Enhancing Volatility Forecasting in Financial Markets: A General Numeral Attachment Dataset for Understanding Earnings Calls

Ming-Xuan Shi,[1] **Chung-Chi Chen,**[2] **Hen-Hsen Huang,**[3] **Hsin-Hsi Chen** [1]
[1] Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[2] AIST, Japan
[3] Institute of Information Science, Academia Sinica, Taiwan
b08902105@csie.ntu.edu.tw, c.c.chen@acm.org,
hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

Abstract

Volatility, a crucial statistical measure in the financial market, serves as an indicator of financial instrument risk. Accurate volatility capture aids in predicting stock movements and is valuable in derivative trading, such as options trading. While recent research focuses on volatility forecasting using earnings call transcriptions, most approaches rely on end-to-end models that directly process textual or vocal data. However, limited efforts have been made to simulate the reading and comprehension processes of financial professionals, thereby enhancing the capabilities of language models. To address this gap, we propose a general numeral attachment dataset designed to train language models to understand earnings calls with the expertise of professionals. Additionally, we introduce a pre-training process that improves the semantic understanding of earnings calls. Experimental results demonstrate that our pre-trained language model enhances the accuracy of 3-day volatility forecasting.

1 Introduction

A key element in understanding financial narratives is pinpointing the specific target linked with every numeral, especially considering the prevalent use of numerals in these texts (Chen et al., 2021a). The notion of "numeral attachment" was first introduced by Chen et al. (2019). This concept seeks to clarify the connection between numerals and particular stocks. Yet, this methodology was specifically devised for a financial social media setting, making it less adaptable to more formal, general documentation. Addressing this gap, we put forth a broader numeral attachment framework, free from such specific constraints. Instead of developing an entirely new dataset, we amplify the existing ECNum dataset (Chen et al., 2021b) with augmented annotations. Consider the nuanced difference between "revenue decreased 20%" and "revenue decreased 2%" — two statements with varying implications for investors. Furthermore, the distinction between "revenue decreased 20%" and "cost decreased 20%", where the numeral remains the same but the narrative diverges, exemplifies the essence of our endeavor. Our central goal is to refine the model's precision in recognizing and distinguishing situations inherent to general numeral attachment.

When making investment decisions, professional investors carefully consider financial risk. Estimating the risk associated with financial instruments is crucial due to the inherent trade-off between potential returns and risks involved. Volatility, which quantifies financial risk as the second-moment measure of price return, serves as a widely used indicator in this regard. Although previous studies have explored various models and data sources to improve volatility forecasting, they have largely neglected enhancing the semantic capabilities of language models for this task. In this paper, we introduce a novel approach that simulates the reading process of financial professionals, aiming to enhance the accuracy of volatility forecasting.

Earnings calls, which involve teleconferences among managers and investors to discuss company operations, have garnered significant attention in recent research. While prior studies have primarily focused on constructing multimodal models for earnings calls, they have often overlooked the finer semantic aspects. In this work, we demonstrate the value of understanding general numeral attachment in the context of earnings calls, particularly for improving volatility forecasting—an important downstream task in financial analysis.

Our contributions are threefold:

1. We introduce the novel task of general numeral attachment, addressing a previously overlooked issue.

2. We provide additional annotations on the publicly-available ECNum dataset.[1]

[1] http://gen-numattach.nlpfin.com/

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 37–42
November 1–4, 2023.

3. We propose an approach to enhance the language model's understanding of numerals, leading to improved performance in 3-day volatility forecasting.

2 Related Work

Language model pretraining with specialized semantic understanding tasks has demonstrated its effectiveness in improving model performance. Notably, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a prominent example. Numerous subsequent studies have utilized BERT or similar Transformer-based architectures for pretraining or fine-tuning language models in various domains. For instance, Clinical BERT (Alsentzer et al., 2019) is pretrained with clinical corpora, while SciBERT (Beltagy et al., 2019) is fine-tuned with scientific publications. FinBERT (Araci, 2019) is specifically pretrained with financial corpus, while BioBERT (Lee et al., 2020), mBERT (Devlin et al., 2019), and patentBERT (Lee and Hsiang, 2019) are fine-tuned with biomedical, multilingual, and patent corpora, respectively.

