MCASP: Multi-Modal Cross Attention Network for Stock Market Prediction

Kamaladdin Fataliyev and **Wei Liu**
University of Technology Sydney, Sydney, Australia
kamaladdin.fataliyev@student.uts.edu.au, wei.liu@uts.edu.au

Abstract

Stock market prediction is considered a complex task due to the non-stationary and volatile nature of the stock markets. With the increasing amount of online data, various information sources have been analyzed to understand the underlying patterns of the price movements. However, most existing works in the literature mostly focus on either the intra-modality information within each input data type, or the inter-modal relationships among the input modalities. Different from these, in this research, we propose a novel Multi-Modal Cross Attention Network for Stock Market Prediction (MCASP) by capturing both modality-specific features and the joint influence of each modality in a unified framework. We utilize financial news, historical market data and technical indicators to predict the movement direction of the market prices. After processing the input modalities with three separate deep networks, we first construct a self-attention network that utilizes multiple Transformer models to capture the intra-modal information. Then we design a novel cross-attention network that processes the inputs in pairs to exploit the cross-modal and joint information of the modalities. Experiments with real world datasets for S&P500 index forecast and the prediction of five individual stocks, demonstrate the effectiveness of the proposed multi-modal design over several state-of-the-art baseline models.

1 Introduction

Stock market movements are inherently affected by a multitude of data sources, encompassing historical price data, technical indicators (Vargas et al., 2017), financial news (Schumaker and Chen, 2009), social media (Chen et al., 2018), and official announcements (Feuerriegel and Gordon, 2018). It has been established that analyzing these multiple data modalities together enables the capture of underlying patterns in stock movements, rendering stock market prediction a multi-modal learning task (Akita et al., 2016). The efficacy of employing effective multi-modal representation and learning techniques to uncover the joint influence of these data modalities is pivotal for model performance (Li et al., 2020). Simultaneously, it is important to extract the intra-modal information within each data source. Early information fusion techniques combine raw input features initially and then construct a prediction model, which aids in capturing the combined influence of modalities but neglects intra-modal information. Late fusion techniques, conversely, analyze input features separately and subsequently employ a fusion layer for prediction. While this approach facilitates a focus on modality-specific features, it may overlook inter-modal information. Balancing the capture of intra-modal and inter-modal information from input modalities is essential.

Researchers have identified that pairs of data modalities, such as financial news and market prices, as well as market prices and technical indicators (Vargas et al., 2017), both impact price movements. However, existing models, while striving to capture the joint influence of all modalities together, may overlook the underlying bi-modal relationships between various data inputs. Therefore, in addition to capturing their collective influence, it is also crucial to understand the bi-modal relationships among pairs of input modalities.

To address these challenges, various methods have been developed, primarily categorized as inter-modality and intra-modality-based techniques. Inter-modality methods aim to capture the underlying relationships among input modalities but may miss the connections within each modality. Conversely, intra-modality techniques focus on uncovering modality-specific relations but tend to disregard the inter-modal connections across input modalities. Combining modality-specific features with inter-modal connections can synergize and enhance overall analysis. Hence, exploring a uni-

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 67–77
November 29 - December 1, 2023 ©2023 Association for Computational Linguistics

fied framework capable of capturing both inter-modality and intra-modality relations within the input data is imperative.

Motivated by these challenges, we present a novel Multi-Modal Cross-Attention Network for Stock Market Prediction (MCASP). MCASP forecasts the direction of price movements by jointly modeling inter-modality and intra-modality relationships within the input data (i.e., financial news, market data, and technical indicators) within a unified deep learning framework. To achieve this, we construct two distinct attention networks: a self-attention network and a cross-attention network, designed to capture intra-modal and inter-modal relationships, respectively.

The self-attention module focuses on extracting modality-specific features from the input modalities. We first employ two separate Long Short-Term Memory (LSTM) networks to extract latent features from market data and technical indicators. Simultaneously, we leverage FinBERT (Liu et al., 2020b) to encode textual data (i.e., financial news). Within the self-attention network, the LSTM network outputs are processed by two Transformer (Vaswani et al., 2017) units, while the encoded textual data undergo analysis via a Convolutional Neural Network (CNN).

The cross-attention module involves creating three pairs by concatenating representations of news and market data, news and technical indicators, and market data and technical indicators. These pairs are then fed into three separate Transformer units. The outputs from the self-attention and cross-attention modules converge in the Fusion Layer to generate a combined feature vector. Finally, we employ a fully-connected layer to predict the direction of price movements.

