Dual-path Transformer: Aligning Text Embeddings with Market Movements for a New Paradigm of Cross-modal Financial Time Series Prediction

Haohan Zhang[1(✉)*], Hao Kong[2*], and Saizhuo Wang[2*]

[1] The Hong Kong University of Science and Technology (Guangzhou)
hzhang760@connect.hkust-gz.edu.cn
[2] The Hong Kong University of Science and Technology
{hkongab,swangeh}@connect.ust.hk

Abstract. The diversification of data modalities across various industries presents new challenges for time series prediction. Finance serves as a key example for this trend: quantitative data such as stock prices and financial indicators coexist with alternative data like sentiment embedded in news or social media. This raises a critical question: how can we fully harness the potential of this cross-modal data? In this paper, we propose a novel framework for cross-modal analysis, using finance as a representative case study.

We explore two distinct approaches to integrating these data types. First, we quantify the alternative data modality by using a neural network to align text embeddings directly with market movements. Second, we introduce a dual-path transformer architecture designed to capture the cross-modal attention between quantitative market data and text-based financial news. Finally, we demonstrate the effectiveness of this cross-modal approach through comprehensive back-testing, where results show that the dual-path transformer, leveraging effectively both modalities, outperform models using purely quantitative data within our experimental framework.

While the financial domain serves as our use case, the methodologies we develop are applicable to any scenario where multiple data modalities converge. By establishing this paradigm, we aim to provide key insights to this process.

1 Introduction

At the inception of financial time series predictions, a predominant school of thought, epitomized by the Efficient Market Hypothesis (EMH) [6], held that all necessary information for evaluating or predicting the performance of listed

* These authors contributed equally.

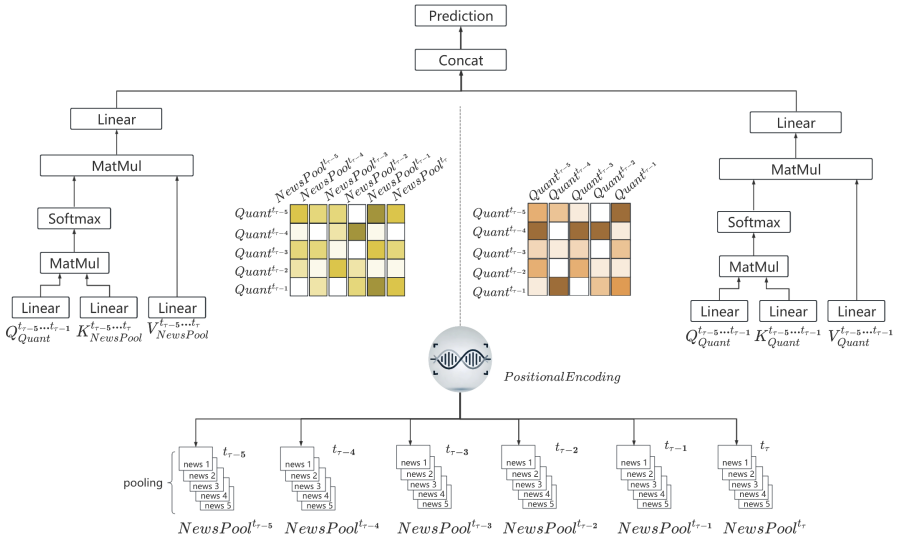

Fig. 1. The Dual-path Transformer Architecture

companies was inherently reflected in data directly related to stock prices. While this perspective has been foundational, the proliferation of multi-modal online data and the growing influence of social media discussions about listed companies have rendered a shift from this perspective inevitable. Traditional quantitative data (e.g., open price, close price, trade volume), once sufficient, is now increasingly unable to capture the full complexity of modern market climates, trends, and movements.

On one side, works in the realm of behavioral finance [8] [9] have argued that market sentiments—much of which is derived from discussions in textual formats—offer unique indicators that cannot be matched by quantitative data alone[7]. On the other side, researchers and practitioners are more and more incorporating alternative data, such as sentiment from news and social media, into their research and investment strategies to enhance their understanding of market behavior.

Given this background, financial time series prediction is no longer solely about analyzing numerical data or quantitative indicators; it is becoming increasingly cross-modal, requiring the integration of both quantitative and textual information to more accurately understand and predict market behavior. Henceforth, in this work we primarily study the sentiment extraction from textual information compounded with existing quantitative indicators as a baseline for cross-modal financial time series predictions.

Sentiment textual analysis was drastically transformed and innovated when LLMs such as ChatGPT[10] and BERT[11] were introduced. Now LLMs have

the ability to directly comprehend the content of news and generate a rating or evaluation within the context of future stock price movements. [12] and [13] became the first works to investigate the ability of ChatGPT to respectively predict U.S. and Chinese A stock price movements from analysis of news texts. For LLMs with BERT-like architectures, it is also possible to connect the embedding layers directly to a classifier and train the classifier with artificially constructed labels assigning each news item to a tier of sentiment from a finite set (e.g., neutral, positive, negative)[16].

