UnitedANT: A Multimodal Deep Learning Framework for Predicting Financial Risk from Acoustic, Numeric, and Textual Cues in Earnings Conference Calls

Research-In-Progress Paper

Sen Yan
Shanghai University of Finance and Economics, China
yansen@163.sufe.edu.cn

Yang Bao
Shanghai Jiao Tong University, China
baoyang@sjtu.edu.cn

Hui Fang
Shanghai University of Finance and Economics, China
fang.hui@mail.shufe.edu.cn

Abstract

Earnings conference calls have been recently recognized as significant information events to the market due to its less constrained fashion and direct interaction between managers and analysts. However, it is a non-trivial task to fully exploit the information contained in these conference calls due to its multimodality. To tackle this problem, we develop a general multimodal deep learning framework called UnitedANT (A, N, T stands for Acoustic, Numeric, and Textual information respectively) which could simultaneously leverage acoustic, numeric, and textual information of conference calls for predicting corporate financial risk. Empirical results on a real-world dataset of S&P 500 companies demonstrate the superiority of our proposed method over competitive baselines from the extant literature. Our ablation study presents evidence that all three modalities are useful for financial risk prediction, and the exclusion of any one or two of them will lead to a drop in model performance.

Keywords: *multimodal deep learning, financial risk prediction, earnings conference calls*

Introduction

Earnings conference calls held in conjunction with earnings releases have become an increasingly important form of voluntary corporate disclosures. In these 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 (Matsumoto et al. 2011), these conference calls have been recognized as significant information events to the market. Although researchers and practitioners have recognized their importance for market prediction, it is a non-trivial task to fully exploit the information contained in earnings conference calls due to its multimodality. Specifically, the conference calls contain three types of multimodal information, including the acoustic information (i.e., conference audios), numeric information (e.g., the accompanying financial variables), and textual information (i.e., conference transcripts), but the two types of non-numeric information can not be directly used by conventional machine learning or econometric models. While some previous studies have attempted to use one or two types of multimodal information for market prediction (Kogan et al. 2009; Qin and Yang 2019), there is no existing model that could combine all three modalities of earnings conference calls as far as we know.

To fill the aforementioned research gap, we develop a general multimodal deep learning framework called UnitedANT (A, N, T stands for Acoustic, Numeric, and Textual information respectively; and our framework is called UnitedANT because unity is strength for ants) which could simultaneously leverage acoustic, numeric, and textual information of conference calls for predicting corporate financial risk. Inspired by the recent success of Transformer architecture (Vaswani et al. 2017), we use the self-attention mechanism rather than the conventional recurrent or convolutional

neural networks for modeling the sequential acoustic and textual data. The other feature of our framework is that we use the latest pre-trained language model LongFormer (Beltagy et al. 2020) rather than the commonly used BERT (Devlin et al. 2019) for processing long documents of conference calls since the input sequence length of BERT model is limited to 512 tokens.

To evaluate model performance, we use a real-world dataset of quarterly earnings conference calls of S&P 500 companies. We measure the firms' financial risk by using their stock return volatility - one of the most commonly used measures in prior research (Kogan et al. 2009). We compare our proposed model with two types of baseline models from the extant literature. The first is the SVR (Support Vector Regression) based regression model which is commonly used for text-based risk prediction problems (Frankel et al. 2016; Kogan et al. 2009). Such shallow models typically use the bag-of-words assumption and term weighting methods (e.g., TFIDF) for representing textual documents as vectors, and train a conventional SVR model for prediction. The second type of methods is the state-of-the-art multimodal deep learning models (Qin and Yang 2019) which usually perform better than the traditional shallow machine learning models. Empirical results demonstrate the superiority of our proposed method over those competitive baselines. Our ablation study presents evidence that all three modalities are useful for financial risk prediction, and the exclusion of any one or two of them will lead to a drop in model performance.

