Causality-Guided Multi-Memory Interaction Network for Multivariate Stock Price Movement Prediction

Di Luo[1]*, **Weiheng Liao**[1]*, **Shuqi Li**[1]*, **Xin Cheng**[2] and **Rui Yan**[1,3]†
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Wangxuan Institute of Computer Technology, Peking University
[3]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education
{di_luo,shuqili,ruiyan}@ruc.edu.cn,
liaoweiheng@gmail.com, chengxin1998@stu.pku.edu.cn

Abstract

Over the past few years, we've witnessed an enormous interest in stock price movement prediction using AI techniques. In recent literature, auxiliary data has been used to improve prediction accuracy, such as textual news. When predicting a particular stock, we assume that information from other stocks should also be utilized as auxiliary data to enhance performance. In this paper, we propose the Causality-guided Multi-memory Interaction Network (CMIN), a novel end-to-end deep neural network for stock movement prediction which, for the first time, models the multi-modality between financial text data and causality-enhanced stock correlations to achieve higher prediction accuracy. CMIN transforms the basic attention mechanism into Causal Attention by calculating transfer entropy between multivariate stocks in order to avoid attention on spurious correlations. Furthermore, we introduce a fusion mechanism to model the multi-directional interactions through which CMIN learns not only the self-influence but also the interactive influence in information flows representing the interrelationship between text and stock correlations. The effectiveness of the proposed approach is demonstrated by experiments on three real-world datasets collected from the U.S. and Chinese markets, where CMIN outperforms existing models to establish a new state-of-the-art prediction accuracy.

1 Introduction

Financial services, known for their competitiveness, have always been at the forefront of adopting data science techniques to drive investment decisions. Quantitative trading, a specific field within it, has drawn immense interest from both academia and industry over the last few decades. With the rapid advancements in deep learning recently, computer scientists and quantitative researchers have joined forces to apply AI techniques to tackle the challenges within this domain.

Among various tasks, one of the most prominent is stock price movement prediction (Bhardwaj, 2021). The reason for its popularity is self-evident: once a model is able to predict future movement with considerable accuracy, numerous trading strategies can be easily built around it.

Recent studies have shown that deep neural networks are ideal candidates for such prediction models (Yoo et al., 2021; Gunduz, 2021). Supporters of the efficient-market hypothesis (EMH), which posits that asset prices reflect all available information, tackle the task with price information alone (Zhang et al., 2017; Stoean et al., 2019; Sezer and Özbayoglu, 2020). However, an alternative perspective suggests that additional insights can be gained from analyzing news articles and social media posts, which may hold valuable clues about the future (Hu et al., 2018; Xu and Cohen, 2018; Wang et al., 2019b; Tang et al., 2020).

Another intriguing approach analyzes the relationships between different stocks. Clearly positive and negative correlations, or even non-correlations can be immensely useful in constructing a diversified stock portfolio (Borodin et al., 2003). Several studies even empirically demonstrate that exploiting correlations can improve the accuracy of stock price movement prediction (Long et al., 2020; Yoo et al., 2021). However, their correlations are often realized by acquiring industry sector and calculating correlation matrices or attention scores, which are bidirectional and symmetrical, leading to excessive attention on spurious correlations. Due to the lag problem widely existed between two time series, we are more concerned about the dominance of information flow between stocks, specifically, the direction of causality.

Additionally, we have observed that the situation can significantly change when incorporating text information. Let's consider two highly correlated

*Equal contribution.
†Corresponding author: Rui Yan (ruiyan@ruc.edu.cn).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 12164–12176
July 9-14, 2023

companies (A and B) and there is promising news specifically about company A. In such a scenario, it's fairly easy to infer that the current news might still have a substantial impact on company B, despite there being no direct connection between the two companies on paper. However, it's impossible to reach this conclusion by just examining the news about company A or the correlation between A and B alone, which highlights the limitations of relying solely on individual pieces of textual information or traditional correlations between stocks.

Inspired by observations above, we propose the Causality-guided Multi-memory Interaction Network (CMIN), a novel end-to-end deep neural network which captures both financial news as well as the causality-enhanced correlations between stocks for better stock price movement prediction.

To achieve this goal, CMIN incorporates two key components: the *Text Memory Network* and the *Stock Correlation Memory Network*. Both networks utilize a recurrent neural network with nonlinear combination of memory attentions to generate a global memory abstraction. And we introduce a global causality matrix according to the transfer entropy between stock price time series to guide the abstraction process, forming a Causal Attention mechanism to capture the asymmetric correlations. By considering causality, CMIN goes beyond traditional symmetric correlations and captures the true inter-dependencies between stocks. Furthermore, we employ an attention-based fusion mechanism between the two networks, introducing multi-directional interactions through which CMIN learns not only the self-influence within each network but also the interactive influence between them. It captures the interrelationship between textual information and correlations, enhancing the overall predictive power of CMIN.

We further demonstrate the effectiveness of CMIN with experiments conducted on 3 real-world datasets collected from both the U.S. and Chinese markets, where CMIN achieves state-of-the-art prediction accuracy, surpassing existing models in terms of performance.

To summarize, our main contributions are:

- Proposal of a causality-guided multi-memory interaction network for stock movement prediction which is to our best knowledge the first attempt to simultaneously consider causality-enhanced correlations and textual information to achieve higher prediction accuracy;
- Introduction of the attention-based multi-directional interactions, so that CMIN captures not only the self-influence of temporal movements and textual information but also the interactions between these two types of information flows;
- Collection and release of two new datasets: one for the U.S. market and another for the Chinese market. Both datasets include comprehensive financial texts and stock price time series data, which are publicly available at `https://github.com/BigRoddy/CMIN-Dataset`, facilitating further research and benchmarking in the field.

