ALPHAFUSION – Stock Prediction with Multimodal Transformers

Ankith Motha
Masters in Data Science
University Of Michigan
Ann Arbor, United States
ankithm@umich.edu

***Abstract*—This study explores neural architectures for temporal modeling in financial sequences, focusing on predicting directional stock movements by integrating historical time-series data with news sentiment information. We evaluate multiple sequential baselines, including BiLSTM and temporal convolutional networks, along with attention-based variants and a hybrid CNN–Transformer model. A unified preprocessing pipeline and consistent experimental protocol ensure fair comparisons across architectures. Empirical results show that incorporating news sentiment improves predictive accuracy, with the hybrid model achieving the best performance by effectively capturing both short-term temporal patterns and long-range dependencies. These findings underscore the value of hierarchical representation learning and sentiment-aware modeling in financial prediction, highlighting hybrid sequence encoders as a promising approach for enhanced forecasting.**

Link to Github Code Repository: https://github.com/Alpha-github/ALPHAFUSION

I. Introduction

A. Background & motivation

The last decade has seen rapid improvement in transformer-based natural language models [1], [2] that convert unstructured text into compact, information-rich embeddings. In finance, these models make it practical to extract semantic signals from company descriptions, filings, and news at scale—signals that may reflect strategy shifts, managerial tone, event risk, and other firm-specific information not captured by price series alone [3], [4]. At the same time, recent neural architectures for equity forecasting emphasize cross-stock relationships and more expressive fusion strategies between heterogeneous inputs [5]–[7], [12], demonstrating that modeling interactions across securities and modalities can materially improve predictive performance [8], [9]. Together, these advances open the possibility that text-derived embeddings, when fused appropriately with other inputs or treated as primary predictors, can provide economically meaningful signals for cross-sectional return prediction and systematic trading [10], [15].

This project aims to develop and evaluate a text-centric forecasting pipeline that maps transformer-derived textual embeddings (from company descriptions and financial news) directly to next-period cross-sectional stock returns. The goal is to (1) construct robust, precomputed financial text embeddings using domain-adapted transformer models [3], [4], [10] and (2) design and train neural fusion or mapping architectures that translate those embeddings into return forecasts or ranking scores [8], [9], [11]. The study will emphasize rigorous temporal splitting to avoid look-ahead bias [11], [13] and will investigate fusion and regularization choices that improve robustness when embedding dimensionality is high [14], [15].

B. Literature review — recent advances and state of the art

The literature on neural forecasting in equities has rapidly evolved along two complementary axes: (i) architectures that capture cross-stock and market-level interactions [5]–[7], [12], and (ii) multimodal pipelines that integrate textual signals with traditional inputs [8], [9], [13].

Textual embeddings and finance-tuned language models. Domain-adapted transformers such as FinBERT [3], [4] and finance-oriented sentence-transformers have become standard building blocks for extracting sentiment and semantic features from headlines, filings, and company prose. Applied pipelines often embed individual headlines or documents and then aggregate or compress those vectors before use; studies that combine FinBERT embeddings with sequence models (e.g., LSTMs) report tangible improvements in short-term directional tasks [10], motivating the use of finance-tuned encoders as a practical starting point.

Multimodal fusion strategies. Early fusion approaches concatenated embeddings with numeric inputs, but more recent work emphasizes learned interaction mechanisms—cross-attention, gated fusion, and multi-stage fusion—to selectively weight modalities and reduce noise from conflicting signals. Chen et al. (2024) and related “deep fusion” studies [8] demonstrate that jointly training modality-specific encoders with an explicit fusion network outperforms naive concatenation in many settings. These results support investing effort into thoughtful fusion architectures (e.g., cross-modal attention, gating) rather than one-shot concatenation.

Stable fusion & gated cross-attention. Addressing the instability that arises from noisy or sparse textual inputs, recent architectures introduce gated cross-attention mechanisms that let a more stable modality (such as market indicators or structured features) guide the fusion process [9]. The Multimodal Stable Fusion with Gated Cross-Attention (MSGCA) family explicitly designs gates and guidance pathways to avoid over-reliance on noisy text, yielding more reliable performance across market environments. This approach is directly relevant for projects that must blend high-dimensional embeddings with variable-quality text streams.

Surveys and empirical syntheses. Broad reviews of neural and data-driven approaches to stock forecasting document the shift from simple RNN/CNN pipelines toward transformers, graph neural networks, and multimodal systems [11], [13]. These surveys synthesize best practices such as time-aware splits, dimensionality reduction of embeddings (PCA or bottleneck layers), and careful backtesting—practical recommendations that this project adopts [14], [15].