Furthermore, [14] treated stock price fluctuations, news summary, macroeconomic summary as different data modules and employed GPT-4[15] to conduct analysis on an aggregation of these data modules to arrive at trading signals.

As for the task of time series prediction with deep learning, [18], a Transformer[27] variant, reached next-level performance on long time series predictions by innovating sentence patching and segmentation. Others use models based on diffusion[21]. In addition to QuantGan[22], [23] used a three-pronged approach to first use clustering algorithm to figure out dominant patterns within a financial series, used a Markov-like decision process[24] to learn how patterns transition from one to another and used a diffusion network to achieve artificial pattern generation.

Although much effort has been dedicated to this field, we nonetheless identify three main limitations within existing works.

1. **Sentiment mis-alignment**: Whether sentiment is rated by LLMs or quantified using traditional methods, most prior research evaluated sentiment based on semantic meaning rather than its implications for the actual market movements. For example, even if a news item may contain overwhelmingly positive words in the semantic sense, the news itself might not correlate with a strong upward growth of the company's price once it is brought within the context of the broader market movements. This disconnect between sentiment scoring and price fluctuations weakens the predictive power of models.
2. **Insufficient synergy between cross-modal data**: In many agent-based approaches, LLMs interpret textual information or sentiment is transformed into indicators, but there is little exploration of how textual data and quantitative time series interact or attend to each other. The absence of a mechanism that captures the cross-modal influence between these data types limits the models' ability to fully exploit their complementary nature.
3. **Simplistic back-testing environment and scarcity of data**: Many studies overlooked the importance of high-fidelity back-testing, limiting their practicality in financial forecasting. Additionally, works on textual sentiment analysis in finance often rely on datasets of relatively small scale, typically in the range of thousands to tens of thousands of text items, which does not accurately reflect the vast amount of data available in practice. While larger datasets can improve model performance, they also introduce more noise, which amounts significantly to more challenges. Effective models must therefore balance leveraging larger data sets while managing the complexity and noise that ensue.

Addressing these limitations, this work contributes in the following aspects:

1. **Direct market alignment**: We design a neural network that directly aligns news text embeddings with market movements. This allows sentiment analysis to be more closely tied to actual stock price fluctuations, providing a more accurate reflection of the influence of news on market behavior.
2. **Cross-modal synergy**: We introduce a dual-path transformer (the architecture of which is illustrated in Fig. 1) that utilizes a cross-attention mechanism between the textual embedding series and quantitative data series. This approach captures the interaction between sequences of these two modalities, enabling the model to fully exploit the complementary information from both sources. By working directly with sentence embeddings and employing the cross-attention mechanism, we develop insights from the ground up, bypassing the drawbacks of using LLMs as agents and relying solely on their interpretations which can often times lead to inaccuracies or hallucinations.
3. **Professional back-testing environment and realistic scale of textual data**: We employ a high-fidelity back-testing environment to demonstrate the effectiveness of our proposed models. This setup allows us to showcase tangible improvements of designed models (e.g., annual returns, sharpe rations, excess returns) in real-world trading scenarios. Additionally, the textual data used in this work for cross-modal analysis comprises tens of millions of textual data samples, far exceeding the scale typically used in existing studies. By working with such a large dataset, we ensure that our models are tested in conditions that more closely resemble the real-world data abundance and associated noise, allowing for a more rigorous and practical evaluation.

2 Data

The textual data used in this study for all of training, validation and testing the models consists of news summary about listed companies to the amount of 19,233,479 in total and spans the time period from March 2020 to February 2023. The data covers around 5,000 (this number is slightly larger than the current number of listed companies from both exchanges as some companies were de-listed during this period) listed Chinese companies from both the Shanghai Stock Exchange and Shenzhen Stock Exchange. For quantitative data about listed companies, we use the daily open, close, volume, money, high, and low information obtained from JoinQuant[19] about the listed companies mentioned in the news summary. In this work, we primarily investigate how news generated before market open affects the inter-day return on open prices, therefore we only keep news generated on each day before market open at 9:30 am. The data contains news from a total of 756 different sources (we use Figure 3 to illustrate the top 10 data sources by decreasing percentage) and for each item of news, we convert it to a 1024 dimensional embedding vector using the BGE-zh-1.5 Chinese sentence embedding conversion model [20].

The dataset is partitioned into four distinct subsets. The training set comprises stock price data and associated news items covering the period from March 1, 2020, to July 1, 2022. Subsequently, the first validation set spans from July 1, 2022, to August 1, 2022, and is primarily utilized to implement early stopping criteria (if loss on validation set does not improve after 5 epochs, early stopping criteria is met). A second validation set, ranging from August 1, 2022, to September 1, 2022, serves the purpose of stock selection within our universe based on the industry average Information Coefficient (IC). Finally, the period from September 1, 2022, to March 1, 2023, is designated as the test set, during which back-testing is conducted to assess the model's performance.

