DialogueGAT: A Graph Attention Network for Financial Risk Prediction by Modeling the Dialogues in Earnings Conference Calls

Yunxin Sang and **Yang Bao**[*]
Antai College of Economics and Management
Shanghai Jiao Tong University
sangyunxin@gmail.com baoyang@sjtu.edu.cn

Abstract

Financial risk prediction is an essential task for risk management in capital markets. While traditional prediction models are built based on the hard information of numerical data, recent studies have shown that the soft information of verbal cues in earnings conference calls is significant for predicting market risk due to its less constrained fashion and direct interaction between managers and analysts. However, most existing models mainly focus on extracting useful semantic information from the textual conference call transcripts but ignore their subtle yet important information of dialogue structures. To bridge this gap, we develop a graph attention network called DialogueGAT for financial risk prediction by simultaneously modeling the speakers and their utterances in dialogues in conference calls. Different from previous studies, we propose a new method for constructing the graph of speakers and utterances in a dialogue, and design contextual attention at both speaker and utterance levels for disentangling their effects on the downstream prediction task. For model evaluation, we extend an existing dataset of conference call transcripts by adding the dialogue structure and speaker information. Empirical results on our dataset of S&P1500 companies demonstrate the superiority of our proposed model over competitive baselines from the extant literature.

1 Introduction

Financial risk prediction is an essential task for risk management in capital markets since risk is one of the most important variables for making investment decisions in tasks such as portfolio selection, asset pricing, and so on (Poon and Granger, 2003). Due to its importance in risk assessment, the accurate prediction of financial risk is of great interest to academic and industry stakeholders in artificial intelligence, economics, and finance (Kogan et al., 2009).

Traditionally, the risk prediction models are built based on the hard information of numerical data such as the historical stock return volatility (Kogan et al., 2009). But recent studies have shown that the soft information of textual data in related corporate disclosures is incrementally informative over the conventional numerical data for predicting corporate risks (Matsumoto et al., 2011; Bao and Datta, 2014). In this study, we focus on financial risk prediction using a particular type of textual corporate disclosures – the transcripts of earnings conference calls. Earnings conference calls in conjunction with earnings releases have become an increasingly important form of voluntary corporate disclosure. In conference calls, managers (e.g., CEO, CFO, or other executives) can voluntarily present information of firm performance during the quarter, and interested participants such as analysts and investors can also directly engage in information disclosure in a follow-up Q&A session. Due to its less constrained fashion relative to the mandated corporate disclosures (e.g., annual reports) and direct interaction between managers and analysts, these conference calls have been recognized as significant information events to the market (Matsumoto et al., 2011).

Recently, there have been some models that make use of textual conference call transcripts for predicting financial risk (Qin and Yang, 2019; Theil et al., 2019). Unfortunately, except for very few exceptions (Ye et al., 2020), most existing models mainly focus on extracting useful semantic information from the textual conference call transcripts but ignore their subtle yet important information of dialogue structures. Specifically, the earnings conference calls will affect the risk perceptions of market investors not only by what is said (i.e., utterance) but also by who said it (i.e., speaker). For example, the managers inside the company usually hold private information but might be reluctant to disclose the negative information, while

[*] Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1623–1633
December 7-11, 2022 ©2022 Association for Computational Linguistics

the analysts outside the company usually ask harsh questions that may be of interest to the investors.

To bridge the aforementioned gap, we develop a graph attention network called DialogueGAT[1] for financial risk prediction by simultaneously modeling the speakers and their utterances in dialogues in conference calls. Different from previous studies, we propose a new method for constructing the graph of speakers and utterances in a dialogue, and design contextual attention at both speaker and utterance levels for disentangling their effects on the downstream task of financial risk prediction. The designed attention mechanisms also provide our model with reasonable interpretation ability. To evaluate model performance, we extend an existing dataset of quarterly conference call transcripts (Li et al., 2020) by adding the dialogue structure and speaker information. We measure the corporate financial risk by using stock return volatility - one of the most commonly used measures in prior research (Kogan et al., 2009). Empirical results on our extended dataset of S&P1500 companies demonstrate the superiority of our proposed model over competitive baselines from the extant literature. Supplementary studies are also conducted to examine the effectiveness of our model's key components, parameter sensitivity, and interpretability.

The remainder of the paper is organized as follows. First, we review the related works on financial risk prediction and dialogue-based graph neural networks. Then, we define the problem and elaborate on our proposed model. After that, we present and analyze the experimental results. Finally, we conclude the paper.

2 Related Work

This study mainly relates to two areas: (1) text-based financial risk prediction and (2) graph neural network for modeling dialogues.

Financial risk prediction is an essential task in capital markets. The early risk prediction models are built based on the hard information of numerical data, but recent studies have shown that the soft information of textual data is incrementally informative over the conventional numerical data for predicting corporate risks (Kogan et al., 2009). Two types of textual corporate disclosures are commonly used for financial risk prediction, including the annual reports (Bao and Datta, 2014) and earnings conference call transcripts (Matsumoto et al., 2011). Compared with mandated corporate disclosure such as annual reports, voluntary conference calls have been recognized as significant information events to the market due to their less constrained fashion and direct interaction between managers and analysts (Matsumoto et al., 2011). In this line of research, most existing models cast the risk prediction task as a standard text regression problem without considering the dialogue structures of conference calls (Qin and Yang, 2019; Theil et al., 2019; Li et al., 2020). The only exception is a recent model called MRQA (Ye et al., 2020), which proposes a multi-round Q&A attention network for considering the dialogue form. It is worth mentioning that this model ignores the speaker information when modeling the dialogue. We will use the MRQA model as a baseline in our experiment.