2 Related Work

2.1 Stock Movement Prediction

In traditional trading practices, two main frameworks are commonly used to make predictions on future stock prices (Ferreira et al., 2021). The first is fundamental analysis, which aims to assess the intrinsic value of a stock by considering various factors related to it as a whole, such as financial statements, industry trends and economic conditions. The other is technical analysis, which operates under the assumption that the market is efficient (i.e., the Efficient Market Hypothesis holds true) and focuses on analyzing only historical and current price patterns in order to predict future movements.

Although both frameworks have been widely adopted by top hedge funds and investment firms, technical analysis has gained more popularity among AI practitioners, many of whom focus on employing long short-term memory networks and other innovative architectures to model stock price history alongside technical analysis indicators (Nelson et al., 2017; Zhang et al., 2017; Stoean et al., 2019; Sezer and Özbayoglu, 2020). This is primarily because processing a single stream of price data is relatively simpler than analyzing and synthesizing a range of diverse data sources with varying frequencies and characteristics.

2.2 Predicting with the Help of Text Data

The recent advancement of natural language processing (NLP) techniques has opened up new possibilities for analyzing large volumes of text data in the context of stock movement prediction. Many researchers have recognized the potential value of incorporating news articles, analysis, commentaries and even social media posts (Xu and Cohen, 2018), which are believed to provide valuable insights about the future. Some studies focus solely on textual information. For example, (Hu et al., 2018) leverages attention mechanism at multiple levels within a deep structure to identify the most important news articles and predict price trends. Others adopt a two-step approach. First, they extract features (e.g. investor sentiment) from financial texts. Then they fuse these features with price information to make predictions such as (Li et al., 2017) and (Jin et al., 2020). This integration of text analysis with quantitative techniques holds promise for enhancing the accuracy and effectiveness of stock movement prediction models.

2.3 Exploiting the Relations between Stocks

Another important trading framework takes advantage of the correlations between different stocks. Portfolio selection, particularly pairs trading, is a well-known and successful trading strategy that exploits the correlated nature of stocks, whether positive or negative. In fact, as early as (Borodin et al., 2003) pointed out that stock correlations based portfolio selection could beat any strategy that relied on predicting trends or specific targets.

The incorporation of correlations in stock movement prediction has gained attention in recent years, drawing inspiration from several existing works. For example, (Yoo et al., 2021) utilizes transformer to learn dynamic correlations between stocks in an end-to-end manner. (Long et al., 2020) employs knowledge graphs and graph embedding techniques to model the relationships between stocks. These studies have achieved admirable results, potentially due to effective feature engineering however, because the direct benefit of stock correlations in predicting future prices lacks fundamental logic.

In this paper, we propose constructing a single model to handle both textual data and stock correlations simultaneously, aiming to shed light on the success of correlation-based approaches with the help of financial texts. We also introduce a novel causal attention mechanism to interpret the underlying logic behind stock correlations, leveraging transfer entropy to provide insights. We further model the multi-directional interactions between texts and correlations so that we could uncover not only relevant texts for prediction through correlations, but also the hidden stock correlations through texts. By integrating text data and stock correlations within a unified model, we aim to provide a comprehensive understanding of the relationship between the two and discover valuable insights for stock movement prediction.

3 Problem Formulation

This paper is dedicated to predict the price movement of a target stock. To this end, we leverage both the correlations between stocks and textual information to make prediction.

Consider a target stock with numerical features denoted as $P_{\text{target}} \in \mathbb{R}^{k \times d}$, where k represents the number of time steps in the monitoring window and d represents the dimension of price features, such as the highest and the closing prices. The prices of n other relevant stocks are denoted as: $\mathcal{P} = \{P_1, P_2, \cdots, P_n\} \in \mathbb{R}^{n \times k \times d}$.

Besides, we have financial documents associated with the target stock, which are represented as $\mathcal{M} = \{M_1, M_2, \cdots, M_k\} \in \mathbb{R}^{k \times l \times w}$, where l denotes the number of documents in a time step and w is the maximum number of words in a document. In cases where a specific stock has fewer than l documents at a given time step, zero padding values are added to align the lengths. Similarly, if a document contains fewer than w words, zero padding is applied to ensure uniform length across all documents (Ang and Lim, 2022).

We formulate the task as a binary classification problem whose goal is to predict the movement of the target stock at the next time step, denoted as $\hat{y}_{\text{target}}$. Here, $\hat{y}_{\text{target}} = 1$ indicates a rise in the price while $\hat{y}_{\text{target}} = 0$ indicates a fall.

4 Proposed Method

4.1 Model Overview

Figure 1 presents an overview of the Causality-guided Multi-Memory Interaction Network (CMIN). It is consisted of three main modules: feature embedding module, multi-memory networks and multi-directional interaction module.

(1) The feature embedding module includes two encoders, one for embedding the textual information and another for embedding the price time series. Additionally, a global causality matrix is introduced to capture the asymmetric correlations using transfer entropy, which then guides the calculation of attention weights in the multi-memory networks.