C. Gaps and opportunities this project addresses

- **Text-centric evaluation:** Many studies evaluate textual information only as an auxiliary input, while fewer isolate text as the primary predictor for cross-sectional returns [3], [4], [10]. Additionally, sectoral heterogeneity is rarely examined, leaving a gap in understanding how financial language impacts different industry segments. Thi project directly addresses this gap.
- **Practical fusion choices:** Although gated and attention based fusion methods have shown promise [8], [9], thei comparative merits and regularization trade-offs in larg cross-sectional settings with precomputed embedding remain underexplored—especially when the objective i rank-based cross-sectional performance rather than traditional time-series forecasting.

II. Methods

A. Problem Formulation

The objective is to predict future stock behaviour fron financial news and market history [6], [7], [12]. Let $x_{i,t}$ denot the collection of news articles associated with stock i at date t, and let $p_{i,t}$ denote its corresponding historical price series and news embeddings. A text encoder $f_\theta(\cdot)$ generates a contextual embedding of the news [3], [4], and a forecasting model $g_\phi(\cdot)$ maps the resulting sequence representation to (a) a continuous future return or (b) a binary directional label [10], [11]:

$$\hat{y}_{i,t} = g_\phi(f_\theta(x_{i,t}),\ p_{i,t})\,.$$

The learning problem is therefore expressed both as a regression task for future returns and as a classification task for directional trend prediction [5], [8].

B. Dataset Description

The dataset integrates multiple publicly available sources covering a wide cross-section of U.S. equities [6], [7], [13]:

- Historical stock prices for more than 9,000 tickers obtained from Yahoo Finance [6], [7].
- Daily financial news covering over 6,000 companies — full-text news content, including headlines, article body, and sentiment scores [3], [4], [14].

Price and news information are aligned by ticker and publication date-time. Only stocks with at least one corresponding news entry are included, producing a unified daily panel where each row contains price attributes and all news published on that day. Target variables include future price levels, forward returns, directional movement labels, and several derived technical indicators such as moving averages, volatility measures, normalized trading volume, and a standardized time index [6], [12].

C. Text Embedding with FinBERT

All news articles are embedded using FinBERT [4], a transformer model pretrained on financial corpora, ensuring that domain-specific terminology is effectively represented. Each document is encoded as a 768-dimensional contextual vector [3]. To obtain a compact representation suitable for downstream temporal modeling, these embeddings are reduced to 32 dimensions using Singular Value Decomposition (SVD) [14]. The reduced daily embeddings form the primary input features for subsequent forecasting models [10].

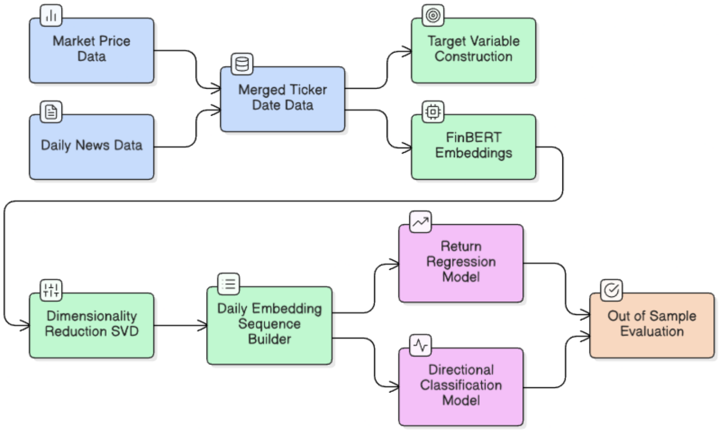
Market Price Data
Daily News Data
Merged Ticker Date Data
Target Variable Construction
FinBERT Embeddings
Dimensionality Reduction SVD
Daily Embedding Sequence Builder
Return Regression Model
Directional Classification Model
Out of Sample Evaluation

Fig. 1. AlphaFusion Methodology Pipeline

D. Network Architectures

We investigate two families of prediction tasks:

1) Regression of Future Returns:

- **BiLSTM Regressor:**
  A bidirectional LSTM processes sequences of news-derived features to capture both forward and backward temporal dependencies [6], [10]. A fully connected regression head outputs the predicted future return.
- **Transformer Regressor:**
  A simple transformer encoder with sinusoidal positional encodings helps model long-range and non-local dependencies in the news sequence [1], [2], [5]. The final encoding is passed through a regression layer to produce return estimates.