Our work is also related to recent studies using GNN (Graph Neural Network) for modeling dialogues. GNN is a type of neural network which directly operates on the graph structure, and it has achieved great success in many natural language processing tasks (Scarselli et al., 2008). Recently, some GNN models have been proposed for modeling dialogues, and the main challenge is how to construct a graph of utterances from dialogue in an effective manner (Hu et al., 2019; Banerjee and Khapra, 2019). In this line of research, DialogueGCN is a state-of-the-art model for dialogue-based emotion recognition without using external knowledge (Ghosal et al., 2019). This model is closely related to our model because we are one of the few dialogue-based GNN models that consider not only utterances but also speakers when constructing the dialogue graph. We will use the DialogueGCN model as a baseline in our experiment.

3 Methodology

In this section, we formulate our problem of finance risk prediction and then present our proposed DialogueGAT model.

3.1 Problem Definition

To measure the corporate financial risk, we use the stock return volatility, which is a commonly used measure in prior research (Kogan et al., 2009). This volatility measure reflects the degree of variation of stock prices, and higher volatility indicates that the

[1] The code is available at https://github.com/sangyx/DialogueGAT.

firm's stock is riskier. Formally, the stock return volatility $v_{[t,t+\tau]}$ over the time period from the day t to the day $t+\tau$ is defined as:

$$v_{[t,t+\tau]} = \sqrt{\sum_{i=0}^{\tau} \left(r_{t+i} - \bar{r}\right)^2 / \tau} \qquad (1)$$

where r_t is the dividend-adjusted return of a specific stock on the day t and $\bar{r}$ is the mean of dividend-adjusted returns over the time period from day t to day $t+\tau$. The dividend-adjusted return is defined as $r_t = \frac{P_t}{P_{t-1}} - 1$, where P_t is the dividend-adjusted closing price on the day t. We set the time window size τ to 3, 7, and 15 days based on the PEAD (Post-Earnings-Announcement Drift) theory in accounting literature (Bernard and Thomas, 1989) – the stock's cumulative abnormal returns tend to drift for several weeks following the earnings announcement. We conjecture that the return volatility with a smaller window size τ is more difficult to predict because the mean reversion theory in finance posits that the stock prices are more volatile in the short run and will eventually revert to the long-term average.

We formulate our financial risk prediction problem as a supervised regression task and propose to leverage the dialogue structures of earnings conference calls for improving the prediction of stock return volatility. More specifically, given a dialogue with M speakers and N utterances in a firm's conference call held on the day t, we aim to predict the firm's future stock return volatility over the period from the trading day t to $t+\tau$. We use both presentations by managers and Q&A between managers and analysts for constructing the dialogues. Each dialogue of a conference call is a list of pairs of speakers and utterances that are sorted in the temporal order of dialogue. It is worth noting that each utterance consists of all the sentences uttered by a speaker in each dialogue round.

3.2 Proposed Model

We develop a model called DialogueGAT (Dialogue Graph Attention Network) for jointly modeling the dialogue structure of utterances and speakers in order to improve the prediction of financial risk. As shown in Figure 1, the architecture of our proposed model is composed of four main modules: (1) an utterance encoder which uses the TextCNN (Text Convolutional Neural Network) for learning the representation of utterances, (2) a graph encoder that uses the GAT (Graph Attention Network) for jointly learning the better representation of utterances and speakers in our constructed graph, (3) two contextual attention layers which aggregate the embedding vectors of utterances and speakers in a conference call, and (4) an output layer which fuses different types of information for the downstream task of financial risk prediction. Next, we describe these modules in detail.

3.2.1 Utterance Encoder

We employ the TextCNN for learning the embedding vector $u_i \in \mathbb{R}^d$ for utterance i. The input word tokens are first represented using the pre-trained 300-dimensional GloVe embedding vectors[2], and then fed into the TextCNN with default parameter settings as in (Kim, 2014). A max-overtime pooling operation is applied over the feature maps, and the pooled features are concatenated to obtain the embedding vector u_i of the utterance i.

3.2.2 Graph Encoder

The core component of our model is a graph encoder for jointly learning the representation of utterances and speakers. In this module, we propose a new simple yet effective method for constructing the graph of utterances and speakers for each dialogue. Specifically, we treat both utterances and speakers as graph nodes and generate the edges between nodes in the following two ways: (1) each utterance node is connected to its previous and next utterance nodes for capturing the local context, and (2) each utterance node is connected to its speaker node for capturing the speaker context. This way, we construct an undirected dialogue graph shown in the GAT module in Figure 1. We also provide a more detailed illustration of our dialogue graph in Figure A1 in Appendix. It is worth mentioning that we keep our model implementation parsimonious and effective by converting our original heterogeneous graph to a homogeneous graph.