(2) The multi-memory networks consist of the *Text Memory Network* and *Stock Correlation Memory Network*, which are designed to select and re-

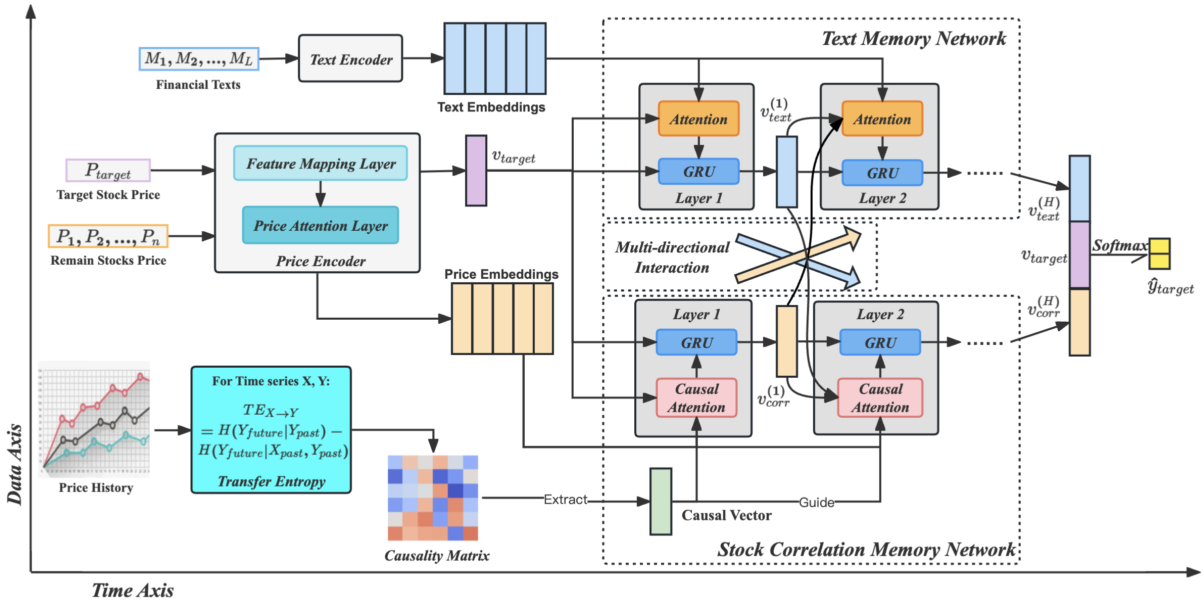

Figure 1: The structure of Causality-guided Multi-Memory Interaction Network (CMIN), which includes two encoders: *Text Encoder* and *Price Encoder*, a global causality matrix between stocks calculated by price history (changing as the monitoring window slides) and two memory networks: *Text Memory Network* and *Stock Correlation Memory Network* with multi-directional interactions between them.

tain the most relevant and influential information (textual and correlational) for the target stock.

(3) The multi-directional interaction module facilitates the interaction between the textual and correlational information. This interaction allows the two types of information to reinforce each other and leverage the advantages of different information flows for better prediction performance, enhancing the predictive capabilities of the CMIN.

4.2 Feature Embedding

Self-attention mechanisms have proven to be effective in capturing long-term dependencies and modeling complex sequential patterns, particularly in the Transformer architecture (Vaswani et al., 2017). Given the significance of historical information in financial documents and stock prices for stock price movement prediction, we employ attention mechanisms to summarize this information.

4.2.1 Text Encoder

The *Text Encoder* focuses on processing the financial documents $\mathcal{M}$ to extract meaningful representations for stock movement prediction. We firstly use a popular word representation tool Glove (Li et al., 2018) to generate the word embedding tensor $\mathcal{M}_{\text{word}} \in \mathbb{R}^{k \times l \times w \times d_w}$, where d_w is the size of word embeddings. Each word in the financial documents is represented as a d_w-dimensional vector.

Then the word embeddings are passed through a text embedding layer. Here we adopt the bidirectional Gated Recurrent Unit (Bi-GRU) (Li et al., 2022) to capture both preceding and succeeding contexts within each document. The average of the last hidden vectors is taken as the text embeddings $\mathcal{M}_{\text{text}} \in \mathbb{R}^{k \times l \times d_m}$, or equivalently $\mathcal{M}_{\text{text}} \in \mathbb{R}^{s \times d_m}$, where s is the total number of documents in the monitoring window.

After that, the text attention mechanism is applied to summarize all historical documents across time steps. The text embedding of the last time step $\mathcal{M}_{\text{text},-1} \in \mathbb{R}^{l \times d_m}$, serves as the query matrix, while the entire text embeddings $\mathcal{M}_{\text{text}} \in \mathbb{R}^{s \times d_m}$ acts as both the key and value matrices. Soft scaled dot-product attention is used to compute the attention weights, which are then applied to the text embedding to obtain a representation $E_{\text{text}} \in \mathbb{R}^{l \times d_m}$ enhanced by the history state attention:

$$E_{\text{text}} = \text{softmax}(\frac{\mathcal{M}_{\text{text},-1}\mathcal{M}_{\text{text}}^T}{\sqrt{d_m}})\mathcal{M}_{\text{text}}. \quad (1)$$

The resulting E_{text} is the textual embedding that contains highly concentrated information from the stock's related texts. This embedding serves as a summary of the historical text data and is used for further processing in the multi-memory networks and multi-directional interaction module of CMIN.