This study contributes to a growing body of literature on multimodal-based prediction models. First, we develop a new multimodal deep learning framework that could flexibly combine acoustic, numeric, and textual features for supervised learning tasks. Although our framework is designed for financial risk prediction, it is general enough to be applied for other prediction tasks. Second, we are among the first to access the usefulness of different types of information for predicting corporate financial risk. Our empirical study demonstrates the acoustic, numeric, and textual information of earnings conference calls are all useful for predicting the future stock return volatility.

The remainder of the paper is organized as follows. In the next section, we review two areas of related works. After that, we elaborate our proposed framework and present experimental settings and results. Finally, we provide some concluding remarks.

Literature Study

Our work is related to two areas: (1) financial risk prediction, and (2) multimodal machine learning.

Financial Risk Prediction

Most of the prior studies focus on exploring textual data and numerical financial data for financial risk prediction. They find that textual corporate disclosures are incrementally informative over the accompanying numerical financial data (Kogan et al. 2009; Matsumoto et al. 2011). Two types of textual corporate disclosures are most commonly used for financial risk prediction, including annual reports (Kogan et al. 2009; Rekabsaz et al. 2017; Tsai and Wang 2017; Wang et al. 2013) and transcripts of earnings conference calls (Qin and Yang 2019; Theil et al. 2019; Wang and Hua 2014). For example, Kogan et al. (2009) combine numerical features (i.e., historical volatility) and textual features extracted from annual reports, and these textual features are represented using the bag-of-words assumption and conventional term weighting methods such as TF, TFIDF, and LOG1P. Some studies (Qin and Yang 2019; Rekabsaz et al. 2017; Theil et al. 2019) also use the recent word embedding models, e.g., Word2Vec and FastText, to learn the distributional representation of textual words and documents. While textual information is widely used for financial risk prediction, acoustic information is seldom used. To the best of our knowledge, (Qin and Yang 2019) is the only existing work that attempts to predict financial risk by using acoustic information in addition to the commonly used textual information of earnings conference calls. But it cannot handle the numeric financial features such as historical stock return volatility.

In terms of the machine learning algorithms, support vector regression (SVR) is used in most existing works for financial risk prediction (Kogan et al. 2009; Rekabsaz et al. 2017; Tsai and Wang 2017; Wang et al. 2013), while deep learning algorithms are adopted in some recent studies (Qin and Yang 2019; Theil et al. 2019).

Multimodal Machine Learning

Multimodal machine learning (Baltrušaitis et al. 2018) has attracted much attention in recent years due to its ability to extract useful information from multimodal data (e.g., text, image, and audio). It has been widely used in many real-world applications, such as the early audio-visual speech recognition (Yuhas et al. 1989), and the recent image

captioning (Hodosh et al. 2013) and sentiment analysis (Poria et al. 2019; Zadeh et al. 2017) tasks.

Recently, two multimodal deep learning models have been proposed for the task of financial risk prediction, including the PROFET (Theil et al. 2019) and MDRM (Qin and Yang 2019) models. The first PROFET model uses the bidirectional LSTM to encode the textual information of earnings conference calls, and could combine the textual features with the numeric financial features. The second MDRM model combines verbal and vocal features using a deep learning framework for financial risk prediction. Our proposed framework UnitedANT differs from them in two important ways. First, our framework is able to combine acoustic, numeric, and textual information while those two models can only combine two modalities. Second, we use the recent attention mechanism for modeling the sequential acoustic and textual data while those two models only use the conventional recurrent neural networks.

Proposed Model

We develop a general multimodal deep learning framework called UnitedANT that can flexibly combine acoustic, numeric, and textual features for the downstream regression or classification tasks. As shown in Figure 1, our proposed framework is composed of three encoders (i.e., *acoustic encoder*, *numeric encoder*, and *textual encoder*) for learning informative features from the corresponding modality, and a *fusion module* for combining the three types of features for financial risk prediction. We elaborate on the details of the four modules below.