Once the dialogue graph is constructed, we utilize the GAT model (Veličković et al., 2018) for jointly learning the representation of utterance and speaker nodes. We initialize the utterance node i by using its embedding vector $u_i^{(0)}$ obtained from the TextCNN in utterance encoder (note: the superscript indicates the layer). To capture the speaker information within and across conference calls, we assign a trainable speaker embedding vector p_j to

[2] The pre-trained GloVe model used in this study is available at `https://nlp.stanford.edu/data/glove.840B.300d.zip`.

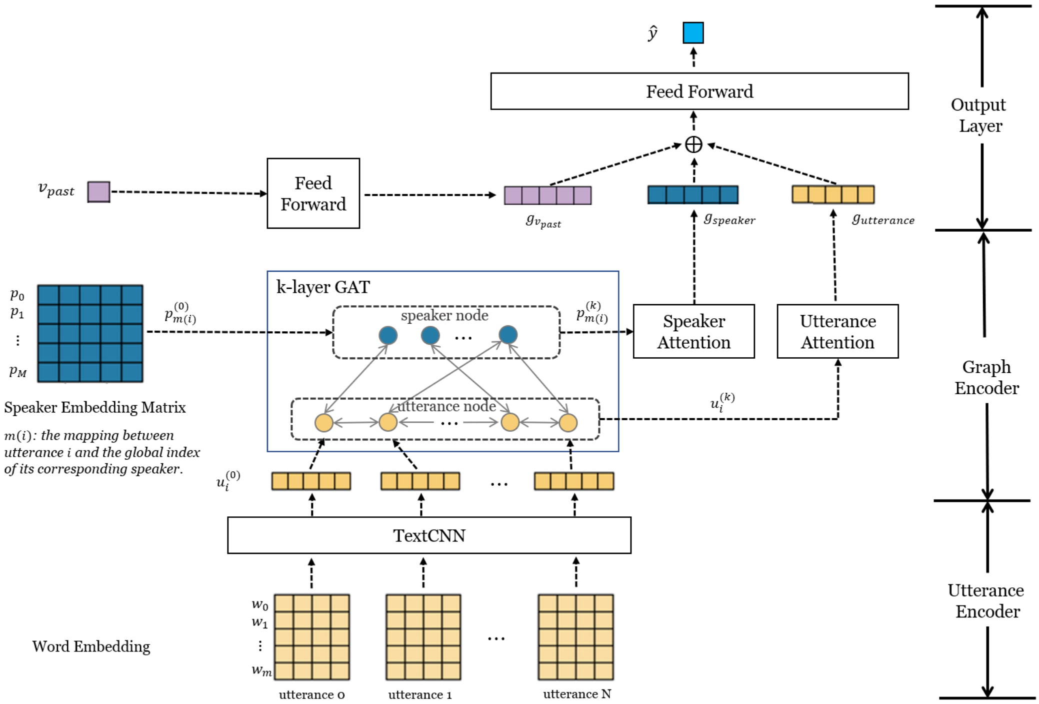


Figure 1: The architecture of our proposed DialogueGAT model.

each speaker j and perform orthogonal initialization for these vectors. Each speaker embedding vector will be dynamically updated every time the corresponding speaker (either manager or analyst) attends a conference call. The node embedding vector $x_i^{(l+1)}$ at the layer $l+1$ is updated by using an attention mechanism to aggregate the one-hop neighborhood from the layer l:

$$x_i^{(l+1)} = \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} W_x x_j^{(l)} \tag{2}$$

where x is either the utterance node u or the speaker node p, $\mathcal{N}(i)$ is the set of one-hop neighbors of node i, and $W_x \in \mathbb{R}^{d \times d}$ is a trainable transformation matrix. α_{ij} is the pairwise attention score between the nodes i and j which is calculated as:

$$\begin{aligned} e_{ij} &= \text{LeakyReLU}(W_a(W_f x_i \oplus W_f x_j)) \\ \alpha_{ij} &= \frac{\exp(e_{ij})}{\sum_{o \in \mathcal{N}(i)} \exp(e_{io})} \end{aligned} \tag{3}$$

where $\oplus$ denotes concatenation.

By defining a multi-layer (i.e., k-layer) GAT model, we are able to aggregate the multi-hop neighborhoods for learning node embedding vectors. We also attempt to avoid overfitting by adding a dropout layer between GAT layers.

3.2.3 Utterance Attention and Speaker Attention

We design two contextual attention layers (i.e., utterance attention and speaker attention) for separately aggregating the node embedding vectors of utterances and speakers in a conference call. In this way, we can disentangle the effects of utterances and speakers on the downstream task of financial risk prediction. Specifically, we calculate the attention score β_i for each utterance node embedding in the vector set $\{u_0^{(k)}, u_1^{(k)}, \cdots, u_N^{(k)}\}$ obtained from the graph encoder as follows:

$$h_i = \tanh\left(W_u u_i^{(k)} + b_u\right) \tag{4}$$

$$\beta_i = \frac{\exp\left(h_i^\top u_{context}\right)}{\sum_{j=0}^{N} \exp\left(h_j^\top u_{context}\right)} \tag{5}$$

where h_i is the hidden representation of $u_i^{(k)}$ by using a dense layer, and $u_{context}$ is a context vector which is randomly initialized and jointly learned during the training process. The global utterance vector $g_{utterance} \in \mathbb{R}^d$ of a conference call is then calculated as the following weighted sum:

$$g_{utterance} = \sum_{i=0}^{N} \beta_i u_i^{(k)} \tag{6}$$

The global speaker embedding vector $g_{speaker} \in \mathbb{R}^d$ of a conference call is calculated in the same way as $g_{utterance}$ by using the speaker attention.