2 Related Work

In this section, we review related work in stock market prediction, multimodal machine learning and the attention mechanism.

2.1 Stock Market Prediction

Financial news, market data, social media data, official company announcements have been widely used for market analysis research. It has been shown by Shi et al. (2019a) that using only news titles is better than using the whole article text. Schumaker et al. (2012) proposed the Arizona Financial Text (AZFinText) system, focusing on sentiment analysis using propoer nouns. In another study, Vargas et al. (2017) represented news headlines using Word2Vec word embeddings and constructed a multimodal prediction model using Convolutional Neural Networks (CNN) and Long Short-term Memory (LSTM) networks. Meanwhile, Huynh et al. (2017) designed a prediction model using the Bidirectional Gated Recurrent Unit (BGRU) architecture, extracting news headlines and representing them using word embedding vectors.

The paper by Nuij et al. (2014) used ViewerPro to extract events from news articles and incorporated them with technical indicators. Matsubara et al. (2018) employed paragraph vectors for news data representation, and Ding et al. (2015) introduced a CNN-based event embeddings model where the authors constructed a neural tensor network to learn event embeddings from financial news data.

2.2 Multimodal Machine Learning

Multimodal learning architectures have been widely utilized in various fields including robotics (Lee et al., 2018), healthcare (Ghulam et al., 2021), multimedia (Liang et al., 2018), and sentiment analysis (Zadeh et al., 2018). A multimodal paper by Barnum et al. (2020) applies early fusion in the multimodal representation of audio and visual inputs and another research (Federici et al., 2020) employs structured image and textual to construct multimodal concept taxonomies. Researchers have also utilized various RNN structures for multimodal representations for different kinds of applications such as human behaviour analysis (Rajagopalan et al., 2016) and time-series data analysis (Liang et al., 2018; Zadeh et al., 2018).

One popular technique for combined utilization of multimodal data is early fusion (Morency et al., 2011; Pérez-Rosas et al., 2013). Early fusion concatenates low-level features from individual modalities to be utilized with any learning framework for downstream machine learning tasks. Moreover, early fusion performs poorly when feature fusion among non-interacting modalities (such as voice and fingerprint) is performed. These limitations are slightly addressed in Zadeh et al. (2016), where shared embeddings (latent space) among individual modalities are learned. These shared representations outperform the early fusion but require careful parameter tuning.

There also exists a stream of work that perform outer-product-based neural frameworks for multi-modal data fusion. In Lin et al. (2015) a bilinear-CNN is proposed to obtain bi-modal interactions among features obtained from two heterogeneous CNNs. This is accomplished by taking a neural-based bilinear product of high-level features. The bilinear layer required parameter estimation of a quadratic number of neurons and hence prone to over-fitting. This limitation is alleviated in Fukui et al. (2016); Hu et al. (2017a) which introduced an alternate formulation of the bilinear layer and obtains its compact representation by utilizing sophisticated neural-based factorization schemes.

2.3 Attention

The attention mechanism has found success in a wide range of domains, including natural language processing (NLP) (Bahdanau et al., 2014; Vaswani et al., 2017), image captioning (You et al., 2016), image classification (Xiao et al., 2014), visual question answering (Lu et al., 2016), and more (Rush et al., 2015; Li et al., 2015). Notably, the Transformer model (Vaswani et al., 2017) introduced the self-attention mechanism, which explores intra-modal relationships, such as the relationships between words in machine translation.

Taking inspiration from the Transformer model (Vaswani et al., 2017), the self-attention mechanism has been applied in various works, extending its utility to visual question answering (Yu et al., 2019), video analysis (Wang et al., 2017), and image-text matching (Wu et al., 2019).

In recent years, attention mechanisms have also made their way into multi-modal learning problems. While architectures like BERT (Devlin et al., 2019) were originally designed for NLP tasks, they have been adapted for multi-modal challenges as well (Chen et al., 2019; Lu et al., 2019). For instance, some approaches, like the dual attention network in Nam et al. (2016), focus on learning inter-modal relationships between visual regions and textual elements within sentences. Others, like the co-attention framework in Lu et al. (2016), tackle tasks like visual question answering by jointly learning image and question attentions. Additionally, in Paulus et al. (2017), a combination of inter-modal and intra-modal attentions is leveraged within deep reinforcement learning for text summarization.

3 Model Design

In this section, we provide a detailed description of the architecture of the proposed MCASP model. The design of our MCASP model is demonstrated in Figure1.