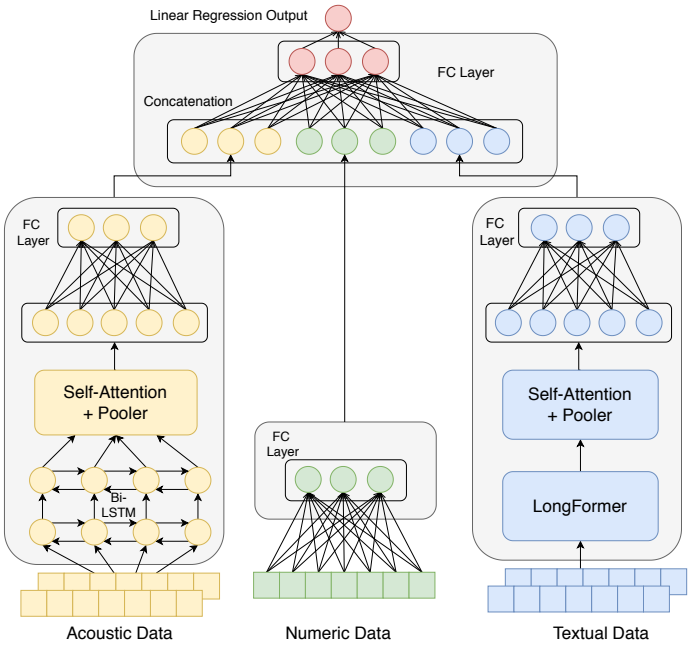

Figure 1. Our proposed multimodal deep learning framework UnitedANT.

Acoustic Encoder. As shown in the left part of Figure 1, this module is used to learn the high-level features of acoustic data via deep learning. Specifically, we first process the raw acoustic data by splitting each audio file of the conference call into fragments (i.e., sentences), and extracting 6,373 acoustic features (the size of this feature vector is denoted as d_a) from each fragment using the open-source software openSMILE with the configuration file IS13-ComParE (Eyben et al. 2010). For each audio file of conference call, the sequence of the d_a-dimensional feature vectors of its fragments are fed into a two-layer Bi-LSTM (Bidirectional LSTM) (Hochreiter and Schmidhuber 1997) model for capturing the sequential patterns of acoustic data. The output of the Bi-LSTM model is the hidden states $\mathbf{H_A} \in \mathbb{R}^{l_a \times 2d_{a\_h}}$, where l_a is the number of audio fragments and d_{a_h} is the dimension of Bi-LSTM's hidden layer. To further capture the sequential information of acoustic fragments, we use a self-attention layer (Vaswani et al. 2017) to encode H_A as:

$$\mathbf{A_A} = \mathbf{softmax}(\frac{\mathbf{Q_A K_A^T}}{\sqrt{\mathbf{2d_{a\_h}}}})\mathbf{V_A} \tag{1}$$

where $\mathbf{A_A} \in \mathbb{R}^{l_a \times 2d_{a\_h}}$, $\mathbf{Q_A} = \mathbf{W_{Q_A} H_A} + \mathbf{b_{Q_A}}$, $\mathbf{K_A} = \mathbf{W_{K_A} H_A} + \mathbf{b_{K_A}}$, $\mathbf{V_A} = \mathbf{W_{V_A} H_A} + \mathbf{b_{V_A}}$, $\mathbf{Q_A}$, $\mathbf{K_A}$, and $\mathbf{V_A} \in \mathbb{R}^{l_a \times 2d_{a\_h}}$, $\mathbf{W}_*$ and $\mathbf{b}_*$ are the corresponding parameters of weights and bias in a self-attention layer (Vaswani et al.

2017). Finally, we apply a dense layer on the average pooling of $\mathbf{A_A}$ to represent acoustic data as embedding vectors:

$$\mathbf{E_A} = \mathbf{W_{E_A}}\mathbf{avg\_pool}(\mathbf{A_A}) + \mathbf{b_{E_A}} \tag{2}$$

where $\mathbf{E_A} \in \mathbb{R}^{d_{a\_e}}$ is the final output of *acoustic encoder*, d_{a_e} is the acoustic embedding dimension, $\mathbf{W_{E_A}}$ and $\mathbf{b_{E_A}}$ are the weight matrix and bias vector, and $avg\_pool(*)$ computes the mean of all the row vectors in $\mathbf{A_A}$.