3 Quantification of Alternative Data: Sentiment Factor from Text Embedding Alignment

This work ultimately aims to introduce a dual-path transformer for cross-modality financial time series forecasting; however, quantifying alternative data first serves as a crucial intermediary step. By aligning sentiment derived from textual embeddings with numerical market data, we establish a measurable relationship between the two modalities. This step allows us to rigorously evaluate the impact of textual sentiment as an independent factor influencing market movements. To assess this, we implement two sets of vanilla transformer models that rely solely on self-attention: one with the addition of this sentiment factor and one without it. The distinct performance impact of sentiment as a quantified, traditional feature is further analyzed in subsequent sections.

To acquire this factor, we initially train a neural network to align text embeddings directly with future returns. Our focus centers on examining the relationship between news embeddings published before the market opens on a given trading day and the subsequent open returns between that day and the next. The underlying assumption is that certain news events occurring before the market opens could potentially influence significant market movements and these information cannot be captured by the quantitative factors that were available at market close on the day before.

We use a rolling window approach, for each trade day and for each listed company, we find 5 items of news summary embeddings generated on this same day before market open and thus acquire a 5 by 1024 matrix. We then flatten this matrix and pass it into a multi-layer perceptron (MLP) whose architecture, also illustrated in Figure 2, can be described as follows: The architecture of the Multilayer Perceptron begins with an input transformation, where the input matrix is flattened into a one-dimensional vector of 5120 elements to facilitate linear transformations. We employ a deep residual neural network to shrink the dimension from 5120 to 512, preparing the data for final output prediction. The final stage is an output layer that maps the 512-dimensional vector to a single scalar value, representing the predictive outcome. Upon training the MLP on the training set and meeting the stopping criteria on Validation I, we observe a distinctive pattern wherein the Information Coefficient (IC) varies significantly

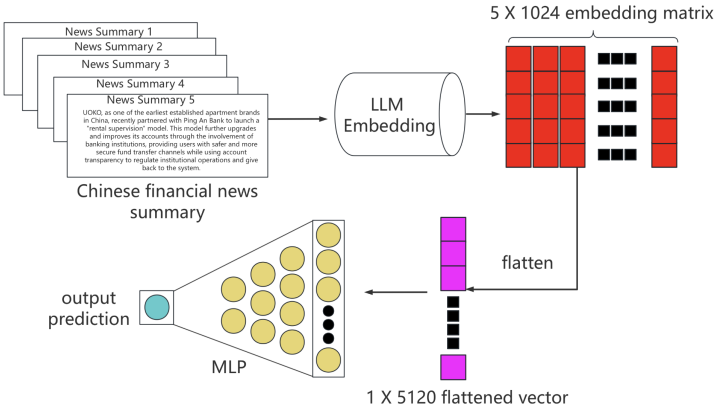

Fig. 2. Architecture of news embedding to return alignment (the original news texts are Chinese, here they are translated into English for illustration purposes)

across different industries. This variation can be attributed to the broad differences in the inherent characteristics of companies across industries, which result in varying levels of sensitivity to sentiment. For the stock universe we construct for subsequent experiments, we select representative stocks from each industry in proportion to their average IC performance on Validation II so that industries with higher IC contribute more to the final selection. In Figure 4, we show the average sentiment of different industries on the second validation set.

4 Dual-path Transformer for Concurrent Multi-modal Processing

4.1 Dual-path Transformer's Architecture

In this work, we aim to investigate the methodology for conducting cross-modal time series prediction in finance from the ground up, rather than employing LLMs agents. Our architecture is designed to concurrently process sequences of quantitative trading data—including metrics such as open, close, high, low prices, volume, and money—and corresponding sequences of news text embeddings. This dual-input approach enables a comprehensive analysis by leveraging the inherent temporal and contextual relationships both within and across these distinct data modalities. The overall architecture is illustrated in Figure 1.

Suppose we were making the prediction of open price return between $t_{\tau+1}$ and t_τ. Starting with the self-attention branch, illustrated by the right part of Figure 1, we start with the sequence of quantitative data (e.g., open, close, volume, money) of $t_{\tau-5}...t_{\tau-1}$ as an input, raise the dimension of this input into 512 dimensions and apply a linear positional encoding to maintain the temporal sequence integrity. Then, we add the CLS token ahead of the sequence and pass

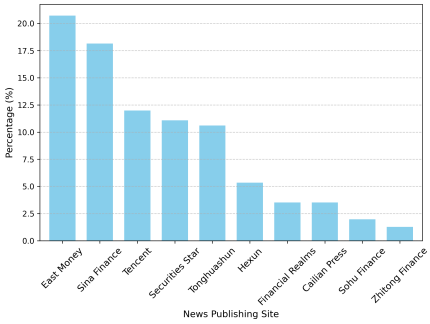

Fig. 3. Top 10 news publishing sites by percentage

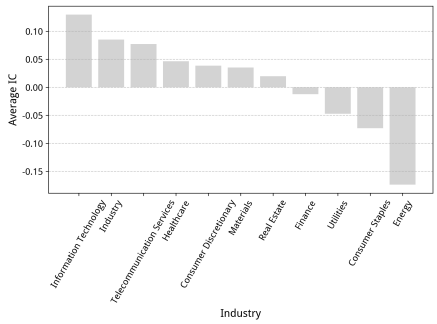

Fig. 4. IC average by industry on Validation II

the whole sequence into a multi-headed self-attention, during which the model learns how the various aspects of quantitative data influence each other and is essentially computing an attention filter between different time-steps of the quantitative sequence.