3.1 Input Representation

We start by using historical market data and financial news as our primary data sources. From the market data, we derive a set of seven technical indicators. We employ three distinct data modalities for stock market prediction: market data, technical indicators, and financial news. To process these modalities, we employ three separate deep networks.

We construct two LSTM networks to handle the market data and technical indicator modalities, respectively. Additionally, we utilize text embeddings to encode the news data. For this purpose, we leverage BERT and FinBERT embeddings.

The latent features obtained from the LSTM networks and the sentence embeddings from FinBERT are then fed into the self-attention and cross-attention modules to capture both intra-modal and inter-modal relationships.

3.2 Self-Attention Module

The primary objective of the attention process is to discern the relationship between two states and focus on the most crucial features. This is achieved by assigning higher weights to the most pertinent elements within the input vectors. The attention layer consists of three key components: the query, keys, and values, with these elements being identical in the self-attention context. The attention mechanism can be conceptualized as mapping a query and a set of key-value pairs to an output, where the output is a weighted sum of the values. The weight matrix, determining the weight assigned to each value, is defined using the query and the key. Several options for the attention function are available, including the dot product, multi-layer perceptron, and scaled dot product.

The self-attention network is used to capture intra-modality relations, employing two separate Transformer units (Li et al., 2014) for market data and technical indicators, along with a CNN for financial news data. In the Transformer model, we employ the scaled dot product to compute the weight matrix. This module encompasses both multi-head self-attention and position-wise feed-

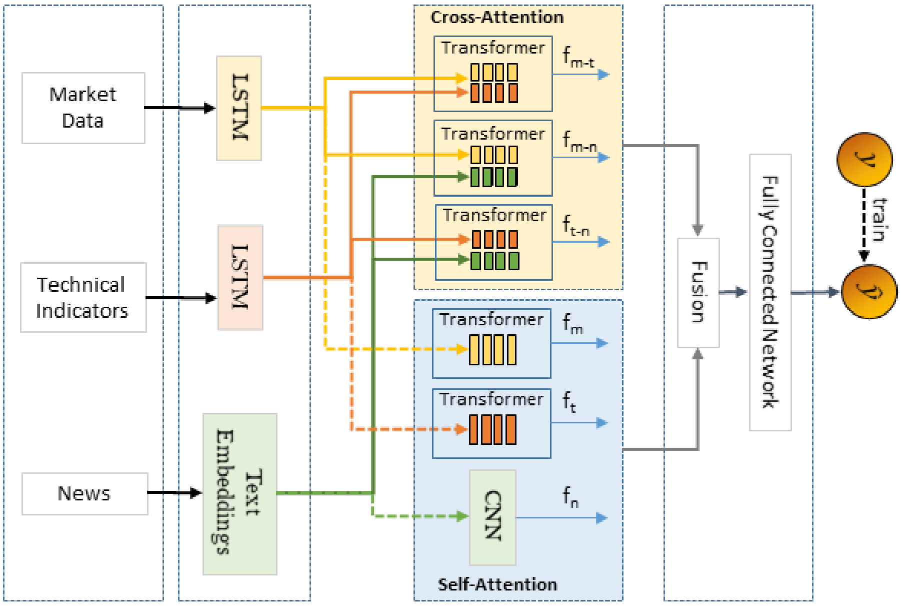

Figure 1: Demonstration of the MCASP architecture design.

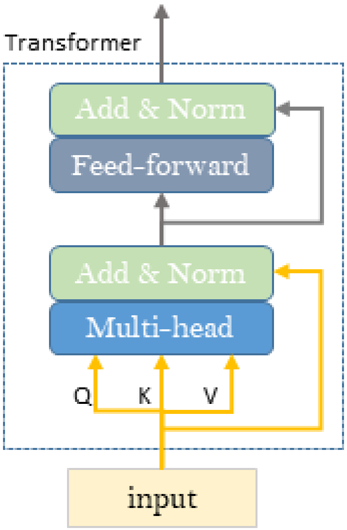

Figure 2: Design of the Transformer model

forward layers, as depicted in Figure 2. The term 'multi-head attention' implies that attention is computed multiple times. The attention calculation is as follows:

$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (1)$$

Where the d_k represents the dimension of the queries and the keys. In the Transformer module, multiple parallel attention values are computed where each output is called a head. The i^th head is calculated as:

$$head_i = A(QW_i^Q, KW_i^K, VW_i^V) \qquad (2)$$

We then concatenate these heads to obtain the multi-head attention.

$$MT(Q, K, V) = Concat(head_1, ., head_h)W^0 \qquad (3)$$

In our self-attention module, the two Transformers for market data and technical indicators modalities, we get the following two outputs:

$$\begin{aligned} f_m &= MT(Q_m, K_m, V_m) \\ f_t &= MT(Q_t, K_t, V_t) \end{aligned} \qquad (4)$$

For the textual data modality, we utilize the outputs of the BERT embeddings. The BERT model incorporates multiple Transformers and is proficient at capturing intra-modality information. Subsequently, we employ a CNN to extract local latent features denoted as f_n.