4.2.2 Price Encoder

The *Price Encoder* is introduced to utilize multivariate features from historical prices and capture their temporal interrelationships. Firstly we employ a feature mapping layer to project them into a latent space of dimension d_p, aiming to improve the learning capacity (Yoo et al., 2021). For target stock price $P_{\text{target}} \in \mathbb{R}^{k\times d}$, the historical price embeddings $\tilde{P}_{\text{target}} \in \mathbb{R}^{k\times d_p}$ can be formulated as:

$$\tilde{P}_{\text{target}} = ReLU(P_{\text{target}} W_t + b_t), \tag{2}$$

where $W_t \in \mathbb{R}^{d\times d_p}, b_t \in \mathbb{R}^{d_p}$ are parameters.

Moreover, recognizing that historical patterns can repeat themselves sometimes, we incorporate a multi-head price attention layer to capture each stock's distinctive changing patterns. The price embedding of the target stock at the last time step is donated as $\tilde{\mathcal{P}}^{-1}_{\text{target}} \in \mathbb{R}^{d_p}$. Then we employ the multi-head attention mechanism with the query $\tilde{\mathcal{P}}^{-1}_{\text{target}}$ and the key/value $\tilde{\mathcal{P}}_{\text{target}}$ as follows:

$$v_{\text{target}} = \text{MultiheadAtt}(\tilde{\mathcal{P}}_{\text{target}}, \tilde{\mathcal{P}}^{-1}_{\text{target}}) \tag{3}$$

v_{target} is a key vector that serves as the initial hidden state for the two memory networks, playing a crucial role in the final prediction. Similarly, we process the remaining stocks and obtain the correlational embedding $E_{\text{corr}} \in \mathbb{R}^{n\times d_p}$. Notably, the shared parameters across all stocks ensure the stability and generality of the extracted features (Wang et al., 2019a).

4.2.3 Causality Matrix

When it comes to detecting causal relationships and conducting predictive analysis, transfer entropy, a non-linear generalization of Granger causality (Seth, 2007), serves as a conceptually neat and mathematically rigorous method. It has been considered as an important tool for causality analysis and successfully applied in diverse domains including financial markets (Sandoval Junior et al., 2015).

Transfer entropy is derived from Shannon Entropy: $H = -\sum_{i=1}^{N} p_i \log p_i$. In this context, considering the time series of a stock, we can partition the possible values into different bins and calculate the probabilities at each time step. Transfer entropy from series X to another series Y can be defined as the average amount of information contained in the source X but not contained in Y's past:

$$\begin{aligned} TE_{X\to Y} =& H(Y_{future}|Y_{past}) \\ &- H(Y_{future}|X_{past}, Y_{past}) \end{aligned} \tag{4}$$

Based on this principle, for each monitoring window, we calculate the transfer entropy between all stocks using their historical closing prices and generate a transfer entropy matrix, referred to as the Causality Matrix $C \in \mathbb{R}^{n\times n}$, which illustrates the asymmetric flow of information from one stock to another. Specifically, $C[i, j]$ represents the transfer entropy from stock i to stock j, and $C[i, j] > C[j, i]$ indicates that stock i provides more predictive information about the movement of stock j than j to i. This Causality Matrix will next serve as a guide for the memory networks, enabling the identification of causal dependencies between multivariate stocks.

4.3 Multi-memory Networks

We introduce a *Text Memory Network* and a *Stock Correlation Memory Network* (Sukhbaatar et al., 2015) to manage the textual and correlational information separately. They each maintain a continuous representation and update it iteratively using multiple computational steps (hops), ultimately producing a global memory abstraction.

As shown in Figure 1, each layer of the memory network comprises an attention unit and a GRU unit, which receive textual or correlational embeddings as inputs and are supervised by the continuous representation generated in the previous layer. To initialize the continuous representations of each network, we use the target stock vector v_{target} (generated from Eq.3):

$$v^{(0)}_{\text{text}} = v^{(0)}_{\text{corr}} = v_{\text{target}}. \tag{5}$$

4.3.1 Text Memory Network

In each layer $h \in [1, H]$ of the *Text Memory Network*, we input the textual embeddings E_{text} (Eq.1) and the continuous representation from the previous layer $v^{(h-1)}_{\text{text}}$. We utilize an attention unit (Eq.3) to identify important information within the textual embeddings. Subsequently, a non-linear GRU cell unit (Xu et al., 2019) acts as an information aggregator, determining the amount of text information to retain:

$$v^{\text{Att}(h)}_{\text{text}} = \text{MultiheadAtt}(E_{\text{text}}, v^{(h-1)}_{\text{text}}), \tag{6}$$

where $v^{(h-1)}_{\text{text}}$ is the query matrix and E_{text} represents the raw form of the key and value matrices.

Then the GRU cell unit updates the current hidden state into the next hidden state and outputs it to the next layer as the new continuous representation:

$$v^{(h)}_{\text{text}} = GRU(v^{\text{Att}(h)}_{\text{text}}, v^{(h-1)}_{\text{text}}). \tag{7}$$

4.3.2 Stock Correlation Memory Network

The *Stock Correlation Memory Network* is employed to dynamically identify stock relationships and update the continuous representation of stock correlations in an intuitive and asymmetric manner.

However, the use of unsupervised attention weights in previous models can be problematic as they may be inevitably misled by the dataset bias, resulting in excessive attention on spurious stock correlations. To address this, we introduce extra knowledge in the form of Transfer Entropy-based causality to guide the attention weights and mitigate potential confounding effects.