3.2.4 Output Layer

The last component of our model is an output layer that combines different types of feature representations for predicting stock return volatility. In addition to utterance and speaker embedding vectors, we also allow the inclusion of past stock return volatility v_{past} as input because it is a strong predictor, as shown in prior research (Kogan et al., 2009). As shown in Figure 1, we align the dimensions of different feature vectors via a fully-connected neural network layer. The aligned feature vectors $g_{utterance}, g_{speaker}, g_{v_{past}} \in \mathbb{R}^d$ are concatenated together and then fed into a fully-connected layer for predicting the stock return volatility:

$$\hat{y} = W_g(g_{utterance} \oplus g_{speaker} \oplus g_{v_{past}}) + b_g \quad (7)$$

4 Experimentation

In this section, we conduct empirical studies to evaluate model performance. We first describe our dataset and baselines and then present and analyze the experimental results.

4.1 Dataset

Although there are some recent datasets of earnings conference calls (Qin and Yang, 2019; Li et al., 2020), none of them has the speaker information required for speaker-aware dialogue models. To tackle this problem, we extend the most comprehensive MAEC (Multimodal Aligned Earnings Conference Call) dataset[3] (Li et al., 2020) by adding the dialogue structure and speaker information. Like the MAEC dataset, our extended dataset consists of conference call transcripts of S&P1500 companies from February 25, 2015 to June 21, 2018. We parse the dialogue structure of each conference call transcript, and add the speaker information of managers and analysts collected from the SeekingAlpha website (`https://seekingalpha.com`). Following (Li et al., 2020), we preserve the temporal order of conference calls and split the dataset into training/validation/testing sets in the ratio of 7:1:2 on a yearly basis (the years 2017 and 2018 are merged). Table 1 summarizes the descriptive statistics of our dataset. It is worth mentioning that our dataset contains more sentences than the MAEC dataset because the latter drops some sentences when aligning the textual and audio data. We drop 28 conference call samples that lack speaker information.

4.2 Baselines

We compare our model with the following competitive baseline models from the extant literature on financial risk prediction using conference calls (Ye et al., 2020; Yang et al., 2021). We noticed that some baselines ignore the historical stock return volatility v_{past} – a strong predictor of financial risk (Kogan et al., 2009). Hence, to compare models on an equal footing, we use the historical stock return volatility v_{past} as an additional feature for all models. Specifically, we use a fully-connected layer to expand the dimension of v_{past} and concatenate it with the output vector of baselines for the downstream prediction.

(1) v_{past} & *SVR*$_{v_{past}}$. These are the two simple yet competitive baselines without using textual data of conference call transcripts (Kogan et al., 2009). v_{past} directly uses the historical stock return volatility v_{past} in the past τ days to predict the return volatility in the future τ days. *SVR*$_{v_{past}}$ is the SVR model with linear kernel using v_{past} as the only feature variable.

(2) *HAN*. This baseline uses the HAN (Hierarchical Attention Networks) model (Yang et al., 2016) to exploit the textual data of conference calls for financial risk prediction. This model considers the hierarchical structure of documents and has two levels of attention mechanisms at both word and sentence levels. But this model does not consider the dialogue structure of conference call transcripts.

(3) *ProFET*. This is a competitive model for Predicting the Risk of Firms from Event Transcripts (PRoFE) (Theil et al., 2019). It combines the BiLSTM and an attention module to predict the stock return volatility by using both financial and textual features of earnings conference calls.

(4) *BERT & XLNet*. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) and XLNet (Yang et al., 2019) are the two widely used pre-trained language models which have achieved the start-of-the-art results on many downstream prediction tasks. We use the pre-trained *bert-base-cased* and *xlnet-base-cased* models to encode textual utterances in conference calls. For the BERT model, the hidden state of the

[3] The dataset is available at `https://github.com/Earnings-Call-Dataset/MAEC-A-Multimodal-Aligned-Earnings-Conference-Call-Dataset-for-Financial-Risk-Prediction`.

Table 1: Descriptive statistics of our extended dataset.

Year	2015	2016	2017 - 2018
#Companies	523	897	736
#Speakers	5,840	8,768	7,006
#Utterances	59,549	108,714	79,253
#Sentences	102,142	184,936	145,206
Training set (#Samples)	25/02/2015 - 22/10/2015 (531)	05/01/2016 - 03/08/2016 (968)	17/01/2017 - 07/11/2017 (890)
Validation set (#Samples)	22/10/2015 - 29/10/2015 (75)	03/08/2016 - 12/08/2016 (138)	07/11/2017 - 15/02/2018 (127)
Testing set (#Samples)	29/10/2015 - 17/12/2015 (153)	15/08/2016 - 15/11/2016 (278)	15/02/2018 - 21/06/2018 (255)

Table 2: Model performance in terms of MSE by varying the window size τ.