These three outputs from our self-attention module, namely f_m, f_t, and f_n, are later employed to predict the movement of closing prices.

3.3 Cross-Attention Module

We introduce a novel cross-attention to model both intra-modality information and the interconnectedness of the modalities, achieved by implementing three separate Transformer units. By modeling both intra-modality and inter-modality relationships, we aim to capture the joint effect of the input modalities while retaining modality-specific features.. Our aim is to capture the interactions across the input modalities by applying the cross-attention function to the outputs of the input representation layer. Initially, we establish three distinct pairs from the modalities to implement the attention mechanism: from market data to technical indicators $(m-t)$, from market data to financial news $(m-n)$, and from technical indicators to financial news $(t-n)$. Market data and the derived technical indicators have a significant influence on market movements, which justifies prioritizing these pairings with higher weights.

The calculation of these three cross-attention values is as follows:

$$\begin{aligned} A_{m-n}(Q_m, K_n, V_n) &= softmax(\frac{Q_m K_n^T}{\sqrt{d_k}})V_n \\ A_{m-t}(Q_m, K_t, V_t) &= softmax(\frac{Q_m K_t^T}{\sqrt{d_k}})V_t \\ A_{t-n}(Q_t, K_n, V_n) &= softmax(\frac{Q_t K_n^T}{\sqrt{d_k}})V_n \end{aligned} \tag{5}$$

Here, A_{m-n}, A_{m-t}, and A_{t-n} represent the cross-attention between market data and news, market data and technical indicators, and technical indicators and news modalities, respectively. Furthermore, Q_m and Q_t denote the query vectors for the market data and technical indicators modalities, while K_t and K_n represent the key vectors, and V_t and V_n denote the value vectors for the technical indicators and news modalities, respectively.

With these cross-attention terms in place, we proceed to compute the attention values for each head as follows:

$$\begin{aligned} head^i_{m-n} &= A_{m-n}(Q_m W_i^{Q_m}, K_n W_i^{K_n}, V_n W_i^{V_n}) \\ head^i_{m-t} &= A_{m-t}(Q_m W_i^{Q_m}, K_t W_i^{K_t}, V_t W_i^{V_t}) \\ head^i_{t-n} &= A_{t-n}(Q_t W_i^{Q_t}, K_n W_i^{K_n}, V_n W_i^{V_n}) \end{aligned} \tag{6}$$

These terms represent each head in each cross-attention pair. Subsequently, we combine these head values for each pair to obtain the multi-head attention for each cross-attention block:

$$\begin{aligned} MT_{m-n} &= Concat(head^1_{(m-n)}, ., head^h_{(m-n)})W^0_{m-n} \\ MT_{m-t} &= Concat(head^1_{(m-t)}, ., head^h_{(m-t)})W^0_{m-t} \\ MT_{t-n} &= Concat(head^1_{(t-n)}, ., head^h_{(t-n)})W^0_{t-n} \end{aligned} \tag{7}$$

Putting all these together, our cross-attention module produces the following three outputs:

$$\begin{aligned} f_{m-n} &= MT_{m-n} \\ f_{m-t} &= MT_{m-t} \\ f_{t-n} &= MT_{t-n} \end{aligned} \tag{8}$$

3.4 Fusion Layer

In the fusion layer, we amalgamate the feature vectors from the self-attention and cross-attention modules to form a combined feature vector.

$$f_{merged} = [f_m, f_t, f_n, f_{m-n}, f_{m-t}, f_{n-t}] \tag{9}$$

We then employ a fully connected layer with ReLU as the activation function to process the feature vector f_{merged}. In the final step, another fully connected layer is employed to make predictions. The overall network is a binary classification model used for predicting the movement direction of stock closing prices, and the model weights are optimized by minimizing the binary cross-entropy loss:

$$L = -(y \log(\hat{y}) + (1-y)\log(1-\hat{y})) \tag{10}$$

where y represents the target class for the movement direction, and $\hat{y}$ signifies the prediction obtained from MCASP. The movement direction is defined as the difference between the closing prices on day $t+1$ and day t. The labels are categorized into two classes: Class 1 indicating an upward movement and Class 0 indicating a downward movement in the closing prices.