For each target stock, we extract a causal vector $v_{\text{causal}} = C[:, \text{target}]$ from the pre-calculated causality matrix, which quantifies the degree of information flow from other stocks to it. Then we modify the traditional attention mechanism into Causal Attention by incorporating causal guidance:

$$\mathcal{S} = \text{softmax}(\frac{\mathcal{QK}^T}{\sqrt{d}}), \quad \tilde{\mathcal{S}} = f(\mathcal{S}, v_{\text{causal}}). \quad (8)$$

Here, f is a function that aggregates the attention weight $\mathcal{S}$ and the causal vector v_{causal} to produce a causality-guided attention weight $\tilde{\mathcal{S}}$. We use the average aggregation method for simplicity (i.e., $f(\mathcal{S}, v_{\text{causal}}) = (\mathcal{S} + v_{\text{causal}})/2$). To better balance them, one can introduce a hyperparameter $\lambda \in [0, 1]$. Then f() updates to $f(\mathcal{S}, v_{\text{causal}}) = \lambda\mathcal{S} + (1-\lambda)v_{\text{causal}}$. We believe that different degrees of causal attention can impact the model's performance, and leave it for future exploration.

The continuous representation is gradually updated through the Causal Attention, indicating the influence of causal relationships on movement prediction and the self-influence on the flow of correlation information:

$$v_{\text{corr}}^{\text{Att}(h)} = \text{CausalAtt}(E_{\text{corr}}, v_{\text{corr}}^{(h-1)}) \quad (9)$$

$$v_{\text{corr}}^{(h)} = GRU(v_{\text{corr}}^{\text{Att}(h)}, v_{\text{corr}}^{(h-1)}) \quad (10)$$

It is important to note that although we design multiple layers within each memory network to learn deep representations, different layers of the same memory network share the same unit. This enables the network to focus on crucial information that affects the movement of the target stock, thereby enhancing the continuous representation.

4.4 Multi-directional Interactions

In reality, textual information and correlations have an impact on each other when it comes to stock price movement prediction. For instance, news about a technological breakthrough in the new energy sector may uplift the prices of most stocks in that industry, thereby affecting the correlations among those stocks.

To simulate this phenomenon and enhance the synergy between textual and correlational information, we introduce a multi-directional interaction module. This module allows textual and correlational information to reinforce each other and amplify the advantages of different information flows for better prediction performance.

Take the *Text Memory Network* as an example, in each layer we firstly calculate the self-influence by using $v_{\text{text}}^{(h-1)}$ as the query:

$$v_{\text{text}->\text{text}}^{\text{Att}(h)} = \text{MultiheadAtt}(E_{\text{text}}, v_{\text{text}}^{(h-1)}) \quad (11)$$

Next we consider the interactive influences from correlations to texts using $v_{\text{corr}}^{(h-1)}$ as the query:

$$v_{\text{corr}->\text{text}}^{\text{Att}(h)} = \text{MultiheadAtt}(E_{\text{text}}, v_{\text{corr}}^{(h-1)}) \quad (12)$$

Finally, we produce a new attentional continuous representation by averaging these two influences:

$$v_{\text{text}}^{\text{Att}(h)} = \frac{v_{\text{text}->\text{text}}^{\text{Att}(h)} + v_{\text{corr}->\text{text}}^{\text{Att}(h)}}{2}, \quad (13)$$

which means that we replace Eqs. 6 with Eqs. 11-13 to obtain the new attention-aggregated vector.

The workings of *Stock Correlation Memory Network* are quite similar.

Consequently, the fusion of different information flows is promoted due to the multi-directional interaction mechanism in which CMIN learns not only the influences from text/correlation to movement prediction within each information flow but also the interactive influences between different information flows, representing the interrelationship between text and correlations.

4.5 Learning Objective

With the continuous representations $v_{\text{text}}^{(H)}$ and $v_{\text{corr}}^{(H)}$ from the last layer of each memory network, along with the target stock representation v_{target}, we concatenate them and apply a softmax function to generate the final prediction vector $\hat{y}$:

$$\hat{y} = \text{softmax}(W_y[v_{\text{text}}^{(H)}, v_{\text{target}}, v_{\text{corr}}^{(H)}] + b_y). \quad (14)$$

The objective is to minimize the cross entropy loss:

$$\mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{\text{n}} (y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)) \quad (15)$$

where n is the size of the training set.

5 Experiments

In this section, we empirically evaluate our CMIN model with three real-world datasets collected from the U.S. and Chinese stock markets.

5.1 Experimental Settings

5.1.1 Datasets

In our experiments we have used three datasets, namely **ACL18**, **CMIN-US** and **CMIN-CN**, spanning different time periods to evaluate our proposed model CMIN against other baselines.

ACL18 (Xu and Cohen, 2018) is a classic dataset with tweets from Twitter as financial texts in the task of text-enhanced stock movement prediction. As there are few existing high-quality datasets containing both texts and price, we are also making available two new benchmark datasets along with this paper from 2018-01-01 to 2021-12-31 in the U.S. and Chinese market named **CMIN-US** and **CMIN-CN**. These two datasets are available at `https://github.com/BigRoddy/CMIN-Dataset` to facilitate further research and enable reproducibility. More details and statistics of those three datasets are in Appendix A.

5.1.2 Baselines

We compare CMIN against the following four baselines, all of which are high-performing stock movement prediction models proposed by recent studies:

•**ALSTM**(Qin et al., 2017) is a dual-stage attention-based recurrent neural network, which selects relevant time series across all time steps.

•**Adv-LSTM**(Feng et al., 2019) uses adversarial training to improve the generalization of ALSTM.

•**Stocknet**(Xu and Cohen, 2018) introduces recurrent continuous latent variables and uses variational inference to address the posterior inference.

•**DTML**(Yoo et al., 2021) is a newly published attention-based model that exploits the correlations between stocks to improve the prediction accuracy.