Year	2015			2016			2017-2018		
Methods	$\tau = 3$	$\tau = 7$	$\tau = 15$	$\tau = 3$	$\tau = 7$	$\tau = 15$	$\tau = 3$	$\tau = 7$	$\tau = 15$
v_{past}	1.0905	0.5441	0.2744	1.3542	0.7300	0.4465	1.1739	0.5681	0.2723
SVR$_{v_{past}}$	0.6576	0.4393	0.2503	0.6440	0.3810	0.2714	0.5643	0.3634	0.2273
HAN	0.5421	0.4259	0.2516	0.5272	0.3390	0.2501	0.5186	0.3554	0.2231
ProFET	0.5902	0.4297	0.2471	0.5737	0.3717	0.2502	0.5341	0.3605	0.2270
BERT	0.5628	0.4214	0.2653	0.5256	0.3429	0.2472	0.5411	0.3596	0.3032
XLNET	0.5537	0.4167	0.2723	0.5385	0.3368	0.2560	0.5396	0.3798	0.2652
DialogueGCN	0.5376	0.4138	0.2462	0.5209	0.3343	0.2472	0.5019	0.3494	0.2204
MRQA	0.5174	0.4126	0.2407	0.5162	0.3314	0.2286	0.4966	0.3443	0.2240
DialogueGAT	**0.4530**	**0.3236**	**0.1898**	**0.4549**	**0.2884**	**0.1810**	**0.4090**	**0.2886**	**0.2036**

last layer on the [CLS] token is used to represent each utterance. For the XLNET model, we use the last token hidden state to represent each utterance. We represent each document of the conference call by averaging all the utterance feature vectors and then use the SVR model to predict the stock return volatility.

(5) *DialogueGCN*. This is a competitive model for emotion recognition in dialogues without using external knowledge (e.g., large BERT-like pretrained models) (Ghosal et al., 2019)[4]. It constructs a graph of utterances in dialogue and uses the GCN (Graph Convolutional Network) for modeling the conversational context. Although this model is originally designed for emotion recognition, it can be directly used to leverage the dialogue structure of conference calls for our task of financial risk prediction. Unlike our method, the DialogueGCN only constructs the graph of utterance nodes and indirectly leverages the speaker information by linking together the utterance nodes of different speakers within a conference dialogue.

(6) *MRQA*. To our knowledge, the MRQA (Multi-Round Q&A) is the state-of-the-art attention-based model for financial risk prediction using conference calls (Ye et al., 2020). Unlike our model, this model does not use speaker information but exploits the dialogue information of the multi-round Q&A structure. Specifically, the MRQA uses the BiLSTM to encode textual features of conference calls and designs two modules for modeling the dialogue structure, including an RSS (Reinforced Sentence Selector) module for selecting important sentences in the Q&A segments, and an RBAN (Reinforced Bidirectional Attention Network) module for exploring the interaction between questions and answers.

4.3 Model Evaluation

Next, we present and analyze our experimental results.

4.3.1 Experimental Settings

To measure the model performance for predicting stock return volatility, we use the evaluation metric MSE (Mean Squared Error), which is commonly used in prior research (Kogan et al., 2009). We split the data sample as aforementioned and train all the models on a single Nvidia RTX 2080 Ti GPU. We use the MSE (with L2 regularization) as a loss function for training and use the Adam algorithm (Kingma and Ba, 2014) for optimizing the loss. We tune the hyper-parameters of all models by performing a grid search on the validation set. The tuned parameters of our DialogueGAT model are as follows: the learning rate is 1e-5, the L2 penalty is 1e-6, the batch size is 4, the head of GAT is 5, the number of layers of GAT is 5, the dimension of feature vectors d is 300, and the dropout rate is 0.1. We train our model for a maximum of 100 epochs and stop training if the validation loss

[4] https://github.com/declare-lab/conv-emotion

does not decrease for 5 consecutive epochs.

4.3.2 Performance Comparison

The main results of our performance comparison are shown in Table 2. To compare all models on an equal footing, we include the historical stock return volatility v_{past} as a feature variable for all models. We summarize our main findings as follows.

First, our proposed DialogueGAT performs best among all models for predicting the future stock return volatility for the three window sizes $\tau = 3, 7, 15$ in all testing years. Paired two-tailed t-tests show that our DialogueGAT significantly outperforms: (1) v_{past}(τ = 3,7), SVR$_{v_{past}}$(τ = 3), HAN(τ = 3), ProFET(τ = 3,15), BERT(τ = 3), XLNET(τ = 3,15), MRQA(τ = 15) at the 1% significance level; (2) SVR$_{v_{past}}$(τ = 7), HAN(τ = 15), ProFET(τ = 7), BERT(τ = 7,15), XLNET(τ = 7), DialogueGCN(τ = 3,15), MRQA(τ = 3) at the 5% significance level; (3) v_{past}(τ = 15), HAN(τ = 7), DialogueGCN(τ = 7), MRQA(τ = 7) at the 10% significance level. This demonstrates the effectiveness of our DialogueGAT model for financial risk prediction.

Second, we observe that the models using both numerical historical stock return volatility and textual conference call transcripts (i.e., HAN, ProFET, BERT, and XLNET) outperform the models that only use the numerical historical stock return volatility (i.e., v_{past} and SVR$_{v_{past}}$). This implies that the textual information of conference calls is incrementally useful for financial risk prediction, which is consistent with prior research (Li et al., 2020).