Numeric Encoder. As shown in the middle part of Figure 1, this module is designed to encode the numerical feature vectors $\mathbf{N} \in \mathbb{R}^{l_n}$ (l_n is the number of numerical features) as the hidden states of a fully-connected dense layer $\mathbf{E_N} \in \mathbb{R}^{d_{n\_e}}$ (d_{n_e} is the dimension of hidden states). Specifically, we directly use the dense layer to align the dimension of numerical feature vectors with the other two modalities' feature vectors:

$$\mathbf{E_N} = \mathbf{W_{E_N}}\mathbf{N} + \mathbf{b_{E_N}} \tag{3}$$

where $\mathbf{W_{E_N}}$ and $\mathbf{b_{E_N}}$ are the parameters of weight matrix and bias vector, and E_N is the final output of *numeric encoder*.

Textual Encoder. As shown in the right part of Figure 1, this module is used to learn the feature representation of each textual conference call transcript. To this end, we propose to leverage the pre-trained contextual embedding language model which has caused a stir by its state-of-the-art performance in a variety of natural language processing tasks. Since the conference call transcripts are usually long documents, we cannot directly utilize the commonly used BERT language model (Devlin et al. 2019) because its input sequence length is limited to 512 tokens. To tackle this problem, we propose to use LongFormer (Beltagy et al. 2020) – a recent state-of-the-art pre-trained language model that can handle long documents – to encode the long conference call transcripts $\mathbf{T} \in \mathbb{R}^{l_t}$ (l_t is the number of tokens of a transcript, and each element in $\mathbf{T}$ is the corresponding token id):

$$\mathbf{H_T} = \mathbf{LongFormer}(\mathbf{T}) \tag{4}$$

where $\mathbf{H_T} \in \mathbb{R}^{l_t \times d_{t\_h}}$ is the output of the LongFormer, d_{t_h} is the dimension of hidden states in LongFormer. To further capture the sequential information of word tokens, we use a self-attention layer to encode $\mathbf{H_T}$ as:

$$\mathbf{A_T} = \mathbf{softmax}(\frac{\mathbf{Q_T}\mathbf{K_T^T}}{\sqrt{\mathbf{d_t}}})\mathbf{V_T} \tag{5}$$

where $\mathbf{A_T} \in \mathbb{R}^{l_t \times d_{t\_h}}$ is the output of the layer, $\mathbf{Q_T} = \mathbf{W_{Q_T}}\mathbf{H_T} + \mathbf{b_{Q_T}}$, $\mathbf{K_T} = \mathbf{W_{K_T}}\mathbf{H_T} + \mathbf{b_{K_T}}$, $\mathbf{V_T} = \mathbf{W_{V_T}}\mathbf{H_T} + \mathbf{b_{V_T}}$, $\mathbf{Q_T}$, $\mathbf{K_T}$, and $\mathbf{V_T} \in \mathbb{R}^{l_t \times d_{t\_h}}$, $\mathbf{W}_*$ and $\mathbf{b}_*$ are the corresponding parameters of weights and bias in a self-attention layer. Finally, we apply a dense layer on the average pooling of $\mathbf{A_T}$ to represent textual transcripts as embedding vectors:

$$\mathbf{E_T} = \mathbf{W_{E_T}}\mathbf{avg\_pool}(\mathbf{A_T}) + \mathbf{b_{E_T}} \tag{6}$$

where $\mathbf{E_T} \in \mathbb{R}^{d_{t\_e}}$ is the final output of *textual encoder*, d_{t_e} is textual embedding dimension, $\mathbf{W_{E_T}}$ and $\mathbf{b_{E_T}}$ are the weight matrix and bias vector, and $avg\_pool(*)$ computes the mean of all the row vectors in $\mathbf{A_T}$.

