Less Is More: Volatility Forecasting with Contrastive Representation Learning (Student Abstract)

Yanlong Huang[*1], **Wenxin Tai**[*1], **Ting Zhong**[1†], **Kunpeng Zhang** [2]

[1] University of Electronic Science and Technology
[2] University of Maryland, College park
hyloong77@gmail.com, amperetai@gmail.com, zhongting@uestc.edu.cn, kpzhang@umd.edu

Abstract

Earnings conference calls are indicative information events for volatility forecasting, which is essential for financial risk management and asset pricing. Although recent volatility forecasting models have explored the textual content of conference calls for prediction, they suffer from modeling the long-text and representing the risk-relevant information. This work proposes to identify key sentences for robust and interpretable transcript representation learning based on the cognitive theory. Specifically, we introduce TextRank to find key sentences and leverage attention mechanism to screen out the candidates by modeling the semantic correlations. Upon on the structural information of earning conference calls, we propose a structure-based contrastive learning method to facilitate the effective transcript representation. Empirical results on the benchmark dataset demonstrate the superiority of our model over competitive baselines in volatility forecasting.

Introduction

Volatility is a statistical measure of variation in a stock's returns over time, which plays an important role in many financial market-related tasks. Motivated by the practical findings from financial economics, there is a burgeoning body of studies that aims to predict a company's risk from its earnings conference call transcripts (Theil, Broscheit, and Stuckenschmidt 2019; Ye, Qin, and Xu 2020). However, one problem that is often ignored by prior studies is that earnings conference call transcripts are long text. An earnings conference call usually lasts for about one hour and its transcript contains around 7-8 thousands word tokens. Therefore, directly representing the entire long transcript as one document and feeding it into a deep model (such as LSTM or BERT) inevitably disregards important risk-relevant information.

According to the cognitive theory (Ding et al. 2020) that a few key sentences in the text store sufficient and necessary information to fulfill most NLP tasks, we try to identify key sentences for robust transcript representation learning. In particular, we calculate the occurrence of words between sentences and utilize TextRank (Mihalcea and Tarau 2004) to roughly find the key sentences. Next, we propose a self-attention module to select the candidates by modeling the semantic correlations between sentences. To learn more distinguishable representations that make our method more interpretable, we exploit the structural information of transcripts and design a contrastive learning paradigm to improve the efficacy of transcript representations. Experiments conducted on real-world datasets show that our model outperforms the strong baselines.

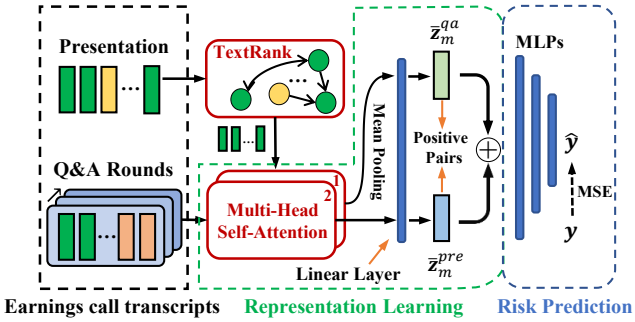

Figure 1: Overview of the proposed model architecture.

Method

We follow prev (Ye, Qin, and Xu 2020) to measure the financial risk using stock return volatility. Formally, we define the stock return volatility as $v_{[t,t+T]} = ln\sqrt{\sum_{i=0}^{T}\left(r_i - \bar{r}\right)^2 / T}$, where T is the size of time window for calculating volatility. r_i denotes the change rate on the i-th day and $\bar{r}$ represents the average of the change rate in this period.

We first search the key sentences using TextRank, a graph-based ranking algorithm which has been successfully applied in automated text summarization and phrase ranking (Ding et al. 2020). In particular, a graph is constructed where the vertices of the graph represent each sentence in a document and the edges between sentences are based on content overlap – i.e., by calculating the number of words that two sentences have in common. After the ranking algorithm, the top-ranked sentences are selected as key sentences. Then, we propose a multi-head self-attention module to screen out the candidates by modeling the semantic correlations.

[*]These authors contributed equally.
[†]Corresponding author.

As the recording of the conference, the transcript restores the *Presentation* addressed by firms' managers and the *Q&A* interaction between senior managers and analysts chronologically. This structure may contain valuable facts since the *Q&A* reveals analysts' concerns referring to additional information that the management team does not disclose in the *Presentation* and managers' reactions to these questions.