Last and most importantly, we find that the models that further consider the dialogue structure (i.e., DialogueGCN, MRQA, and our DialogueGAT) perform better than the models that only exploit the semantic information of textual conference call transcripts (i.e., HAN, ProFET, BERT, and XLNET). This implies that the dialogue structure contains incremental information that is useful for improving financial risk prediction. Our DialogueGAT model performs best among all models that consider dialogue structure, demonstrating its effectiveness for modeling dialogues and the predictive power of speaker information.

4.4 Supplementary Analysis

We further conduct supplementary analysis to examine the effects of our model's key components, parameter sensitivity, and model interpretability.

4.4.1 Ablation Study

To examine the effects of our model's key components, we conduct the ablation study by evaluating the following variants of our DialogueGAT model: (1) *w/o* v_{past}: DialogueGAT model without using the feature of historical stock return volatility v_{past}. (2) *w/o speaker*: DialogueGAT model without using the speaker-related modules in Figure 1. (3) *random speaker embedding*: DialogueGAT model which removes the trainable speaker embedding matrix and uses a random vector to initialize the embedding of the speaker node. (4) *position embedding*: DialogueGAT model which adds a trainable position embedding to each utterance node embedding as in (Vaswani et al., 2017) for examining the usefulness of sequential information of utterances.

The results of our ablation study are shown in Table 3. It is worth noting that we only present the results in 2015 due to space limitation, but untabulated results show that the results remain in other testing years. We have the following observations from Table 3:

(1) The inclusion of v_{past} feature can improve our model performance, and its importance increases as the window size τ gets larger than 7.

(2) Our proposed speaker-related modules play a vital role in our DialogueGAT model since their performance will drop significantly if those modules are removed or replaced. Specifically, we can observe that the performance of the variant model *random speaker embedding* performs much worse than our full DialogueGAT model when the trainable speaker embedding vectors are replaced with the random embedding vectors. Moreover, the performance of the variant model *w/o speaker* will further drop because this variant not only removes the speaker information, but also destroys the structure of dialogue graph by removing all speaker nodes.

(3) The sequential information of utterances in conference calls is unnecessary and even harmful to model performance. This observation is consistent with the finding in a recent study (Yang et al., 2021). Hence, we do not use the position embedding in our full DialogueGAT model.

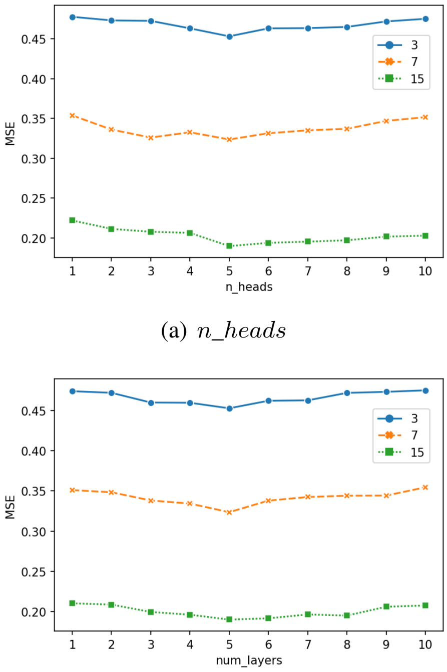
4.4.2 Parameter Sensitivity

We examine whether our model is sensitive to the parameters of our key module – the Graph Encoder in Figure 1. There are two important parameters in this module: (1) the number of heads n_heads in multi-head attention, and (2) the num-

Table 3: The performance of variants of our DialogueGAT model.

Variants of DialogueGAT	$\tau = 3$	$\tau = 7$	$\tau = 15$
w/o v_{past}	0.4711	0.3338	0.2276
w/o speaker	0.5493	0.4288	0.2473
random speaker embedding	0.5036	0.3590	0.2430
position embedding	0.4894	0.3569	0.2238
DialogueGAT (full model)	0.4530	0.3236	0.1898

ber of GAT layers n_layers. Here, the parameter n_layers indicates that each node in the dialogue graph can aggregate the information from its n-hop neighbors. We measure the model performance of our DialogueGAT by varying these two parameters while fixing all the other parameters, and the results in 2015 are shown in Figure 2. As can be seen, our model will perform better as the parameter n_heads or n_layers takes a larger value but tend to perform worse when n_layers and n_heads are larger than 5 because of the over-fitting and over-smoothing issue (Li et al., 2018). In addition, when each node in the dialogue graph can aggregate the information from its 5-hop neighbors (i.e., $n\_layers = 5$), our designed dialogue graph guarantees that all utterance nodes can obtain the local context information of their preceding and following speaker nodes.



(a) n_heads

(b) n_layers

Figure 2: The performance of our DialogueGAT model in 2015 by varying the parameters n_heads and n_layers.