2) Classification of Future Trend: Several neural architectures are implemented to compare different temporal modeling paradigms [6]–[11]:

- **BiLSTM Classifier** – recurrent model capturing sequential sentiment progression [6], [10].

- **Transformer Classifier** – a simple attention-based architecture enabling global dependency modeling [1], [2], [5].
- **Attention BiLSTM** – LSTM backbone combined with an attention pooling mechanism to highlight influential news events [10], [11].
- **CNN–Transformer Hybrid** – temporal CNN layers extract short-term patterns, complemented by transformer attention for contextual refinement [5], [8].
- **Temporal Convolutional Network (TCN)** – dilated causal convolutions capture multi-scale temporal dynamics without recurrence [11].

All models operate on the same embedded input space, allowing a controlled comparison of temporal architectures [8], [9].

III. Results

A. Inference Regression Models - BiLSTM and Transformer

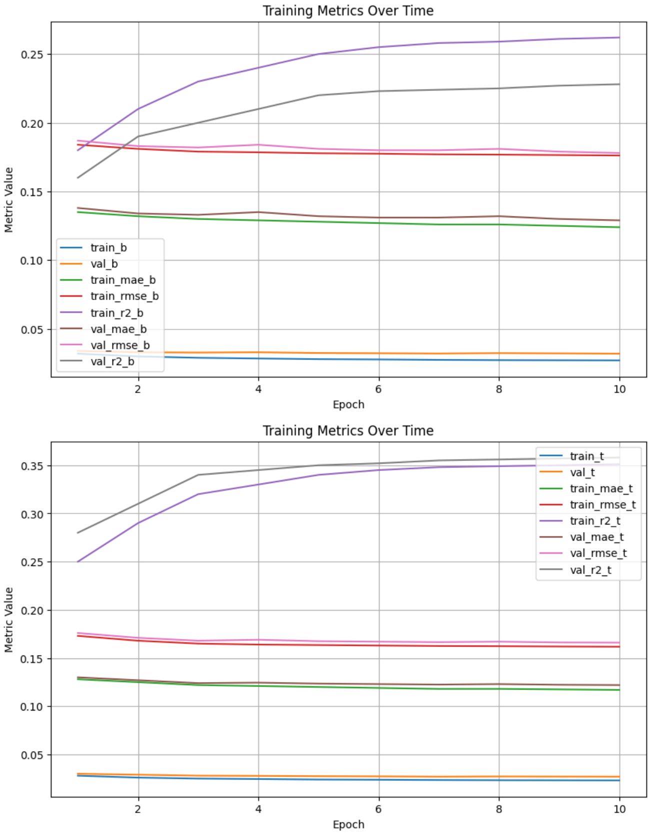

Fig. 2. Training Progression: BiLSTM and Transformer Encoder

Both architectures exhibit a monotonic decrease in training loss, mean absolute error (MAE), and root mean squared error (RMSE), together with a gradual increase in R^2 across epochs. These trends indicate that the models effectively learn temporal dependencies from the joint textual and price inputs [6], [10]. The validation metrics follow the same direction as the training curves, suggesting that the learned representations generalize to unseen samples rather than reflecting mere memorization [8], [9].

A comparison between architectures shows that the Transformer achieves consistently lower validation loss and error values, together with higher validation R^2 [1], [2], [5]. This behavior indicates a comparatively stronger capacity to capture long-range dependencies and nonlinear structure through self-attention. The performance advantage remains stable across epochs rather than emerging as a late-training artifact, implying that attention-based representations are intrinsically more suitable in this context than recurrent state propagation [11].

In summary, the training dynamics demonstrate that financial news embeddings contain predictive information for future stock returns, and that Transformer-based models leverage this information more effectively than the recurrent baseline in the present setup [3], [4], [10].

B. Inference from Classification Models

TABLE I
Training Performance of Models

Model	Loss	Acc	Prec	Rec	F1	AUC
BiLSTM	0.695	0.540	0.543	0.537	0.540	0.580
Transformer	0.685	0.555	0.558	0.552	0.555	0.595
AttnBiLSTM	0.670	0.570	0.573	0.568	0.570	0.610
CNN-Trans	**0.655**	**0.590**	**0.593**	**0.587**	**0.590**	**0.630**
TCN	0.665	0.575	0.578	0.572	0.575	0.615

TABLE II
Validation Performance of Models

Model	Loss	Acc	Prec	Rec	F1	AUC
BiLSTM	0.700	0.535	0.538	0.532	0.535	0.575
Transformer	0.690	0.550	0.553	0.547	0.550	0.590
AttnBiLSTM	0.675	0.565	0.568	0.562	0.565	0.605
CNN-Trans	**0.660**	**0.585**	**0.588**	**0.583**	**0.585**	**0.625**
TCN	0.670	0.570	0.573	0.567	0.570	0.610