This is followed by the application of multi-headed self-attention, a mechanism that computes the attention between different time-steps of quantitative data. It is expected that such a procedure may uncover and understand both the short-term fluctuations and the long-term trends present in the financial market data, enabling it to identify patterns and dependencies that are not immediately obvious.

For the cross-attention branch, we begin by collecting all the news taking place between $t_{\tau-5}$ and t_τ. This is slightly different than the quantitative series because since the news we use are before market open, we can include the news generated on t_τ itself. To summarize the influence of news taking place on each day, we take the average pooling of embeddings of news occurred on the same day to form the news embedding sequence. How cross attention works in conjunction with self attention will be explained in subsequent sections. At the center of Figure 1, we conceptualize two attention heat maps, corresponding to cross attention and self attention which illustrates how much entries within input sequences attend to each other.

4.2 Injection of Positional Encoding into Text Embedding

To effectively capture the timing of news events, our model uses a custom positional encoding that translates the publication time of each news item into minutes past midnight. This encoding leverages two sinusoidal functions: a 60-minute cycle for capturing minute-level fluctuations and a broader 570-minute cycle to reflect the entire trading day. For a nuanced integration of these temporal signals, we assign the sum of short-term sine functions to even dimensions and the sum of corresponding cosine functions to odd dimensions of the embeddings. This dual-period approach allows the model to recognize both immediate

and extended temporal influences within the news data, effectively encoding the time-sensitive nature of news impact on the market.

4.3 The Forward Propagation Process

At this juncture, we have two sequences: one of quantitative data featuring metrics such as open, close, and so on, and the other of pooled daily news embeddings. For the cross-modal cross attention mechanism, the quantitative data serving as queries (Q) and the news embeddings as keys (K) and values (V). This setup allows the model to interrogate, "Given the current and past market conditions, what news information is most relevant?" By treating the quantitative sequence as queries, the model dynamically identifies and emphasizes the news information that is most pertinent to the given market conditions. The keys and values, derived from the news embeddings, represent the universe of potential impacts that news events could have on market movements. The attention scores calculated during this process highlight the relevance of each news item to the market conditions, enabling the model to synthesize a weighted representation of news information that is most likely to influence future market behaviors.

The proposed model incorporates both self-attention for quantitative data and cross-attention between quantitative data and news embeddings. These mechanisms are crucial for capturing both internal dependencies within the quantitative data and interactions between quantitative and textual information. For the quantitative data sequence X_{quant}, the multi-head attention (MHA) mechanism computes queries, keys, and values using learned weight matrices for each attention head as follows:

$$Q_{\text{quant},i} = W_{Q_{\text{quant},i}} X_{\text{quant}},\ K_{\text{quant},i} = W_{K_{\text{quant},i}} X_{\text{quant}},\ V_{\text{quant},i} = W_{V_{\text{quant},i}} X_{\text{quant}}$$

where $W_{Q_{\text{quant},i}}, W_{K_{\text{quant},i}}, W_{V_{\text{quant},i}}$ are the learned query, key, and value matrices for the i-th head. The attention scores for each head are calculated as:

$$A_{\text{quant},i} = \text{softmax}\left(\frac{Q_{\text{quant},i} K_{\text{quant},i}^{\top}}{\sqrt{d_k}}\right) V_{\text{quant},i}$$

where d_k is the dimensionality of the key vectors. The outputs of all heads are concatenated and passed through a linear projection:

$$\text{MHA}_{\text{self}}(X_{\text{quant}}) = \text{Concat}(A_{\text{quant},1}, A_{\text{quant},2}, \ldots, A_{\text{quant},h}) W_{O_{\text{quant}}}$$

where $W_{O_{\text{quant}}}$ is the output projection matrix. The final self-attention output for the quantitative data sequence is obtained through recursive layers, each applying attention and feed-forward processing as:

$$H_{\text{quant}}^{l} = \text{LN}(\text{MHA}_{\text{self}}(Z_{\text{quant}}^{l-1})) + Z_{\text{quant}}^{l-1}, \quad Z_{\text{quant}}^{l} = \text{LN}(FF(H_{\text{quant}}^{l})) + H_{\text{quant}}^{l}$$

where $Z^0_{\text{quant}} = X_{\text{quant}}$, and this recursion continues for each layer l. Layer Normalization (LN) in this context operates by normalizing the input as:

$$\text{LN}(x) = \frac{x - \mu(x)}{\sigma(x) + \epsilon} \cdot \gamma + \beta$$

where $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation of the input x, and γ and β are learned scaling and shifting parameters.