Despite the availability of domain-specific BERT models, none of them have fine-tuned the BERT model with numeral-related tasks to enhance models' numeracy. In this paper, we propose a novel pretraining task focused on understanding the numerals in earnings calls at a more granular level. We release the associated dataset and pretrained language model, referred to as NumBERT, for future research and applications.

3 Dataset

3.1 Annotation Process

We utilize the ECNum dataset (Chen et al., 2021b), which consists of 9,034 annotated instances from earnings calls. Each instance is associated with three annotations related to numeral information, including the target numeral category label and two domain-specific annotations. In this study, we extend the dataset by providing additional general numeral attachment labels for the ECNum instances.

Each instance was assigned to three annotators. We have regular meetings with all annotators to resolve the inconsistencies. Over a span of three weeks, they were diligently coached to adhere to the same guidelines. Regular discussions were held to clarify doubts and resolve ambiguities, culminating in a strong alignment in their annotations.

	Earnings call		Investor's report	
Rank	Entity	Freq.	Entity	Freq.
1	revenue	767	revenue	855
2	Q	371	EPS	481
3	sales	255	gross margin	326
4	EPS	221	profit	275
5	earnings	165	operating margin	115
6	years	154	price target	108
7	free cash flow	110	operating profit	59

Table 1: Numeral attachment in different narratives.

During the annotation process, annotators with financial backgrounds were presented with a target numeral and a paragraph from an earnings call. Their task was to select the attached entity, such as a named entity or accounting account, that is relevant to the given numeral within the paragraph. The annotators exhibited high consistency in their annotations, with approximately 78.60% of instances receiving fully consistent annotations, 7.13% of instances having two different annotations, and only 14.26% of instances showing completely different annotations. After resolving inconsistencies and removing instances without attached entities, the dataset for the general numeral attachment task consists of 6,735 instances from ECNum.

3.2 Annotations Analysis

Table 1 presents the frequently attached entities observed in the annotations. To facilitate a comprehensive comparison of entity usage, particularly accounting accounts, in both managers' and investors' narratives, we include additional annotations from the NumClaim dataset (Chen et al., 2020), which comprises professional analysts' reports.

Based on the statistics depicted in Table 1, several key findings emerge. Firstly, managers predominantly report operational data, encompassing revenue, sales, EPS, and earnings. Secondly, investors exhibit interest not only in quantitative operational results such as revenue and EPS, but also in accounting ratios like gross margins and operating margins. Thirdly, while managers seldom mention the stock price, investors frequently engage in discussions regarding it, including price targets.

4 Implementation

4.1 Research Questions

In this paper, we investigate the following two research questions:

(RQ1): How does pre-training with the proposed general numeral attachment task impact the perfor-

mance of volatility forecasting?

(RQ2): Which strategy, discrete or continuous, yields better results with the proposed model?

We start by fine-tuning the RoBERTa language model (Liu et al., 2019) with the proposed general numeral attachment task, resulting in the NumBERT model.

To address **(RQ1)**, we employ NumBERT for the volatility forecasting task. Specifically, we train separate models for the 3-day, 7-day, 15-day, and 30-day volatility forecasting tasks.

To explore **(RQ2)**, we develop both a discrete forecasting model and a continuous forecasting model based on the proposed GNA-Vol architecture. The discrete strategy involves simultaneous predictions for the 3-day, 7-day, 15-day, and 30-day volatility forecasting tasks, utilizing a multi-task model. On the other hand, the continuous strategy treats volatility forecasting as a sequential prediction task, making predictions for volatility from $t+1$ to $t+T$ days. Here, we set T to 30. We consider the continuous strategy due to the well-known phenomenon of volatility clustering in the financial market, where large (small) volatility tends to be followed by large (small) volatility. We hypothesize that the continuous strategy can capture this pattern and further improve the performance of volatility forecasting.

4.2 Model Architecture

In the general numeral attachment task, an input paragraph $P \in \{P_1, P_2, \ldots, P_M\}$ is tokenized into words using the RoBERTa tokenizer: $P = [x_1, x_2, \ldots, x_N]$, where x_i represents the i^{th} word and N is the maximum number of words in any paragraph. To ensure the target numeral is not split by the tokenizer, we replace it with the `[MASK]` token. The hidden state of the i^{th} word, denoted as h_i, is obtained from the last layer of RoBERTa, and h_{mask} represents the hidden state of the mask token. Start and end vectors S and E are defined. To incorporate the target numeral information and generate the probability distribution of the start of the answer span, we concatenate h_i and h_{mask} to form a new vector c_i. The probability p_i of word i being the start of the answer span is computed as the softmax of the dot product between c_i and S: $p_i = \frac{e^{S \cdot c_i}}{\sum_j e^{S \cdot c_j}}$. The fine-tuning loss function is cross-entropy. We perform fine-tuning for ten epochs using a learning rate of 1e-5 and a batch size of 4.