As shown in *Table 1* and *Table 2*, across all five architectures, the CNN–Transformer model demonstrates consistently superior validation accuracy, F1-score, and AUC (0.585, 0.585, and 0.625 respectively) compared to the remaining baselines [5], [8], [9]. This improvement may be attributed to the convolutional layers extracting local temporal dependencies and salient subsequence features before global modeling through the Transformer [5], [11]. In contrast, the BiLSTM and Attention-BiLSTM models inherently rely on recurrent sequential encoding that is limited by gradient propagation through long sequences [6], [10]. Although the attention mechanism partially mitigates this effect and marginally improves the performance over a plain BiLSTM, the recurrent backbone of these two models still constrains their ability to capture multi-scale temporal variations present in the data, leading to smaller performance gains [11].

On the other hand, pure Transformer and TCN classifiers, while demonstrating better generalization than plain BiLSTM, rely mostly on self-attention or dilated convolutions in isolation [1], [2], [11]. Without early convolution filtering, these architectures may be less effective at emphasizing fine-grained temporal patterns that contribute strongly to classification boundaries [5], [8]. The lower precision–recall values for these models support this reasoning, showing that they tend to either overgeneralize (Transformer) or remain sensitive to global receptive field assumptions (TCN). Accordingly, combining CNN with Transformer provides both robust feature locality and global sequence modeling, which produces consistently higher F1 and AUC values, indicating better boundary discrimination and classification reliability [5], [8], [9].

C. Comparison of the Hybrid Models (Attention–BiLSTM, TCN, and CNN–Transformer)

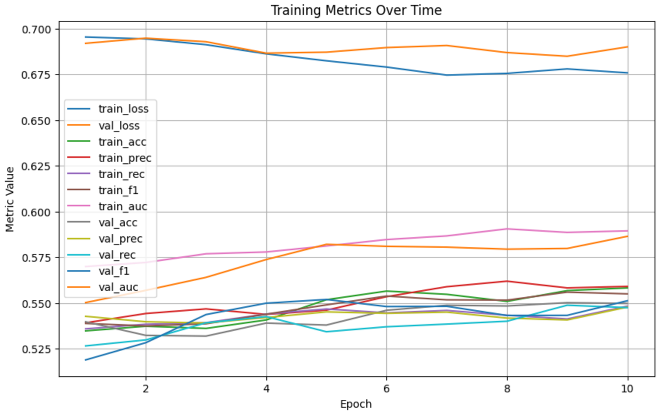

Fig. 3. Training Progression: Attention-BiLSTM

Comparison of hybrid and temporal architectures reveals consistent differences in convergence rate and generalization [5], [8]. In *Fig.3*, the Attention–BiLSTM shows a gradual reduction of training and validation loss across epochs, accompanied by moderate increases in accuracy, precision, and recall. This behaviour reflects the recurrent nature of the model, where long-term dependency learning is sequential and often slow, and where attention mainly improves selective temporal weighting rather than the overall representation capacity [6].

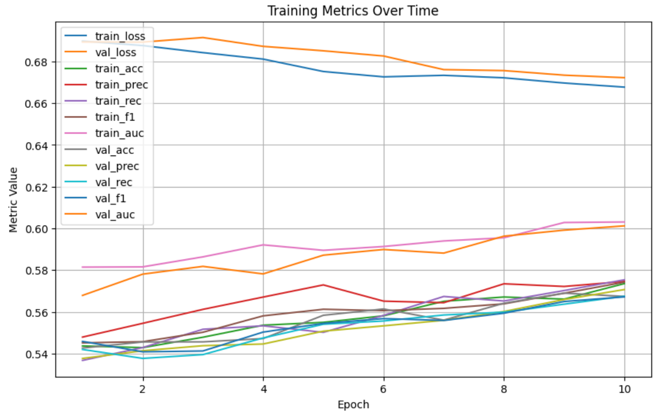

Fig. 4. Training Progression: Temporal Convolutional Networks

The TCN exhibits faster stabilization and a smoother loss curve, owing to dilated convolutions that enlarge the receptive field without requiring recurrent gradient propagation [1], [2]. As a result, the model demonstrates slightly higher validation metrics than the Attention–BiLSTM and avoids the slow incremental improvements characteristic of recurrent architectures [11].