Similarly, cross-attention is applied between the quantitative data X_{quant} and news embeddings X_{news}. The queries are computed from the quantitative data, and keys/values from the news embeddings:

$$Q_{\text{quant},i} = W_{Q_{\text{quant},i}} X_{\text{quant}}, \; K_{\text{news},i} = W_{K_{\text{news},i}} X_{\text{news}}, \; V_{\text{news},i} = W_{V_{\text{news},i}} X_{\text{news}}$$

The cross-attention scores are computed as:

$$A_{\text{cross},i} = \text{softmax}\left(\frac{Q_{\text{quant},i} K^\top_{\text{news},i}}{\sqrt{d_k}}\right) V_{\text{news},i}$$

The output from the cross-attention is then:

$$\text{MHA}_{\text{cross}}(X_{\text{quant}}, X_{\text{news}}) = \text{Concat}(A_{\text{cross},1}, A_{\text{cross},2}, \ldots, A_{\text{cross},h}) W_{O_{\text{cross}}}$$

This cross-attention is applied recursively across layers, with the final layer producing the updated quantitative sequence:

$$H^l_{\text{cross}} = \text{LN}(\text{MHA}_{\text{cross}}(Z^{l-1}_{\text{quant}}, X_{\text{news}})) + Z^{l-1}_{\text{quant}}, \; Z^l_{\text{quant}} = \text{LN}(FF(H^l_{\text{cross}})) + H^l_{\text{cross}}$$

The recursive application of self-attention and cross-attention allows the model to capture both intra-quantitative dependencies and the interaction between quantitative and news data, making it highly flexible in incorporating cross-modal information across layers.

4.4 Baseline Model: Vanilla Transformer with Self-Attention on Quantitative Series

In addition to evaluating the dual-path Transformer, we establish a baseline model using a vanilla transformer architecture that solely processes quantitative sequential data. This baseline model operates on a rolling window of quantitative features, capturing recent price and volume movements within a fixed period. Unlike the dual-path Transformer, which leverages cross-attention to incorporate sentiment and news data, the baseline model applies only self-attention to the quantitative series, thus focusing exclusively on internal dependencies within the financial time series. By comparing the performance of the dual-path Transformer against this quantitative-only vanilla transformer, we assess the added

value of cross-modal analysis for capturing sentiment-driven market dynamics. To illustrate the effect of sentiment factor through quantification of alternative data, we use two sets of vanilla transformers one with the sentiment factor derived in Section 4 one relying solely on the quantitative market features.

5 Experimental Results

In this section, we employ our back-testing system to evaluate the effectiveness of the proposed dual-path transformer model compared to several temporal models, including a transformer model with and without the sentiment factor, across a consistent experimental setting. Using data from listed companies from March 2020 to February 2023, we designated March 2020 to August 2022 as the training and validation set, September 2022 and after as the test set.

Unlike traditional factor ranking-based stock portfolio adjustment methods, which rank stocks at each trading interval and trade based on calculated factors, our approach uses time series forecasting for individual stocks. At the beginning of the simulation, an equal buy-in is conducted for all stocks in the selected pool. Each stock is then assigned a threshold buy or sell value, with trades triggered when the forecasted factor meets these thresholds.

For stock selection, we calculate the sequential information coefficient between each stock's factors and its future returns on the training set. We then select the top 250 stocks with the strongest information coefficients, ensuring a tradable and effective stock pool. During back-testing, equal capital allocation is applied, distributing the initial capital equally among these 250 stocks. A long position is opened for each stock at its long entry threshold and held until the long exit threshold is met. Transaction costs are set at 0.1% for both buys and sells, which are crucial for obtaining more accurate backtesting results.These costs play an important role in simulating realistic trading scenarios, as they reflect the impact of real-world market frictions on the profitability of strategies. We select the CSI300 as the benchmark due to its alignment with the characteristics and style of our chosen stock universe, ensuring a comparable and relevant performance standard. In addition to incorporating commonly used financial time series models, such as LSTM[26] and Transformer[27], for comparison, we also integrate more state-of-the-art time series models. Specifically, we employ the Temporal Convolutional Network (TCN)[28], which incorporates convolutional structures into time series analysis, as well as the Mixer model, which leverages information mixing techniques to capture complex dependencies in time series data. The results, as shown in Table 1 In addition to calculating the Information Coefficient (IC), which measures the strength of the relationship between predicted and actual returns, we also compute several other key performance metrics that are essential for evaluating the practical effectiveness of the trading strategy. Some of these metrics assess the strategy's ability to generate returns while managing associated risks, such as: AR (Annualized Return),RetPerTrade (Average Return per Trade),SR (Sharpe Ratio),MD (Maximum Drawdown), and Win Rate. Other metrics provide valuable insights into