	Validation Data	Test Data
FinBERT	86.24%	87.28%
LinkBERT-base	89.65%	87.71%
RoBERTa-base	89.40%	88.01%
LinkBERT-large	88.92%	88.74%
RoBERTa-large	90.99%	90.02%

Table 2: Experimental results on general numeral attachment task.

We employ the trained language model to encode earnings call transcriptions for volatility forecasting. We introduce a matrix $T = [h_1, h_2, \ldots, h_N]$, where h_i represents the hidden state of the i^{th} `[MASK]` token in the input transcription, and N is the maximum number of numerals in any transcription. To leverage multiple numeral information for volatility prediction, we construct a transformer-based model followed by a fully connected layer. The transformer model requires three inputs: Q, K, V, which are obtained by linear transformations of T: $Q = T \times A$ and $K = V = T \times B$. In the case of continuous volatility forecasting, we replace the fully connected layer with an LSTM.

5 Experiments

5.1 Language Model Selection

In this section, we discuss our choice of using RoBERTa as the base language model instead of the domain-specific language model, FinBERT (Araci, 2019). Our selection is based on the performance of the general numeral attachment task. We divided the instances into training, validation, and test sets, with proportions of 70%, 10%, and 20% respectively, for the general numeral attachment task. The evaluation metric for the span identification task is F-score. Table 2 presents the performance of different language models on the general numeral attachment task. We observe that RoBERTa performs well in our proposed pre-training task. Additionally, we experiment with LinkBERT (Yasunaga et al., 2022), the latest well-performing pretrained language model that surpasses BERT in several benchmark datasets. However, RoBERTa outperforms LinkBERT in the general numeral attachment task. Therefore, we choose RoBERTa as the base language model for constructing NumBERT.

5.2 Experimental Results

For the volatility forecasting task, we utilize the same dataset and follow the separation criterion employed in previous works (Qin and Yang, 2019; Yang et al., 2020; Chen et al., 2021b). The mean

square error (MSE) is employed as the evaluation metric. We compare the performance of our proposed approach with the following baseline models: (1) **MDRM** (Qin and Yang, 2019): This model utilizes GloVe embeddings to represent textual data and feeds both textual and vocal features into a Bi-LSTM-based model for volatility forecasting. (2) **HTML** (Yang et al., 2020): This model uses BERT to extract textual features and aligns sentence-level audio features with the textual data. Volatility forecasting is then performed based on the concatenated features. (3) **FinBERT** (Araci, 2019): This BERT-based model is pre-trained on a financial corpus and has demonstrated effectiveness in various financial tasks. (4) **NAM** (Chen et al., 2021b): This model highlights the importance of extracting numeral features from earnings calls for volatility forecasting. It leverages the extracted numeral features with a Transformer architecture for making predictions.

Table 3 presents the experimental results. Firstly, our proposed GNA-Vol achieves the top rank in the 3-day and 7-day volatility forecasting tasks, highlighting the effectiveness of our pretraining task and model. Given the dynamic nature of the financial market and the rapid incorporation of new information into asset prices, short-term forecasting is of utmost importance. These results demonstrate that our general numeral attachment task enables models to capture critical information in earnings conference calls for short-term risk forecasting. Secondly, despite the intuitive appeal of the continuous strategy for time series data prediction, we observe that the discrete strategy outperforms the continuous strategy in most cases. This suggests that the vanilla continuous strategy may not adequately capture the latent information underlying the volatility clustering phenomenon discussed in Section 4.1.

To assess the impact of our proposed pretraining process, we conduct an ablation analysis, the results of which are presented in Table 4. We observe that without pretraining using our proposed task, the performance significantly deteriorates in most cases, irrespective of the strategy employed. These results underscore the effectiveness of our pretraining task in risk forecasting.