5.1.3 Evaluation metrics

As we have formulated stock price movement prediction as a classification problem, we choose two classic metrics: Accuracy (Acc.) and Matthews Correlation Coefficient (MCC), similar to the previous work (Xu and Cohen, 2018; Yoo et al., 2021).

$$Acc. = \frac{tp + tn}{tp + tn + fp + gn} \tag{16}$$

$$MCC = \frac{tp \times tn - fp \times tn}{\sqrt{(tp + fp)(fn + tp)(fn + tn)(fp + tn)}} \tag{17}$$

Models	ACL18		CMIN-US		CMIN-CN	
	Acc.	**MCC**	**Acc.**	**MCC**	**Acc.**	**MCC**
ALSTM	51.81	0.032	51.64	0.006	53.35	0.023
Adv-LSTM	52.75	0.052	51.73	0.012	53.49	0.025
Stocknet	58.23	0.081	52.46	0.022	54.53	0.045
DTML	57.44	0.191	52.06	0.031	54.42	0.083
CMIN	**62.69**	**0.209**	**53.43**	**0.046**	**55.28**	**0.111**

Table 1: A comparison of prediction accuracy between CMIN and other baselines on three different datasets, where CMIN achieves state-of-the-art performance across both Acc. and MCC metrics.

5.1.4 Implementation details

We set our model for daily price prediction, with a history market window size $k = 5$ and the number of price features $d_p = d = 3$, namely the highest, the lowest and the closing prices. We limit the maximum number of financial texts in one single day to be $l = 20$, and the maximum length of a text document $w = 30$. Within the *Text Encoder*, we set the size of word embedding vector $d_w = 50$ and the hidden state of Bi-GRU network $d_m = 50$.

We implement the CMIN with Pytorch on a NVIDIA Tesla V100 and train it with an Adam optimizer (Kingma and Ba, 2015). All parameters of our model are initialized with Xavier Initialization (Glorot and Bengio, 2010). We search the hyperparameters of CMIN as follows: number of layers of each memory network H in $\{1, 2, 3, 4, 5\}$, dropout rate in $\{0.1, 0.2, 0.3\}$, number of epochs in $\{10, 20, 50\}$, and size of the price hidden state d_p in $\{3, 10, 50\}$. For baselines, we use their default parameters and fine-tune them to fit our data.

5.2 Performance Analysis

The results are summarized in Table 1.

Among all models, ALSTM and Adv-LSTM performed poorly with little improvement over random prediction. This could be attributed to the fact that these models rely solely on stock prices as the basis for decision-making. The Stocknet and DTML incorporate additional information beyond stock prices, demonstrated significant improvements over ALSTM and Adv-LSTM, which highlights the importance of utilizing financial texts and stock correlations for this challenging task. CMIN outperformed all baselines and achieved state-of-the-art performance on both two metrics across all datasets, showing its excellent capabilities to leverage both financial texts and stock correlations, as well as capture their interrelationship.

Models	ACL18		CMIN-US	
	Acc.	MCC	Acc.	MCC
CMIN-TE	52.88	0.0394	50.96	0.0134
CMIN-PR	57.83	0.0498	52.55	0.0162
CMIN-CM	54.76	0.1474	52.98	0.0279
CMIN-MI	60.22	0.1535	53.38	0.0380
CMIN	**62.69**	**0.2090**	**53.43**	**0.0460**

Table 2: Performance of CMIN variants on ACL18 and CMIN-US datasets, showing every component makes an important contribution to the excellent performance.

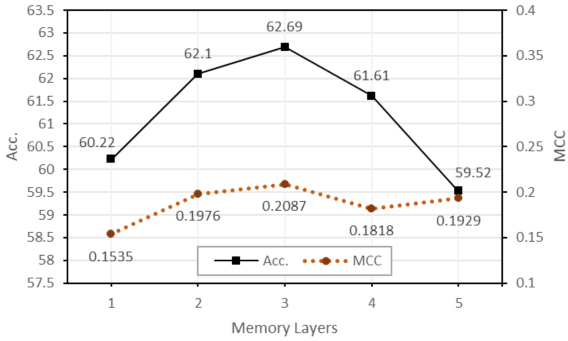

Figure 2: Performance of CMIN with a different number of memory network layers H on ACL18. CMIN achieves its best performance with 3 memory layers.

5.3 Ablation Studies

To evaluate the contribution of CMIN's different components, we compare against several variants:

•**CMIN-TE**: CMIN without the Text (TE), which makes decisions just based on stock prices.

•**CMIN-PR**: CMIN without the Price (PR), which makes decisions just based on related texts.

•**CMIN-CM**: CMIN without the guide of causality matrix (CM).

•**CMIN-MI**: CMIN without multi-directional interactions (MI) between memory networks.

The results are summarized in Table 2. CMIN-TE only achieves a level of prediction accuracy on par with ALSTM and Adv-LSTM, and is worst among all the variants, again indicating the importance of text data. Similar to the performance of Stocknet, CMIN-PR has a relatively high Acc. but a low MCC, suggesting texts are particularly helpful to predict on one side of the binary classification. By modeling both text data and stock relationships, CMIN-CM reaches a good result. Finally, better performance achieved when causality matrix and multi-directional interactions are introduced into the network. Overall, the ablation studies show that every component makes an important contribution to CMIN, and as a result the full model with all components achieves the best performance.

5.4 Analysis of Memory Network Depth

As introduced before, we propose two memory networks to retain vital features of texts and correlations with multiple computational layers. And we want to understand what would be the ideal number of depths to achieve the best prediction results.