Table 1. Performance metrics for different models

Model	IC	AR	SR	Bars	MD	OCC	Ret/Trade	Trades	Win
Transformer (wo/ sentiment)	0.0634	0.0496	0.6040	1.5972	-0.0533	0.3349	0.0034	2346	0.4974
Transformer (w/ sentiment)	0.0633	0.0830	1.1472	1.8346	**-0.0418**	0.3364	0.0032	2164	0.4556
Dual-path Transformer	**0.0998**	**0.2519**	**2.8149**	**2.0882**	-0.0425	0.3203	**0.0067**	1961	0.4936
Linear	0.0811	0.0773	0.8541	0.5945	-0.0481	0.3222	0.0043	**3467**	0.5169
MLP	0.0752	0.0980	1.1218	0.8920	-0.0457	**0.3546**	0.0037	3332	0.4970
LSTM	0.0826	0.1203	1.3079	1.1590	-0.0521	**0.3736**	0.0027	3151	0.4903
Mixer	0.0710	0.0562	0.7514	0.6882	-0.0566	0.3103	0.0009	3111	0.4851
TCN	0.0804	0.1079	1.2167	0.8752	-0.0515	0.3497	0.0034	3340	**0.5054**

the operational aspects of the strategy, particularly in terms of position management and capital utilization, including Bars (Average Holding Period),OCC (Capital Occupation Rate), and TradeNum (Number of Trades). These metrics allow us to assess the efficiency with which capital is allocated and the typical duration of positions held, providing a fuller picture of the strategy's operational dynamics.Compare signal metrics and portfolio performance across models, the dual-path transformer demonstrates a marked improvement over traditional models, including the vanilla transformer without sentiment. The dual-path transformer consistently achieved higher Information Coefficient(IC), Annualized Returns (AR), Sharpe Ratios (SR), and lower Maximum Drawdown (MD), showing resilience across various market conditions. The model's high Sharpe ratio and controlled maximum drawdown indicate that it is particularly effective at capturing profitable signals while mitigating risk, a crucial aspect for practical trading applications.The improvement is statistically significant based on the Diebold–Mariano test[25], as shown in Table 2.

The addition of the sentiment factor also shows clear benefits, as the vanilla transformer model with the sentiment factor outperforms its counterpart without sentiment in terms of portfolio metrics. This improvement validates the influence of sentiment on market behavior and highlights the value of integrating alternative data into quantitative models. While the transformer with sentiment factor demonstrates gains in both return and Sharpe ratio, the dual-path transformer further amplifies these benefits by incorporating a cross-attention mechanism to dynamically interpret sentiment's impact on price movements.

The cross-attention mechanism itself plays a pivotal role, enabling the dual-path transformer to capture nuanced insights from news data effectively. This process enhances the trading signal quality by aligning news sentiment with price trends, allowing the model to focus on the most impactful interactions within the quantitative series. Consequently, this cross-modal approach produces a robust basis for informed trading decisions, capturing sentiment-driven market patterns that contribute to a stable and effective net value curve across the test period, as shown in Figure 5.

Haohan Zhang(✉)*, Hao Kong*, and Saizhuo Wang*

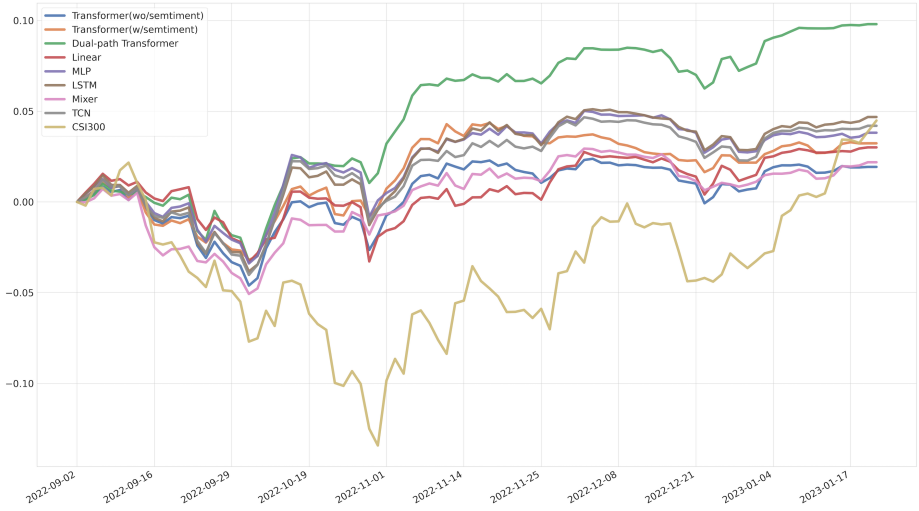

Fig. 5. Back test results with trading cost 0.001

Table 2. Comparison of out-of-sample prediction using Diebold-Mariano tests

Model	Transformer (w/ sentiment)	Transformer (wo/ sentiment)	Linear	LSTM	Mixer	MLP	TCN
Dual-path Transformer	**-9.8730**	**-10.0641**	**-4.4749**	-0.1929	**-12.9882**	**-3.8986**	0.8477
Transformer (w/ sentiment)		0.2132	**5.0407**	**9.3082**	**-3.9069**	**5.9264**	**10.0572**
Transformer (wo/ sentiment)			**4.9882**	**9.4112**	**-3.9642**	**5.9007**	**10.2154**
Linear				**4.4411**	**-8.8724**	0.7686	**5.4310**
LSTM					**-12.5111**	**4.3709**	1.4535
Mixer						**-9.2751**	**13.3196**
MLP							**5.7370**
TCN							

Note: This table reports the pairwise Diebold-Mariano test statistics comparing the out-of-sample prediction performance among the models. Positive values indicate that the column model outperforms the row model, while negative values indicate that the row model outperforms the column model. Bold numbers indicate that the difference is significant at the 5% level or higher in individual tests.