4 Experimental Settings

In our experiments, we utilized real-world datasets encompassing financial news, market data, and technical indicators spanning from January 1, 2010, to December 31, 2019, encompassing a 10-year period. The financial news was sourced from Reuters[1], with each article containing a title, body, and publication date. The publication date was employed to align the articles with the daily market data. We specifically focused on the headlines from the financial news, as research has demonstrated that using news titles can yield superior prediction results compared to using the entire article body (Shi et al., 2019b). The number of news titles per trading day varied; hence, we aggregated all the titles for a given day into a single extended sentence

[1] https://www.reuters.com/business/finance/

and employed FinBERT to encode the textual data into feature vectors. Consequently, we obtained a single sentence embedding vector for each trading day.

We utilize historical market data for S&P index and individual stocks from Yahoo Finance[2] for the corresponding dates. These five companies included Google, Tesla, Amazon, Apple, and Microsoft and the data includes Open, High, Low, Close prices, and Volume. We normalize the market data to be within the range of [0, 1].

We initially employ an 80-20% split for training and testing for index price prediction. We also evaluate the yearly performances of the models by utilizing the first 10 months of each year for training and the last 2 months for testing. We utilize the 80-20% split again for training and testing purposes for individual stock prediction.

Based on the literature (Kim, 2003), we computed seven technical indicators for each trading day using the market data over the preceding five days.

We employ accuracy (Acc) and Matthews Correlation Coefficent (MCC) to evaluate the performance of different models. MCC is generalyy employed when the sizes of classes y = 1 and y = 0 differ.

4.1 Baseline Methods

We compare our approach with the following baselines on predicting individual stocks and S&P500 index.

Recurrent Convolutional Neural Network (RCNN) (Vargas et al., 2017) is a CNN and RNN based stcok forecast model that utilizes technical indicators and financial news. **Event Embeddings (EB-RCN)** (Oncharoen and Vateekul, 2018) is another LSTM and CNN based model that also includes market data and employ event embeddings from (Ding et al., 2015). **Bidirectional Gated Recurrent Unit (BGRU)** (Huynh et al., 2017) uses both online financial news and historical price data to predict the stock movements. **LSTM-based Recurrent State Transition (ANRES)** (Liu et al., 2020a) uses only news events for market movement prediction. **Hybrid Attention Network (HAN)** (Hu et al., 2017b) is a state-of-the-art stock trend prediction model with hierarchical attention that utilizes news data. **Multi-Modality Attention Network** (MMAN) (He and Gu, 2021) **Attention-Based Recurrent Neural Network (At-LSTM)** (Liu, 2018) **Adversarial Attentive LSTM (Adv-LSTM)** (Feng et al., 2018) is a market prediction model using historical market data, where the authors employ attentive LSTMs and utilize adversarial training strategy.

Other than these methods, we also perform ablation studies by constructing different variants of the proposed MCASP model.

5 Results and Analysis

In order to test the effectiveness of our model, we run experiments using real-world dataset including financial news data, historical market data and technical indicators.

5.1 Main Results

We use our dataset to conduct tests for forecasting of the price movements of S&P500 index and five individual stocks. The accuracy results are illustrated in Figure 3, showing that MCASP improves upon the baseline models. The MCC results, presented in Figure 4, echo the same trend, with MCASP exhibiting superior prediction performance for the price movement directions of all five stocks and S&P index compared to the baseline models.

Overall, in our experiments, MCASP consistently achieves the best results in terms of both accuracy and MCC. When compared to the baselines, MCASP demonstrates improvements in prediction performance for both index and individual stock predictions, underscoring the effectiveness of the proposed multi-modal attention design in leveraging intra-modal and inter-modal information from multiple input sources.

Among the baseline models, attention-based prediction models perform better than other baselines in both accuracy and MCC. These results underscore the significance of the attention module in capturing critical latent features from the input data. However, MCASP surpasses the attention-based baseline models, suggesting that its enhanced performance stems not only from the use of the self-attention module but also from its ability to extract inter-modal relationships among input modalities through the novel cross-attention module.