Fusion Module. As shown at the top of Figure 1, this module is designed to fuse the three modalities for the downstream prediction tasks (regression task in our case). Specifically, we concatenate the embedding vectors generated by the three encoders to get the multi-modal feature representation $\mathbf{E} = [\mathbf{E_A}; \mathbf{E_N}; \mathbf{E_T}] \in \mathbb{R}^{d_{a\_e}+d_{n\_e}+d_{t\_e}}$ of conference calls, and then feed them into a two-layer MLP (Multi-Layer Perception) for predicting the stock return volatility $\hat{y}$:

$$\hat{\mathbf{y}} = \mathbf{w_y}\mathbf{ReLU}(\mathbf{W_h}\mathbf{E} + \mathbf{b_h}) + \mathbf{b_y} \tag{7}$$

where $\mathbf{W_h}$, $\mathbf{b_h}$, $\mathbf{w_y}$, $\mathbf{b_y}$ are the weight matrix and bias vector of the MLP.

Experiments

In this section, we first formulate our prediction problem and then present the experimental settings and results.

Financial Risk Prediction Task. We measure the firms' financial risk using stock return volatility - one of the most commonly used measures in prior research (Kogan et al. 2009). More formally, the stock return volatility $v_{[t,t+\tau]}$ over the time period from day t to day $t+\tau$ is defined as:

$$v_{[t,t+\tau]} = log(\sqrt{\frac{\sum_{i=0}^{\tau}(r_{t+i} - \bar{r})^2}{\tau}}) \tag{8}$$

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. To examine both short-term and long-term market reactions, we follow (Qin and Yang 2019) by setting the time window size τ to the values 3, 7, 15, and 30. We formulate our financial risk prediction problem as a supervised regression task, and propose to leverage multimodal information of earnings conference calls for predicting stock return volatility. More specifically, given the acoustic and textual data of a firm's conference call issued on the day t, and the numeric data of historical stock return volatility over the period from day $t-\tau$ to day t, we aim to predict the firm's future stock return volatility over the period from day t to day $t+\tau$.

Experimental Settings. To evaluate model performance, we use a real-world multimodal dataset of quarterly earnings conference calls of S&P500 companies in 2017 which is collected by (Qin and Yang 2019). To preserve the temporal nature of risk prediction, we split the data sample by time in a ratio of 8:2 (the cut-off date is October 21, 2017) - the first 80% samples are used for training and the last 20% samples are used for testing. We train all the models on a single Nvidia Tesla V100 GPU. Our UnitedANT models and the baseline model MDRM are trained for 30 epochs. We use MSE (Mean Squared Error) as loss function, and use Adam algorithm for optimizing the loss. The learning rate is initialized as 2e-5 for the text encoder, and 1e-2 for the other parts of the model. It is worth mentioning that we freeze the parameters of the pre-trained language model LongFormer because this could result in better performance.

Main Results. To measure model performance for predicting the future stock return volatility, we use the commonly used performance metric MSE (Mean Squared Error). We compare our proposed UnitedANT models with competitive baselines from the extant literature (Kogan et al. 2009; Qin and Yang 2019), including: (1) the naive baseline v_{past} which directly uses the historical volatility of previous τ days to predict the future volatility of next τ days, (2) the SVR model with the linear kernel using v_{past} as feature, (3) the SVR model with the linear kernel using the traditional tf-idf textual features, and (4) the state-of-the-art MDRM (Multimodal Deep Regression Model) using acoustic and textual features. The main results of performance comparisons are shown in Table 1, and our main findings are summarized as follows. First, our proposed UnitedANT using all three modalities performs significantly better than all the four baseline models (as shown in the left columns of Table 1) for both short-term and long-term prediction of stock return volatility. Among all models, our proposed UnitedANT (a+n+t) achieves the lowest error rate of 0.7664, 0.2126, 0.1443, and 0.0847 for the window sizes $\tau = 3, 7, 15, 30$, respectively. Second, our ability study (as shown in the right columns of Table 1) demonstrates that all the three modalities (a-acoustic, n-numeric, t-textual) are useful for predicting stock return volatility, and the exclusion of any one or two of them (e.g., a, n, t, n+t, a+t, a+n) will lead to a drop in model performance. Among all modalities, the numeric feature (i.e., the historical stock return volatility) is more informative than the textual and acoustic features.