6 Interpreting the Cross-Attention Mechanism through Visualizing the Attention Heatmap

In this study, we visualize the attention filters within the cross-attention mechanism of our model, designed to capture interdependencies between quantitative time series data and corresponding news embeddings. This visualization provides interpretability by revealing how the model selectively attends to different aspects of these data sources during prediction, offering insights into the complex interactions between structured financial data and unstructured textual information.

To achieve this, we extract attention weights from the cross-attention layers post-inference, where the quantitative series serves as the query (Q), while news embeddings act as both keys (K) and values (V). The resulting attention weights indicate how strongly each attention head aligns specific stock market time steps with relevant news embeddings. Each attention head produces a distinct attention map, where:

- Rows correspond to time steps in the quantitative stock sequence.
- Columns represent news embeddings, pooled at the daily level.

By analyzing these attention maps, we can identify which news events exert greater influence on specific trading days, helping the model dynamically weigh different information sources. Importantly, this structure enables the model to synthesize meaningful patterns across multiple attention heads, enhancing its ability to predict stock movements with greater contextual awareness.

As an example, Figure 6 illustrates an attention heatmap from our model's prediction for Chanhen Chemical Corporation leading up to December 12, 2022. The heatmaps, spanning eight attention heads, reveal distinct attention patterns, where news events occurring near the mid-point of the sequence exhibit the strongest impact on stock price movements, as indicated by bright yellow regions. The variation across attention heads suggests that different aspects of news relevance are captured separately, helping the model distinguish between impactful and less relevant market information. This multi-head structure ensures a more robust and interpretable prediction process, allowing for an improved understanding of how textual information influences stock returns.

7 Conclusions

In this study, we introduced a dual-path transformer framework designed to align textual and quantitative data for enhanced financial time series prediction. Through a novel cross-attention mechanism, we demonstrated that our approach captures the intricate dependencies between news sentiment and stock market indicators, addressing a gap in traditional predictive models that often treat these data modalities separately. Our results, supported by rigorous back-testing and proven as statistically significant, illustrate that this integrated approach

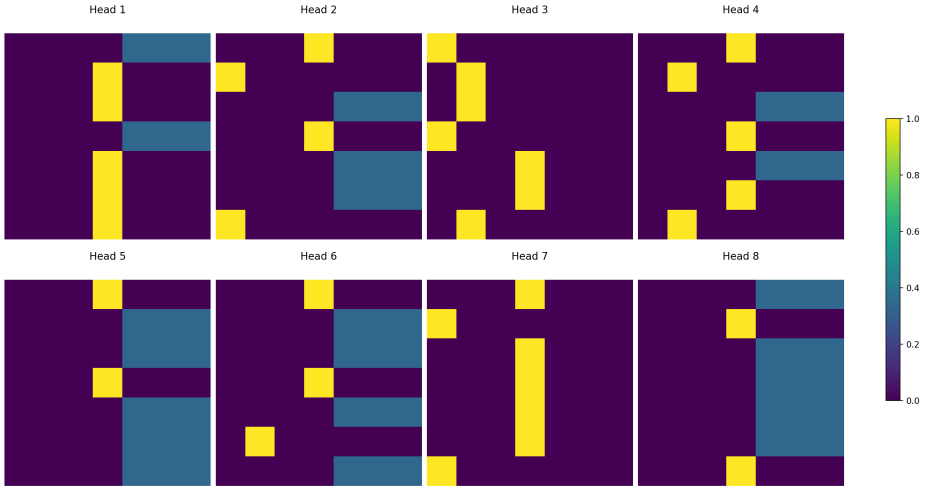

Fig. 6. Attention heat-map of cross-modal data

substantially enhances the predictive power over baseline models that rely solely on either quantitative or sentiment data.

The proposed framework's ability to dynamically align news and price movements provides new insights into the impact of textual information on financial trends, overcoming limitations related to sentiment misalignment and insufficient cross-modal synergy. Furthermore, by leveraging a large-scale dataset and an advanced back-testing environment, we ensured that our approach is both practical and applicable to real-world financial environments, thereby bridging the gap between research and industry application.

This work not only validates the effectiveness of cross-modal integration for financial predictions but also sets the stage for further exploration of cross-modal architectures in domains where multiple data streams converge. Future research could explore additional forms of alternative data and refine the cross-modal attention mechanism to improve robustness across different market conditions.