We also asses the models' yearly prediction performances for S&P 500 index prediction, where we use the first 10 months of each year for training and the last two months for testing. The accuracy re-

[2] https://finance.yahoo.com/

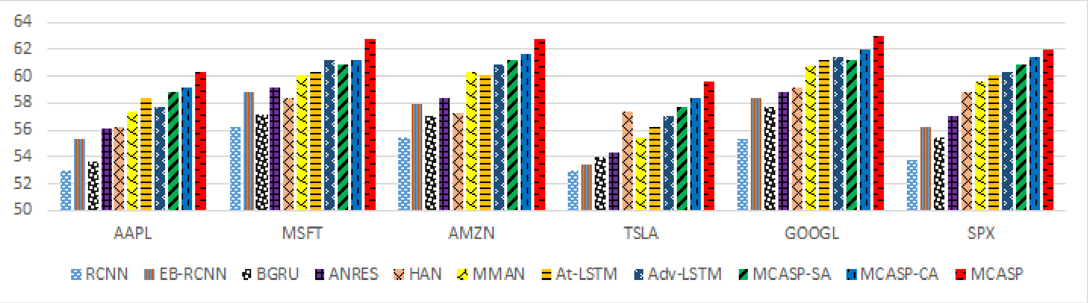

Figure 3: Accuracy results on index and individual stock prediction (the higher, the better).

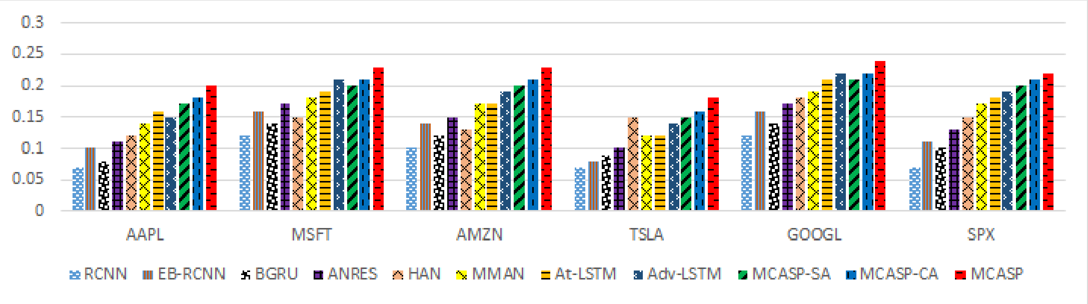

Figure 4: MCC results on index and individual stock prediction (the higher, the better).

sults, given in Figure 5, demonstrate that MCASP consistently outperforms all the baseline models for each year. Although the yearly results are slightly lower than the initial test results, this can be attributed to the smaller test sample size inherent in the yearly setup.

Collectively, the experiments involving S&P500 index prediction and the prediction of price movements for five individual stocks demonstrate that the MCASP model is adept at learning meaningful representations from multiple input modalities, capitalizing on the self-attention network and the innovative cross-attention module.

5.2 Ablation Study

To assess the impact of different components of the MCASP model, we conducted an ablation study using the same real-world dataset. Initially, we evaluated the effectiveness of our two attention modules independently by creating two distinct models. Subsequently, we explored three text embedding techniques to demonstrate the influence of the textual representation method on the overall performance.

Self-attention and cross-attention modules. This experimental study elucidates the individual performance of each module and underscores the significance of capturing both intra-model and inter-model information, in contrast to the prevalent approach of focusing solely on either modality-specific or joint influence of input modalities, as seen in most existing works. To this end, we developed two distinct models - MCASP-SA (MCASP with the self-attention module only) and MCASP-CA (MCASP with the cross-attention module only) - and subjected them to testing using our original dataset.

In our experiments, MCASP consistently outperforms both MCASP-SA (which exclusively employs the self-attention module) and MCASP-CA (which relies solely on the cross-attention module) across both accuracy and MCC metrics. This substantiates the effectiveness of our proposed design in addressing multi-modal problems.

Notably, MCASP-CA yields superior results compared to MCASP-SA. We postulate that this is attributed to the cross-attention module's design, which initially extracts modality-specific features and subsequently captures inter-modal relationships among modalities using the attention mechanism.