Table 1. Model performance in terms of MSE by varying windows size τ.

Method	Naive	SVR	SVR	MDRM	UnitedANT						
Feature	v_{past}	v_{past}	tf-idf	a+t	a	n	t	n+t	a+t	a+n	a+n+t
$\tau = 3$	1.3428	0.9231	1.0753	1.0585	1.2087	0.8665	1.0878	0.8266	1.0188	0.8398	**0.7664**
$\tau = 7$	0.3745	0.2630	0.4256	0.3762	0.4147	0.3126	0.3696	0.2455	0.3331	0.2451	**0.2126**
$\tau = 15$	0.2552	0.1876	0.3065	0.2831	0.2932	0.2078	0.3206	0.1538	0.2634	0.1631	**0.1443**
$\tau = 30$	0.1515	0.1176	0.2092	0.1874	0.2278	0.1081	0.2019	0.0894	0.1960	0.0985	**0.0847**

Conclusion

In this paper, we develop a general multimodal deep learning framework UnitedANT which could simultaneously leverage acoustic, numeric, and textual information of conference calls for predicting corporate financial risk. Empirical results on a real-world dataset of S&P 500 companies demonstrate the superiority of our proposed method over competitive baselines and the usefulness of all three modalities for predicting corporate financial risk. In the future, we will attempt to improve our framework by designing more effective modules for the fusion and alignment of multimodal data and making each module more interpretable.

Acknowledgement

We gratefully acknowledge the support of the National Natural Science Foundation of China (Grant No. 71601116, 71601104, 71771148, 71832088), Shanghai Pujiang Program (Grant No. 16PJC045), and NVIDIA Corporation with the donation of the GPU used for this research.

References

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (41:2), pp. 423–443.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). "Longformer: The long-document transformer," *ArXiv* (abs/2004.05150).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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-HLT),* pp. 4171–4186.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 18th ACM International Conference on Multimedia,* pp. 1459–1462.

Frankel, R., Jennings, J., and Lee, J. (2016). "Using unstructured and qualitative disclosures to explain accruals," *Journal of Accounting and Economics* (62:2-3), pp. 209–227.

Hochreiter, S. and Schmidhuber, J. (1997). "Long short-term memory," *Neural Computation* (9:8), pp. 1735–1780.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research* (47), pp. 853–899.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). "Predicting risk from financial reports with regression," in *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT),* pp. 272–280.

Matsumoto, D., Pronk, M., and Roelofsen, E. (2011). "What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions," *The Accounting Review* (86:4), pp. 1383–1414.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). "MELD: A multimodal multi-Party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 527–536.

Qin, Y. and Yang, Y. (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),* pp. 390–401.

Rekabsaz, N., Lupu, M., Baklanov, A., Dür, A., Andersson, L., and Hanbury, A. (2017). "Volatility prediction using financial disclosures sentiments with Word embedding-based IR models," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 1712–1721.

Theil, C. K., Broscheit, S., and Stuckenschmidt, H. (2019). "PRoFET: Predicting the risk of firms from event transcripts," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI),* AAAI Press, pp. 5211–5217.

Tsai, M.-F. and Wang, C.-J. (2017). "On the risk prediction and analysis of soft information in finance reports," *European Journal of Operational Research* (257:1), pp. 243–250.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is all you need," in *Proceedings of 31st Conference on Advances in Neural Information Processing Systems (NIPS),* pp. 5998–6008.

Wang, C.-J., Tsai, M.-F., Liu, T., and Chang, C.-T. (2013). "Financial sentiment analysis for risk prediction," in *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP),* pp. 802–808.

Wang, W. Y. and Hua, Z. (2014). "A semiparametric gaussian copula regression model for predicting financial risks from earnings calls," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 1155–1165.

Yuhas, B. P., Goldstein, M. H., and Sejnowski, T. J. (1989). "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine* (27:11), pp. 65–71.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* pp. 1103–1114.