8 Future Work

There remains several promising directions future exploration. First, our results indicate that the influence of sentiment varies across industries, suggesting that industry-specific characteristics impact the extent to which textual information drives stock movements. To better capture these variations, future research could incorporate industry embeddings or individual stock embeddings as additional input features, enabling the model to differentiate the impact of sentiment across different market sectors. By embedding industry-level or company/level dynamics, the model could improve its predictive capabilities and further refine the alignment between textual sentiment and market behavior.

Second, while our study focuses on the Chinese stock market and financial news written in Chinese, we anticipate the same cross-modal dynamics to hold across different markets, including those dominated by English-language financial news. The interaction between news sentiment and stock movements is a fundamental aspect of financial markets, independent of linguistic differences. Thus, the next step is to systematically incorporate English-language financial news and conduct equivalent experiments to validate our findings across broader financial contexts. This expansion will further reinforce the universality of our approach and demonstrate its applicability to global markets.

By extending the model to incorporate industry-aware features and validating its performance across global markets, future work can further enhance the applicability and effectiveness of cross-modal financial forecasting frameworks.

References

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, http://www.springer.com/lncs, last accessed 2023/10/25
6. Fama, E.F.: The Behavior of Stock-Market Prices. Journal of Business 38(1), 34–105 (1965).
7. Baker, M., Wurgler, J.: Investor Sentiment in the Stock Market. Journal of Economic Perspectives (2007).
8. Hilton, D.: The Psychology of Financial Decision-Making: Applications to Trading, Dealing, and Investment Analysis. The Journal of Psychology and Financial Markets 2, 37–53 (2001). https://doi.org/10.1207/S15327760JPFM0201_4
9. Nofsinger, J.: Social Mood and Financial Economics. Journal of Behavioral Finance 6, (2003). https://doi.org/10.1207/s15427579jpfm0603_4
10. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language Models are Few-Shot Learners. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), pp. 1877–1901. Curran Associates, Inc., (2020).
11. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, MN (2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
12. Lopez Lira, A., Tang, Y.: Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. arXiv preprint arXiv:2304.07619 (2023). https://doi.org/10.48550/arXiv.2304.07619
13. Zhang, H., Hua, F., Xu, C., Guo, J., Kong, H., Zuo, R.: Unveiling the Potential of Sentiment: Can Large Language Models Predict Chinese Stock Price Movements? arXiv preprint arXiv:2306.14222 (2023). https://doi.org/10.48550/arXiv.2306.14222

14. Fatouros, G., Metaxas, K., Soldatos, J., Kyriazis, D.: Can Large Language Models Beat Wall Street? Unveiling the Potential of AI in Stock Selection. SSRN Electronic Journal (2024). https://doi.org/10.2139/ssrn.4693849
15. OpenAI: GPT-4 Technical Report. OpenAI (2023). arXiv preprint arXiv:2303.08774, https://arxiv.org/abs/2303.08774
16. Zhang, J., Gan, R., Wang, J., Zhang, Y., Zhang, L., Yang, P., Gao, X., Wu, Z., Dong, X., He, J., Zhuo, J., Yang, Q., Huang, Y., Li, X., Wu, Y., Lu, J., Zhu, X., Chen, W., Han, T., Pan, K., Wang, R., Wang, H., Wu, X., Zeng, Z., Chen, C.: Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. CoRR abs/2209.02970 (2022).
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), pp. 6000–6010. Curran Associates Inc., Long Beach, CA, USA (2017).
18. Nie, Y., Nguyen, N., Sinthong, P., Kalagnanam, J.: A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. arXiv preprint arXiv:2211.14730 (2022). https://doi.org/10.48550/arXiv.2211.14730
19. JoinQuant. https://www.joinquant.com/. Accessed: 2023-05-01
20. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N.: C-Pack: Packaged Resources to Advance General Chinese Embedding. arXiv preprint arXiv:2309.07597 (2023).
21. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems (NeurIPS 2020), pp. 6840–6851. Curran Associates Inc., Virtual (2020).
22. Wiese, M., Knobloch, R., Korn, R., Kretschmer, P.: Quant GANs: Deep Generation of Financial Time Series. In: Proceedings of the 36th International Conference on Machine Learning (ICML 2019). Curran Associates Inc., Long Beach, CA, USA (2019). https://doi.org/10.48550/arXiv.1907.06673
23. Huang, H., Chen, M., Qiao, X.: Generative Learning for Financial Time Series with Irregular and Scale-Invariant Patterns. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024). OpenReview, Virtual (2024). https://openreview.net/forum?id=CdjnzWsQax
24. Dueker, M.J.: Markov Switching in GARCH Processes and Mean-Reverting Stock-Market Volatility. Journal of Business & Economic Statistics 15(1), 26–34 (1997).
25. Diebold, F.X., Mariano, R.S.: Comparing predictive accuracy. J. Bus. Econom. Statist. 20(1), 134–144 (2002).
26. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (1997).
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30. Curran Associates, Inc. (2017).
28. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018).