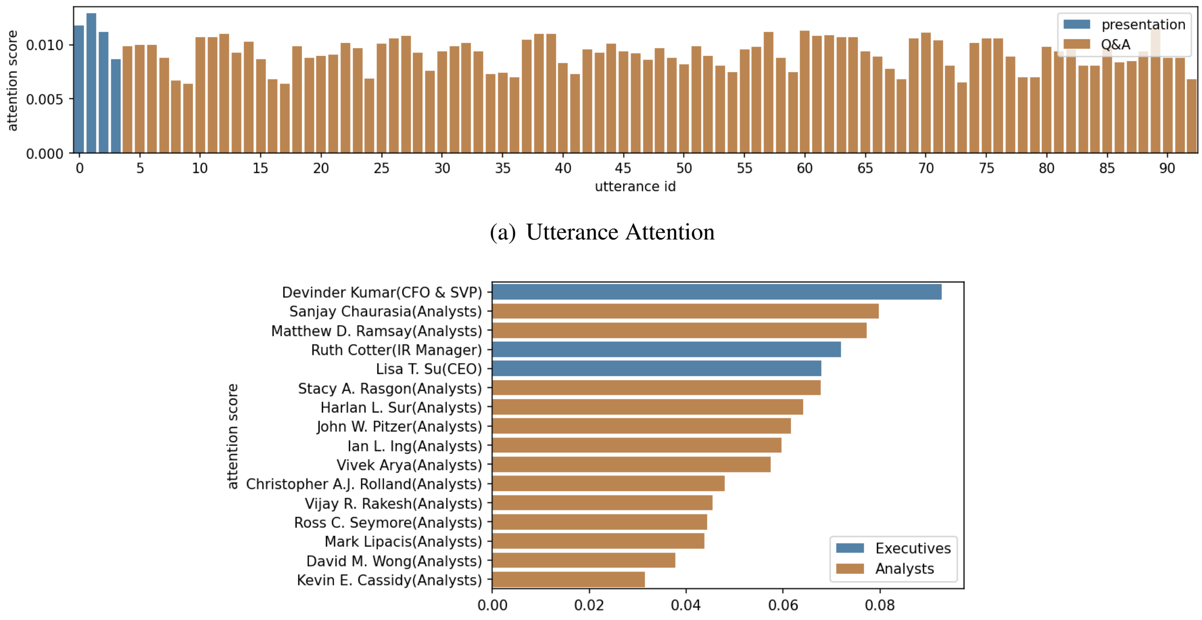
4.4.3 Model Interpretability

Since model interpretability is important in finance and economics, we follow (Wiegreffe and Pinter, 2019) to validate the usefulness of our model's contextual attention layer for model interpretability. Table A1 in Appendix shows the diagnosing results. We also conduct a case study using AMD's 2015 Q2 earnings conference call[5]. We plot the attention scores of utterances and speakers when predicting the stock return volatility in the next three days in Figure 3. Since AMD's financial performance in 2015 Q2 was below the market expectation, its stock price dropped significantly after the release of the earnings conference call. We observe that our model could provide reasonable interpretation ability by its attention mechanisms. As shown in Figure 3 (a), the utterances by managers in the presentation segment are generally more important than those in the Q&A segment of earnings conference call except for certain uninformative messages (e.g., the fourth blue bar in the figure corresponds to the manager's utterance "Thank you, Devinder. Operator, if you could poll the audience for questions, please."). This is perhaps because the managers inside the company usually hold more private information unknown to the outside market participants, and their presentation of the company's potential performance is more informative to the market. As shown in Figure 3 (b), managers and analysts who ask difficult questions that may interest public investors are usually more important than the other analysts. For example, the analysts Sanjay Chaurasia and Matthew D. Ramsay have high attention scores because they dig into the details of the challenges and risks faced by the company. To answer their difficult questions instantly, the manager also provided very informative discussions that may not be disclosed in formal disclosures such as annual and quarterly reports.

5 Conclusion

In this paper, we develop a model called DialogueGAT for financial risk prediction by jointly modeling the utterances and speakers in earnings conference calls. We make a methodological contribution by proposing a new method for constructing the dialogue graph of utterances and speakers and designing two contextual attention mechanisms for both

[5] The transcript of this conference call is available at https://seekingalpha.com/article/3332615-advanced-micro-devices-amd-lisa-t-su-on-q2-2015-results-earnings-call-transcript.



(a) Utterance Attention

(b) Speaker Attention

Figure 3: Attention scores of utterances and speakers generated by our DialogueGAT model in AMD's 2015 Q2 conference call.

utterance and speaker. Our proposed method is general enough to be applied for other dialogue-based prediction tasks such as emotion recognition and sentiment analysis. We also add to the literature by introducing an extended dataset of conference call transcripts with dialogue structure and speaker information. Empirical results on our extended dataset demonstrate the superiority of our proposed model over competitive baselines from the extant literature. We also conduct supplementary analysis for examining the effects of our model's key components, parameter sensitivity, and interpretability.

6 Limitations

This paper is not without limitations. First, we measure the model performance using only the performance metric MSE. Although this metric is widely used for regression models, more performance metrics could be reported for a more thorough evaluation. Second, to compare models on an equal footing, we follow our baselines (e.g., DialogueGCN and MRQA) by only using the pre-trained GloVe model to encode textual utterances. However, it would be interesting to examine whether using more powerful BERT-like models could improve the prediction performance. Third, we run baseline models using their default hyperparameters due to computational constraints.

Acknowledgements

We gratefully acknowledge the support of National Natural Science Foundation of China (Grant #72271160, #71832088, and #71771148).

References

Suman Banerjee and Mitesh M Khapra. 2019. Graph convolutional network with sequential attention for goal-oriented dialogue systems. *Transactions of the Association for Computational Linguistics*, 7:485–500.