Based on this observation, we propose a contrastive paradigm to learn more distinguishable representations. Different from existing contrastive NLP methods such as (Gao, Yao, and Chen 2021) that create positive pairs by generating two views of the same instance via data augmentation, we generate the positive pairs by exploiting the structural relationships. For transcript m, the contrastive loss is:

$$\mathcal{L}_m^{CL} = -\log \frac{\exp\left(\text{sim}\left(\overline{\mathbf{z}}_m^{qa}, \overline{\mathbf{z}}_m^{pre}\right)/\tau\right)}{\sum_{n=1}^{B} \exp\left(\text{sim}\left(\overline{\mathbf{z}}_m^{qa}, \overline{\mathbf{z}}_n^{pre}\right)/\tau\right)\Big)\Big)}, \quad (1)$$

where $\text{sim}(\cdot)$ is the dot product similarity, B is the batch size, and τ is a temperature hyper-parameter. $\overline{\mathbf{z}}_m^{pre}$ and $\overline{\mathbf{z}}_m^{qa}$ denote the representations of *Presentation* and *Q&A* respectively – here mean pooling is used to generate the representation. Finally, we forecast the risk via a three-layer MLP:

$$\hat{y}_m = \text{MLPs}(\overline{\mathbf{z}}_m^{pre} \oplus \overline{\mathbf{z}}_m^{qa}), \quad (2)$$

where $\oplus$ denotes concatenation operation. The optimization objective of our model consists of two parts:

$$\mathcal{L}_m = \mathcal{L}_m^p + \alpha \cdot \mathcal{L}_m^{CL}, \quad (3)$$

where the risk prediction error $\mathcal{L}_m^p$ is based on mean squared error (MSE), and α is a hyper-parameter.

Experimental Results

Datasets. We collect a large-scale earnings transcripts dataset of U.S. firms in four fiscal years (2015-2018). The transcripts are available at SeekingAlaph website[1] and databases such as Thomson Reuters StreetEvents[2]. Note that the number of tokens in a transcript is quite large. For example, the average number of tokens in the *Presentation* is over 3,000, and that in the *Q&A* is over 4,000. It highlights the long-text challenge we aim to address. We choose data from 2015 and 2016 to train the model. Data from 2017 and 2018 are respectively served as the validation and testing sets. This chronically divided data setup can prevent look-ahead bias (Theil, Broscheit, and Stuckenschmidt 2019).

Experimental settings. The top-ranked thirty percent of sentences in the *Presentation* will be served as key sentences. We set the temperature τ to 0.1 and the coefficient α to 1. We use a simple 3-layer MLP to forecast volatility. We consider the regression metrics MSE and Mean Absolute Error (MAE) for evaluation. BERT-SVR (Devlin et al. 2019), ProFET (Theil, Broscheit, and Stuckenschmidt 2019), MRQA (Ye, Qin, and Xu 2020), and SimCSE (Gao, Yao, and Chen 2021) are used as baselines for comparisons.

Performance Comparison. Table 1 summarizes the performance comparisons between our method and baseline approaches, which demonstrates that our model achieves the best performance on both metrics at all forecasting horizons. In particular, MRQA, which uses a reinforced sentence selector to choose important sentences, achieves better result compared to other baselines, verifying the rationality of using key-sentence as the representation of the long-text. In comparison to MRQA, our method introduce contrastive learning to improve the efficacy of transcript representation learning, and achieves significantly better performance.

Method	**MSE**			**MAE**		
	10d	20d	60d	10d	20d	60d
BERT-SVR	0.3070	0.2340	0.1826	0.4295	0.3728	0.3207
ProFET	0.3343	0.2585	0.2130	0.4485	0.3927	0.3574
MRQA	0.3025	0.2311	0.1792	0.4259	0.3699	0.3170
SimCSE	0.3073	0.2330	0.1768	0.4299	0.3739	0.3163
Ours	**0.2964**	**0.2282**	**0.1748**	**0.4228**	**0.3675**	**0.3130**

Table 1: Performance comparisons.

By comparing the non-contrastive method (ProFET) with the contrastive-based methods (SimCSE and ours)[3], we find that both contrastive-based methods outperform the non-contrastive-based method. The result further confirms our claim: instead of directly optimizing the risk prediction objective on the training set, contrastive learning can guide the risk prediction model to learn more risk-relevant representations. Moreover, our approach significantly outperforms SimCSE, indicating that our domain-specific contrastive objective is more effective than sentence-augmentation-based contrastive objective. Now we can conclude that the volatility forecasting error can be significantly reduced by our key sentence extraction and contrastive representation learning.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (Grant No.62072077 and No.62176043) and Natural Science Foundation of Sichuan Province (Grant No. 2022NSFSC0505).

References

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 4171–4186. ACL.

Ding, M.; Zhou, C.; Yang, H.; and Tang, J. 2020. Cogltx: Applying bert to long texts. *NIPS*, 33: 12792–12804.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*, 6894–6910. Association for Computational Linguistics.

Mihalcea, R.; and Tarau, P. 2004. Textrank: Bringing order into text. In *EMNLP*, 404–411.

Theil, C. K.; Broscheit, S.; and Stuckenschmidt, H. 2019. PRoFET: Predicting the Risk of Firms from Event Transcripts. In *IJCAI*, 5211–5217.

Ye, Z.; Qin, Y.; and Xu, W. 2020. Financial Risk Prediction with Multi-Round Q&A Attention Network. In *IJCAI*, 4576–4582.

[1]https://seekingalpha.com

[2]https://www.streetevents.com

[3]This comparison is based on the fact that the three methods follow the same architecture.