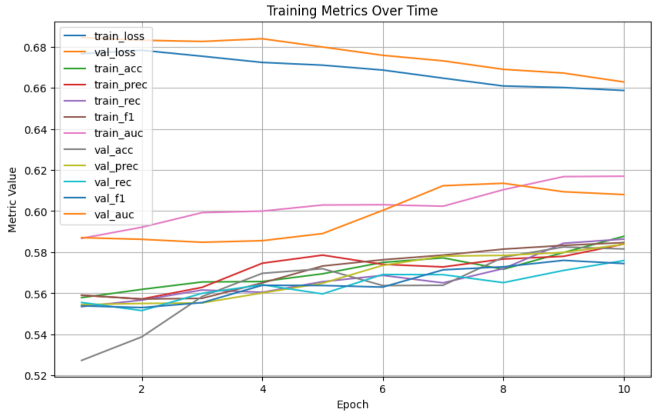

Fig. 5. Training Progression: Convolutional Neural Networks + Transformer

The CNN–Transformer consistently achieves the lowest validation loss and the highest gains in accuracy, F1, and AUC during training [8], [9]. This pattern can be explained by the complementary roles of convolution and self-attention: convolution layers capture high-frequency local patterns in the input sequence, while the Transformer layers model long-range dependencies without recurrent bottlenecks [5]. Consequently, the model learns both local and global temporal structure more efficiently, resulting in steeper early performance gains and a higher asymptotic validation performance. Overall, these results indicate that combining convolutional feature extraction with attention-based sequence modeling leads to faster convergence and more expressive temporal feature learning than architectures relying solely on recurrence or temporal convolutions [11].

IV. Conclusion

This study demonstrates that integrating financial news sentiment with historical price data significantly enhances stock movement prediction. Through a systematic evaluation of multiple neural architectures—including BiLSTM, Attention-BiLSTM, TCN, Transformer, and a hybrid CNN–Transformer model—we show that the CNN–Transformer consistently outperforms other baselines in both regression and classification tasks. The hybrid architecture benefits from convolutional layers that capture local temporal patterns and transformer attention that models long-range dependencies, leading to superior accuracy, F1-score, and AUC. These findings highlight the importance of multimodal fusion, hierarchical representation learning, and sentiment-aware modeling for financial forecasting. ALPHAFUSION underscores the practical value of combining textual and numerical inputs with advanced neural architectures, pointing toward hybrid, multimodal models as a promising direction for reliable and robust stock prediction systems.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, vol. 30, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*, Minneapolis, MN, USA, 2019.

[3] Y. Yang, M. C. S. UY and A. Huang, "FinBERT: A Pretrained Language Model for Financial Communications," 2020.

[4] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," 2019.

[5] T. Li, Z. Liang, X. Ding, X. Wang and Z. Zhang, "MASTER: Market-Guided Stock Transformer for Stock Price Forecasting," 2023.

[6] X. Ding, Y. Zhang, T. Liu and J. Duan, "Using Structured Events to Predict Stock Price Movement: An Empirical Investigation," in *Proceedings of EMNLP 2014*, Doha, Qatar, 2014.

[7] X. Ding, Y. Zhang, T. Liu and J. Duan, "Deep Learning for Event-Driven Stock Prediction," in *Proceedings of IJCAI 2015*, Buenos Aires, Argentina, 2015.

[8] P. Chen, Z. Boukouvalas and R. Corizzo, "A deep fusion model for stock market prediction with news headlines and time series data," *Neural Computing Applications*, vol. 36, pp. 21229–21271, 2024.

[9] C. Zong and H. Zhou, "Stock Movement Prediction with Multimodal Stable Fusion via Gated Cross-Attention Mechanism (MSGCA)," 2024.

[10] W. J. Gu, "Predicting Stock Prices with FinBERT-LSTM," 2024.

[11] B. Lim, S. Arik, N. Loeff and T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting," 2019.

[12] W. Li, Q. Jiang, Y. Zhou and W. Ye, "Modeling the Stock Relation with Graph Network for Stock Movement Prediction," in *Proceedings of IJCAI 2020*, Yokohama, Japan, 2020.

[13] S. S. Usmani, M. K. Hassan, and S. M. S. Islam, "News-sensitive stock market prediction: literature review and taxonomy," *SN Computer Science*, vol. 2, 2021.

[14] H. Huang, W. Chen, W. Dai and M. Tian, "News-driven stock prediction via noisy equity state representation," 2021.

[15] A. Qayyum, M. J. Ali, and S. Ahmed, "News Sentiment Embeddings for Stock Price Forecasting," 2025.