Yang Bao and Anindya Datta. 2014. Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, 60(6):1371–1391.

Victor L Bernard and Jacob K Thomas. 1989. Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27:1–36.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graph-structured network for multi-party dialogues. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5010–5016.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 3543–3556.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 272–280.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 3063–3070.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 3538–3545.

Dawn Matsumoto, Maarten Pronk, and Erik Roelofsen. 2011. What makes conference calls useful? the information content of managers' presentations and analysts' discussion sessions. *The Accounting Review*, 86(4):1383–1414.

Ser-Huang Poon and Clive WJ Granger. 2003. Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, 41(2):478–539.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 390–401.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2931–2951.

Christoph Kilian Theil, Samuel Broscheit, and Heiner Stuckenschmidt. 2019. PRoFET: Predicting the risk of firms from event transcripts. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5211–5217.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations (ICLR)*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20.

Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. 2021. Multi-document transformer for personality detection. In *35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 14221–14229.

Zhilin Yang, Zihang Dai, Yiming Yang, J. Carbonell, R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 5754–5764.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1480–1489.

Zhen Ye, Yu Qin, and Wei Xu. 2020. Financial risk prediction with multi-round q&a attention network. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4576–4582.

Appendix

Illustration of Dialogue Graph

Figure A1 presents a more detailed illustration of our constructed dialogue graph. We construct the dialogue graph for each earnings conference call as described in Section 3.2.2. Specifically, we connect each utterance node to its previous and next utterance nodes and its speaker node. If a speaker attends multiple conferences, the speaker's embedding vector (e.g., speaker 1 in the figure) will be shared for capturing the speaker information across different earnings conference calls.

Model Interpretability

Although there is a debate on whether attention can be used to explain the model, the attention mechanism is commonly used as a tool for understanding the model predictions in recent studies (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019). We validate the usefulness of our model's contextual attention layer for model interpretability by following the experiment proposed by (Wiegreffe and Pinter, 2019). Specifically, we diagnose the attention distributions by guiding simpler models in order to examine the prediction power of attention distributions.

To create a diagnostic model, we first remove the components of our model (i.e., Graph Encoder, Speaker Attention, and Utterance Attention) to create a "clean" setting, where the trained parts of the model have no access to neighboring speakers or utterances. Then, we use the following three methods to generate guide weights to get a global feature vector on utterance embedding vectors obtained from TextCNN and speaker embedding vectors obtained from the pre-trained speaker embedding. We impose the guide weights by following the setting in (Wiegreffe and Pinter, 2019):

(1) *Uniform* – we use simple arithmetic mean to get the global feature vector, which represents the situation without attention;

(2) *Trained MLP* – we use an MLP to learn its own attention parameters;

(3) *Base* – we take the weights learned by the base DialogueGAT's attention layer.

Table A1: Diagnosing attention distributions by guiding simpler models.

Guide weights	$\tau = 3$	$\tau = 7$	$\tau = 15$
Uniform	0.5791	0.4347	0.2562
Trained MLP	0.5568	0.4263	0.2501
Base	**0.5392**	**0.4160**	**0.2482**

The diagnosing results are shown in Table A1. As can be seen, the pre-trained scores from our attention model perform better than other guide weights, which means that they are helpful and consistent for model explainability.

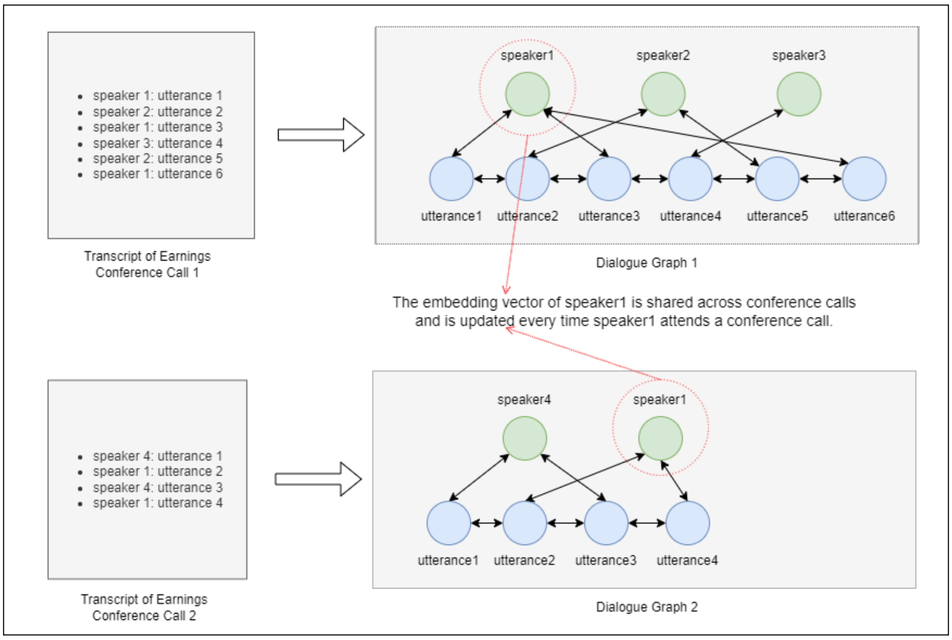


Figure A1: An illustration of our constructed dialogue graph.