodal market information fusion for stock price trend prediction in the pharmaceutical sector." *Applied Intelligence* **55** (1): 1–27.
- [20] Li, S., and Xu, S. (2025). "Enhancing stock price prediction using GANs and transformer-based attention mechanisms." *Empirical Economics* **68** (1): 373–403.
- [21] Iacovides, G., Konstantinidis, T., Xu, M., and Mandic, D. (2024). "FinLlama: LLM-Based Financial Sentiment Analysis for Algorithmic Trading." In *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 134–141.
- [22] Chen, Y., Fang, R., Liang, T., Sha, Z., Li, S., Yi, Y., Zhou, W., and Song, H. (2021). "Stock price forecast based on CNN-BiLSTM-ECA model." *Scientific Programming* **2021** (1): 2446543.
- [23] Sawhney, R., Agarwal, S., Wadhwa, A., and Shah, R. (2020). "Deep attentive learning for stock movement prediction from social media text and company correlations." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8415–8426.
- [24] Reis, P. M. N., and Pinho, C. (2021). "A reappraisal of the causal relationship between sentiment proxies and stock returns." *Journal of Behavioral Finance* **22** (4): 420–442.
- [25] Lee, T. W., Teisseire, P., and Lee, J. (2023). "Effective exploitation of macroeconomic indicators for stock direction classification using the multimodal fusion transformer." *IEEE Access* **11**: 10275–10287.
- [26] Chen, Y. F., and Huang, S. H. (2021). "Sentiment-influenced trading system based on multimodal deep reinforcement learning." *Applied Soft Computing* **112**: 107788.
- [27] Sun, L., Xu, W., and Liu, J. (2021). "Two-channel attention mechanism fusion model of stock price prediction based on CNN-LSTM." *Transactions on Asian and Low-Resource Language Information Processing* **20** (5): 1–12.
- [28] Emami Gohari, H., Dang, X. H., Shah, S. Y., and Zerfos, P. (2024). "Modality-aware Transformer for Financial Time series Forecasting." In *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 677–685.
- [29] Du, Z., Huang, A. G., Wermers, R., and Wu, W. (2022). "Language and domain specificity: A Chinese financial sentiment dictionary." *Review of Finance* **26** (3): 673–719.
- [30] Li, X., Wu, P., and Wang, W. (2020). "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong." *Information Processing & Management* **57** (5): 102212.
- [31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). "Attention is all you need." *Advances in Neural Information Processing Systems* **30**: 1–11.
- [32] Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). "Measuring news sentiment." *Journal of Econometrics* **228** (2): 221–243.
- [33] Araci, D. (2019). "Finbert: Financial sentiment analysis with pre-trained language models." *arXiv preprint arXiv:1908.10063*.