We change the number of layers H of each memory network to find out how the performance fluctuates with it. The results are summarized in Figure 2. When we only have one memory layer, there is no multi-directional information flows between the two memory networks and as a result they only try to identify the vital information in the embeddings related to or having an impact on the movement of the target stock under the supervision of v_{target}. As the number of memory layers increases, the interactions between two memory networks also intensifies. It is intuitive that the performance of CMIN reaches its peak when it has three memory layers. With further increase the number of memory layers, CMIN is prone to overfit.

5.5 Case Study

Here we present an example to illustrate how CMIN considers both financial texts and stock correlations to avoid random noises in time series.

We visualized the causality matrix of ACL18 using a heat map as shown in Figure 3. Stocks are sequenced by their industry sector. The black box on the left shows weak causality, representing weak information flow from Utilities to Materials. On the other hand, the yellow box on the right indicates the relative strong information flow from Materials to Finance and within the Finance industry.

The target stock is Bank Of America (BAC) with a monitor window spanning from 13/11/2015 to 19/11/2015. We employ CMIN to predict BAC's next movement direction on the day of 20/11/2015 and then output the attention scores of texts and causality-guided correlation. The most focused stock by CMIN is Berkshire Hathaway Inc. (BRK-A). It's interesting to note that both are in the same industry sector: Finance, and they do appear to follow a very similar movement pattern in the trading days leading to 20/11/2015, which demonstrates the ability of CMIN to find the dynamic stock correlations with the guidance of Causality Matrix.

The financial text of BAC that obtains the highest attention score is "Beer, Credit Card Debt And

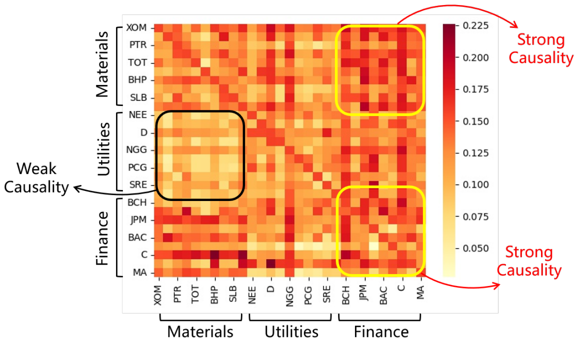

Figure 3: Visualization of partial causal matrix (covering three industries) for ACL18. Each entry in the matrix represents the transfer entropy between two stocks and the redder the square, the stronger the causality.

Other Positives For Bank Of America", the title of an news article[1] which reports the rapidly-improving banking landscape in the U.S.. This text is clearly highly relevant to BAC's subsequent stock performance, which demonstrates that CMIN is able to identify highly relevant texts having a impact on the target stock movement.

Furthermore, it also illustrates the underlying interrelationship between financial texts and stock correlations. Except expressing an optimistic sentiment towards BAC, the news also shows a rapidly improving state of affairs for the wider financial industry. Therefore, through the Multi-directional Interactions mechanism, the text strengthens the model's attention to stocks in the same sector. These two aspects mutually reinforce and complement each other to help the model make the best judgment that BAC's stock price will rise on the next day.

6 Conclusions

In this paper, we proposed CMIN, a causality-guided multi-memory interaction network that simultaneously models financial documents, causality-enhanced stock correlations and the interactions between the two, and recurrently learns a global memory representation for movement prediction. This multi-modality network was designed to enable the concurrent discovery of texts and stock correlations relevant to future price change and we demonstrated, through experiments on three datasets across two distinct markets, that each component of the proposed architecture made significant contributions to the model, leading CMIN to achieve state-of-the-art accuracy.

[1] https://seekingalpha.com/article/3692516

Limitations

We discuss the limitations of our model as follows:

1. Due to the natural uncertainty of financial forecast, although we have taken many methods to improve the generalization performance of the model (such as limiting the depth of memory layers and with the assistance of auxiliary data), creating a trustworthy application requires considering many other factors beyond the algorithmic level. We advise that users monitor the model's performance over time and regularly update it to adapt to ever-changing market conditions.

2. This paper uses Granger causality based on transfer entropy to make a preliminary attempt to introduce causality between time series to model the similarity between stocks more accurately. But this description is junior and classical, and there are lots of more modern methods to measure precise causality in mathematics (like PC algorithm), which we believe would further improve the performance.

3. We only experiment the performance of model on the task of binary classification, leaving more complex tasks (such as regression task and returns prediction) and simulating actual investment to evaluate the capability and potential of the model comprehensively.

Acknowledgements

We would like to express our sincere gratitude to the reviewers for their valuable efforts and insightful contributions. This work was supported by National Natural Science Foundation of China (NSFC Grant No. 62122089), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China. Rui Yan is the corresponding author.

References

Gary Ang and Ee-Peng Lim. 2022. Guided attention multimodal multitask financial forecasting with inter-company relationships and global and local news. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6313–6326.

Kunal Bhardwaj. 2021. Convolutional neural network(cnn/convnet) in stock price movement prediction. *CoRR*, abs/2106.01920.

Allan Borodin, Ran El-Yaniv, and Vincent Gogan. 2003. Can we learn to beat the best stock. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. 2019. Enhancing stock movement prediction with adversarial training. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5843–5849. International Joint Conferences on Artificial Intelligence Organization.