Moreover, when compared to the baseline models, both MCASP-SA and MCASP-CA consistently demonstrate improved accuracy and MCC results in the majority of the tests. This underscores the success of the proposed sequential design for both modules. The results further affirm that leveraging multiple modalities (i.e., financial news, his-

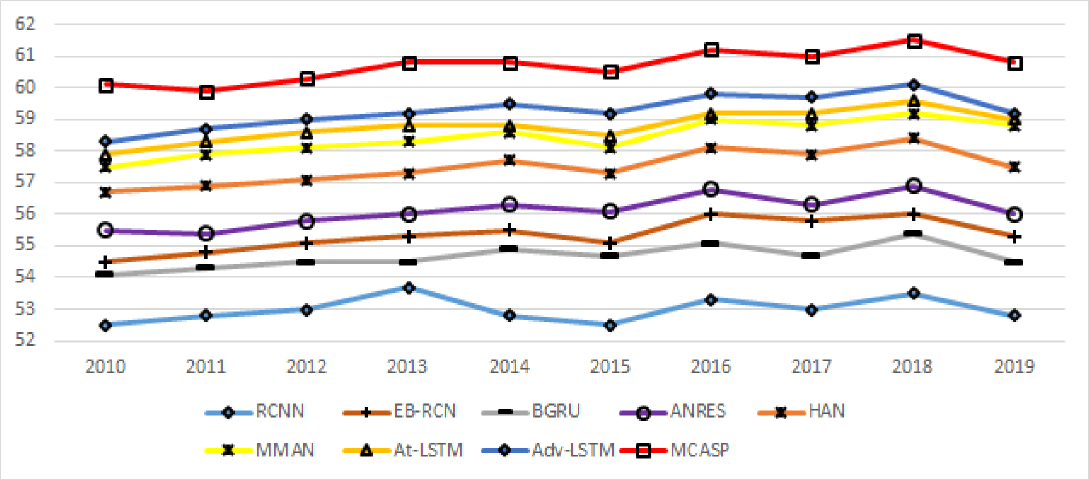

Figure 5: Yearly ACC results on S&P index prediction (the higher, the better).

torical market data, and technical indicators) can enhance model performance.

MCASP with various text embeddings. We subsequently examined the impact of various textual embeddings (Transformer-based BERT and GloVe) on the overall model performance. We employed three distinct textual embedding methods to encode and represent the financial news data, namely GloVe word embeddings, Transformer-based BERT embeddings, and FinBERT embeddings. Our experimental results underscore the significance of selecting an appropriate text embedding method when utilizing financial news data.

The results, presented in Table1 show that Transformer-based BERT and FinBERT embeddings consistently outperformed GloVe embeddings across both accuracy and MCC metrics for S&P index prediction. Furthermore, FinBERT showed improved results compared to BERT embeddings, underscoring the value of domain-specific knowledge in textual data representation.

Table 1: The impact of different text embedding methods.

Embedding Method	***Accuracy***	***MCC***
GloVe	60.91%	0.208
BERT	61.60%	0.215
FinBERT	62.03%	0.228

Notably, predictions using FinBERT as our text embedding method exhibited improvement compared to GloVe and BERT embeddings. This highlights the utility of domain knowledge in comprehending and representing textual data. However, even without domain knowledge and when employing RNN-based GloVe embeddings and general BERT embeddings, MCASP consistently outperformed all baseline methods across both metrics for S&P500 index prediction. These results affirm that while a robust textual representation technique can enhance model performance, the primary factor contributing to improved results lies in the novel multi-modal design, which incorporates both self-attention and cross-attention modules to capture latent features from the input modalities.

6 Conclusion

We have proposed a novel multi-modal cross attention network for stock market prediction that models the intra-modal and inter-modal information from the input modalities in a unified framework. We first analyze the input modalities via three separate deep networks to extract the salient features. We then process these features with the proposed self-attention and cross-attention modules to jointly model the intra-modal and inter-modal information. We analyze financial news, historical market data and technical indicators to predict the movement direction of S&P500 index prices and the prices of five individual stocks. We test the effectiveness of the proposed multi-modal design using real-world dataset from Reuters and Yahoo! Finance and compare its performance against multiple state-of-the-art baseline models. Experimental results show that our model achieves improved performance in stock market prediction.

References

R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara. 2016. Deep learning for stock prediction using numerical and textual information. *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–6.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

George M. Barnum, Sabera Talukder, and Yisong Yue. 2020. On the benefits of early fusion in multimodal representation learning. *ArXiv*, abs/2011.07191.

W. Chen, C. Yeo, C. Lau, and B. S. Lee. 2018. Leveraging social media news to predict stock index movement using rnn-boost. *Data Knowl. Eng.*, 118:14–24.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. *ArXiv*, abs/1909.11740.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

X. Ding, Y. Zhang, T. Liu, and J. Duan. 2015. Deep learning for event-driven stock prediction. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*.

Marco Federici, Anjan Dutta, Patrick Forr'e, Nate Kushman, and Zeynep Akata. 2020. Learning robust representations via multi-view information bottleneck. *ArXiv*, abs/2002.07017.

Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2018. Enhancing stock movement prediction with adversarial training. In *International Joint Conference on Artificial Intelligence*.

S. Feuerriegel and J. Gordon. 2018. Long-term stock index forecasting based on text mining of regulatory disclosures. *Decis. Support Syst.*, 112:88–97.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468.

Muhammad Ghulam, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, Mansour Alsulaiman, and Tiago H. Falk. 2021. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Inf. Fusion*, 76:355–375.

Shwai He and Shihao Gu. 2021. Multi-modal attention network for stock movements prediction. *ArXiv*, abs/2112.13593.

Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, Zheng Lu, Sankha S Mukherjee, Timothy M Hospedales, Neil Martin Robertson, and Yongxin Yang. 2017a. Attribute-enhanced face recognition with neural tensor fusion networks. In *ICCV*, pages 3764–3773.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2017b. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*.

Huy D. Huynh, L. Minh Dang, and Duc Duong. 2017. A new model for stock price movements prediction using deep neural network. *Proceedings of the Eighth International Symposium on Information and Communication Technology*.

Kyoungjae Kim. 2003. Financial time series forecasting using support vector machines. *Neurocomputing*, 55:307–319.

Michelle A. Lee, Yuke Zhu, Krishna Parasuram Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. 2018. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Annual Meeting of the Association for Computational Linguistics*.

Q. Li, J. Tan, J. Wang, and H. Chen. 2020. A multimodal event-driven lstm model for stock prediction using online news. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

X. Li, H. Xie, L. Chen, J. Wang, and X. Deng. 2014. News impact on stock price return via sentiment analysis. *Knowl. Based Syst.*, 69:14–23.

Paul Pu Liang, Liu Ziyin, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *Conference on Empirical Methods in Natural Language Processing*.

Tsung-Yu Lin, Aruni Roy-Chowdhury, and Subhransu Maji. 2015. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457.

Huicheng Liu. 2018. Leveraging financial news for stock trend prediction with attention-based recurrent neural network. *ArXiv*, abs/1811.06173.

Xiao Liu, Heyan Huang, Yue Zhang, and Changsen Yuan. 2020a. News-driven stock prediction with attention-based noisy recurrent state transition. *ArXiv*, abs/2004.01878.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020b. Finbert: A pre-trained financial language representation model for financial text mining. In *International Joint Conference on Artificial Intelligence*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *ArXiv*, abs/1606.00061.

T. Matsubara, R. Akita, and K. Uehara. 2018. Stock price prediction by deep neural generative model of news articles. *IEICE Trans. Inf. Syst.*, 101-D:901–908.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2016. Dual attention networks for multimodal reasoning and matching. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2156–2164.

W. Nuij, V. Milea, F. Hogenboom, F. Frasincar, and U. Kaymak. 2014. An automated framework for incorporating news into stock trading strategies. *IEEE Transactions on Knowledge and Data Engineering*, 26:823–835.

Pisut Oncharoen and Peerapon Vateekul. 2018. Deep learning for stock market prediction using event embedding and technical indicators. *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pages 19–24.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *ArXiv*, abs/1705.04304.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltruaitis, and Roland Göcke. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *ArXiv*, abs/1509.00685.

R. P. Schumaker and H. Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27:12:1–12:19.

R. P. Schumaker, Y. Zhang, C. Huang, and H. Chen. 2012. Evaluating sentiment in financial news articles. *Decis. Support Syst.*, 53:458–464.

L. Shi, Z. Teng, L. Wang, Y. Zhang, and Alexander Binder. 2019a. Deepclue: Visual interpretation of text-based deep stock prediction. *IEEE Transactions on Knowledge and Data Engineering*, 31:1094–1108.

Lei Shi, Zhiyang Teng, Le Wang, Yue Zhang, and Alexander Binder. 2019b. Deepclue: Visual interpretation of text-based deep stock prediction. *IEEE Transactions on Knowledge and Data Engineering*, 31:1094–1108.

M. R. Vargas, B. S. L. P. De Lima, and A. Evsukoff. 2017. Deep learning for stock market prediction from financial news articles. *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pages 60–65.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

X. Wang, Ross B. Girshick, Abhinav Kumar Gupta, and Kaiming He. 2017. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803.

Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning fragment self-attention embeddings for image-text matching. *Proceedings of the 27th ACM International Conference on Multimedia*.

Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. 2014. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6274–6283.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, E. Cambria, and Louis-Philippe

Morency. 2018. Memory fusion network for multi-view sequential learning. In *AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.