	3-day	7-day	15-day	30-day
MDRM (Text Only)	1.431	0.439	0.309	0.219
MDRM (Text + Audio)	1.371	0.420	0.300	0.217
HTML (Text Only)	1.175	0.372	**0.153**	**0.133**
HTML (Text + Audio)	0.845	0.349	0.251	0.158
FinBERT	0.750	0.368	0.241	0.172
NAM (Text Only)	0.745	**0.300**	0.232	0.187
GNA-Vol + Discrete Strategy	**0.700**	0.322	0.252	0.207
GNA-Vol + Continuous Strategy	0.705	0.362	0.250	0.237

Table 3: Experimental results of volatility forecasting task, reported in mean square error.

	3-day	7-day	15-day	30-day
GNA-Vol + Discrete Strategy	0.700	0.322	0.252	0.207
w/o Pretrain	0.730	0.353	0.253	0.186
GNA-Vol + Continuous Strategy	0.705	0.362	0.250	0.237
w/o Pretrain	0.725	0.421	0.275	0.231

Table 4: Ablation analysis. Note the lower the metric MSE, the better the performance.

6 Conclusion

We address the challenge of identifying the target associated with numerals in financial narratives through the novel task of general numeral attachment. By enhancing the publicly-available ECNum dataset with additional annotations, we provide a valuable resource for researchers in this field. Our approach, which simulates the reading process of financial professionals, enhances the semantic understanding of language models and improves the accuracy of 3-day volatility forecasting.

In the realms of tabular QA and math word problems, it's essential to note that earnings conference calls act as a nexus for corporate leaders and adept investors to converse about operational intricacies. Unlike traditional settings, these dialogues typically don't showcase tabular data or pose mathematical queries. This distinct context accentuates the importance of our general numeral attachment task. By deciphering numerical data embedded in such discourses, we augment the semantic grasp of these interactions. Future research can pivot on multiple avenues. Firstly, extrapolating the general numeral attachment task to diverse financial manuscripts and sectors could amplify its scope. Such extensions might delve into associating numerals with an array of financial elements, encompassing market indices, commodities, or macroeconomic markers. Secondly, the exploration of transfer learning or domain adaptation strategies to make our proposed model universally applicable to other financial prediction tasks stands as a promising endeavor. Finally, orchestrating user assessments or real-world appraisals of the improved volatility forecasting mechanisms can shed light on their tangible efficacy. Pursuing these trajectories will further crystallize the role of numeral

attachment in fine-tuning financial examinations.

Acknowledgments

This research is supported by National Science and Technology Council, Taiwan, under grants MOST 110-2221-E-002-128-MY3 and NSTC 111-2634-F-002-023-. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956 and a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Limitations

While this paper introduces a novel task of general numeral attachment and presents valuable contributions, there are some limitations to consider. First, the proposed general numeral attachment task is evaluated on the publicly-available ECNum dataset, which may have its own limitations and biases. The generalizability of the results to other financial documents and datasets needs to be further investigated. Second, the enhanced performance in 3-day volatility forecasting is demonstrated based on the specific approach proposed in this paper. It would be valuable to compare the performance with other existing volatility forecasting models and explore the robustness of the proposed approach across different datasets and market conditions. Third, the proposed approach focuses on enhancing the language model's understanding of numerals. While this is an important aspect, there may be other factors and features that contribute to accurate volatility forecasting. Future research could explore additional contextual information and features to further improve the forecasting accuracy.

Overall, while this paper provides valuable insights and advancements in the field of volatility forecasting through the general numeral attachment task, further research is needed to validate the findings on different datasets, compare with existing models, and explore additional features for enhanced performance.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Numeral attachment with auxiliary tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1161–1164.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Numclaim: Investor's fine-grained claim detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1973–1976.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021a. *From Opinion Mining to Financial Argument Mining*. Springer Nature.

Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, and Hsin-Hsi Chen. 2021b. Distilling numeral information for volatility forecasting. In *Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management*, page 2920–2924, New York, NY, USA. Association for Computing Machinery.

Jacob 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, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieh-Sheng Lee and Jieh Hsiang. 2019. Patentbert: Patent classification with fine-tuning a pre-trained bert model. *arXiv preprint arXiv:1906.02124*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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, pages 390–401, Florence, Italy. Association for Computational Linguistics.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 441–451, New York, NY, USA. Association for Computing Machinery.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.