Fernando G. D. C. Ferreira, Amir H. Gandomi, and Rodrigo T. N. Cardoso. 2021. Artificial intelligence applied to stock market trading: A review. *IEEE Access*, 9:30898–30917.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Hakan Gunduz. 2021. An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination. *Financial Innovation*, 7(1):1–24.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 261–269, New York, NY, USA. Association for Computing Machinery.

Zhigang Jin, Yang Yang, and Yuhong Liu. 2020. Stock closing price prediction based on sentiment analysis and lstm. *Neural Computing and Applications*, 32:9713–9729.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jiahong Li, Hui Bu, and Junjie Wu. 2017. Sentiment-aware stock market prediction: A deep learning method. In *2017 International Conference on Service Systems and Service Management*, pages 1–6. IEEE.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143, Melbourne, Australia. Association for Computational Linguistics.

Xuechen Li, Xinfang Ma, Fengchao Xiao, Cong Xiao, Fei Wang, and Shicheng Zhang. 2022. Time-series production forecasting method based on the integration of bidirectional gated recurrent unit (bi-gru) network and sparrow search algorithm (ssa). *Journal of Petroleum Science and Engineering*, 208:109309.

Jiawei Long, Zhaopeng Chen, Weibing He, Taiyu Wu, and Jiangtao Ren. 2020. An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in chinese stock exchange market. *Applied Soft Computing*, 91:106205.

David M. Q. Nelson, Adriano C. M. Pereira, and Renato A. de Oliveira. 2017. Stock market's price movement prediction with lstm neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1419–1426. IEEE.

Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison W. Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2627–2633.

Leonidas Sandoval Junior, Asher Mullokandov, and Dror Y Kenett. 2015. Dependency relations among international stock market indices. *Journal of Risk and Financial Management*, 8(2):227–265.

Anil Seth. 2007. Granger causality. *Scholarpedia*, 2(7):1667.

Omer Berat Sezer and Ahmet Murat Özbayoglu. 2020. Financial trading model with stock bar chart image time series with deep convolutional neural networks. *Intelligent Automation and Soft Computing*, 26.

Catalin Stoean, Wiesław Paja, and Ruxandra Stoean. 2019. Deep architectures for long-term stock price prediction with a heuristic-based strategy for trading simulations. *PloS one*, 14(10):e0223593.

Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Ning Tang, Yanyan Shen, and Junjie Yao. 2020. Learning to fuse multiple semantic aspects from rich texts for stock price prediction. In *Web Information Systems Engineering – WISE 2019: 20th International Conference, Hong Kong, China, January 19–22, 2020, Proceedings*, page 65–81, Berlin, Heidelberg. Springer-Verlag.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jingyuan Wang, Yang Zhang, Ke Tang, Junjie Wu, and Zhang Xiong. 2019a. Alphastock: A buying-winners-and-selling-losers investment strategy using

Dataset	Country	Stocks	Data Resources		Data Range		
			Price	Text	Train	Development	Test
ACL18	US	87	Yahoo Finance	Twitter	2014-01-01 to 2014-12-31	2015-01-01 to 2015-10-01	2015-10-01 to 2015-12-31
CMIN-US	US	110	Yahoo Finance	Yahoo	2018-01-01 to 2021-04-30	2021-05-01 to 2021-08-31	2021-09-01 to 2021-12-31
CMIN-CN	China	300		Wind			

Table 3: Summary stats on experiment datasets. CMIN-US and CMIN-CN are two new benchmark datasets we are making available alongside this paper.

interpretable deep reinforcement attention networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 1900–1908, New York, NY, USA. Association for Computing Machinery.

Yaowei Wang, Qing Li, Zhexue Huang, and Junjie Li. 2019b. Ean: Event attention network for stock price trend prediction based on sentimental embedding. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 311–320, New York, NY, USA. Association for Computing Machinery.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multi-modal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):371–378.

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1970–1979, Melbourne, Australia. Association for Computational Linguistics.

Jaemin Yoo, Yejun Soun, Yong-chan Park, and U Kang. 2021. Accurate multivariate stock movement prediction via data-axis transformer with multi-level contexts. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 2037–2045, New York, NY, USA. Association for Computing Machinery.

Liheng Zhang, Charu Aggarwal, and Guo-Jun Qi. 2017. Stock price prediction via discovering multi-frequency trading patterns. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 2141–2149, New York, NY, USA. Association for Computing Machinery.

A Dataset Details

The statistics of datasets are summarized in Table 3.

ACL18 (Xu and Cohen, 2018) consists of 87 stocks of 9 industries from the U.S. stock market. It also includes two types of data: tweets from Twitter and historical stock prices from Yahoo finance. We have processed in the same way as described in (Xu and Cohen, 2018).

As there are few existing high-quality datasets containing both texts and prices, we are also making available two new benchmark datasets along with this paper from 2018-01-01 to 2021-12-31: **CMIN-US** includes the top 110 stocks from US by market capitalisation; **CMIN-CN** consists of all 300 constituents of CSI300, a major Chinese stock market index. Similar to ACL18, both CMIN-US and CMIN-CN include financial texts as well as historical stock prices data. The historical price data in both datasets comes from Yahoo Finance. The text data of CMIN-US is collected from Yahoo finance[2] and CMIN-CN from Wind [3]. In our experiments, we have used news headlines instead of the entire texts for efficiency and noise reduction.

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

[3]https://www.wind.com.cn/

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
the last section

☑ A2. Did you discuss any potential risks of your work?
the last section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
the first section

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 5

☑ B1. Did you cite the creators of artifacts you used?
Section 5.1

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 5.1

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 5.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5.1.1 and Appendix A

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5.1.4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1.4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.