Distilling Numeral Information for Volatility Forecasting

Chung-Chi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
cjchen@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
Institute of Information Science, Academia Sinica, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhhuang@iis.sinica.edu.tw

Yu-Lieh Huang
Department of Quantitative Finance, National Tsing Hua University, Taiwan
Center for Research in Econometric Theory and Applications, National Taiwan University, Taiwan
ylihuang@mx.nthu.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
hhchen@ntu.edu.tw

ABSTRACT

The volatility of stock price reflects the risk of stock and influences the risk of investor's portfolio. It is also a crucial part of pricing derivative securities. Researchers have paid their attention to predict the stock volatility with different kinds of textual data. However, most of them focus on using word information only. Few touch on capturing the numeral information in textual data, providing fine-grained clues for financial document understanding. In this paper, we present a novel dataset, ECNum, for understanding the numerals in the transcript of earnings conference calls. We propose a simple but efficient method, **N**umeral-**A**ware **M**odel (NAM), for enhancing the capacity of numeral understanding of neural network models. We employ the distilled information in the stock volatility forecasting task and achieve the best performance compared to the previous works in short-term scenarios.

CCS CONCEPTS

• **Applied computing** → **Forecasting**.

KEYWORDS

Numeracy, volatility forecasting, opinion mining

ACM Reference Format:
Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, and Hsin-Hsi Chen. 2021. Distilling Numeral Information for Volatility Forecasting. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482089

1 INTRODUCTION

Volatility forecasting, a long-term discussed issue, attracts many researchers' attention. Some of them use regression models with price data to predict the movement of volatility [3, 16]. Some of them make prediction based on textual data [8, 17, 19, 24]. Although many works pay attention to predicting stock volatility, few of them take the numerals in the financial document into consideration. In this paper, we provide the exploration in this direction.

Numerals are important in the documents of various domains. In scientific documents, numerals are used for reporting the statistics and experimental results [20]. In clinical notes, physicians use numerals to record the temperature-pulse-respiratory rate and dosage [20]. In financial narratives, numerals occupy a high proportion [6] and are useful for analyzing and comparing the operations of companies [15]. Recently, the transcript of earnings conference call attracts much attention in the NLP community [12, 17, 24]. The earnings conference call is a fiscal quarter meeting held by a company to reveal its operations and the transcript will be publicly available soon after the meeting. However, previous works neither provide the fine-grained annotation on the transcript nor analyze the meaning of the numerals in it. In this paper, we adopt the transcript of earnings conference call to construct a new dataset, ECNum, for understanding the numerals in the transcript with three novel tasks.

Currently, the approach encoding documents by pre-trained language models such as BERT [9] and then predicting with simple multilayer perceptron performs well in many NLP tasks. In this work, we find that instead of encoding the whole document directly, using the distilled numeral information from earnings conference calls can improve the performance of the short-term volatility forecasting task. The proposed approach refines the sentences in earnings conference calls into tuples based on the numeral in the sentences and further use the extracted tuples for predicting stock price volatility. Our experimental results support the usefulness of the distilled information.

Our contributions are threefold as follows. (1) We present a dataset for understanding the numeral information in earnings conference calls. (2) We propose a simple but efficient model for distilling the numeral information, and show that the proposed model performs better than other previous models. (3) We also experiment on a real-world application scenario, volatility forecasting, and find that our method performs well in short-term volatility forecasting tasks.

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482089

Table 1: Statistics of category labels in different financial narratives.

	ECNum: Earnings Conference Calls		NumClaim: Analysis Reports [5]		FinNum: Tweets [7]	
Category	Instances	Ratio	Instances	Ratio	Instances	Ratio
Monetary: *money*	1,859	20.58%	874	16.99%	736	8.30%
Monetary: *quote*	-	-	75	1.46%	1,033	11.65%
Monetary: *change*	428	4.74%	18	0.35%	176	1.98%
Monetary: *buy price*	-	-	-	-	415	4.68%
Monetary: *sell price*	-	-	-	-	135	1.52%
Monetary: *forecast*	-	-	-	-	355	4.00%
Monetary: *stop loss*	-	-	-	-	35	0.39%
Monetary: *support or resistance*	-	-	-	-	302	3.41%
Percentage: *relative*	2,113	23.39%	708	13.76%	767	8.65%
Percentage: *absolute*	579	6.41%	810	15.75%	346	3.90%
Temporal: *date*	1,640	18.15%	2,134	41.49%	2,653	29.92%
Temporal: *time*	8	0.09%	3	0.06%	365	4.12%
Option: *exercise price*	-	-	-	-	132	1.49%
Option: *maturity date*	-	-	-	-	70	0.79%
Indicator	-	-	-	-	216	2.44%
Quantity: *relative*	274	3.03%	278	5.40%	982	11.07%
Quantity: *absolute*	1,299	14.38%				
Product/Version	296	3.38%	136	2.64%	150	1.69%
Ranking	37	0.41%	3	0.06%	-	-
Other	501	5.55%	105	2.04%	-	-
	9,034	100.00%	5,144	100.00%	8,868	100.00%

2 RELATED WORK

In the past, different resources were adopted to predict stock volatility. Rekabsaz et al. [19] present an information retrieval-based model with the sentiment features in 10-K reports. Dereli and Saraclar [8] use Management's Discussion and Analysis (MD&A) as the resources and propose a convolutional neural network-based model (CNN-NTC). Qin and Yang [17] propose a fusion model, which uses both textual and audio features in earnings conference calls for volatility forecasting. Yang et al. [24] propose a multi-task model with the same features as Qin and Yang [17]. However, none of the previous works discuss how to leveraging numeral information to improve the performance on volatility forecasting task. This work provides a pilot exploration in this direction.

When reading financial documents, investors pay attention to the numerals related to the account in financial statements, because many fundamental analysis methods are based on these key numerals. These accounts can be classified into two types, including Generally Accepted Accounting Principles (GAAP) and Non-GAAP. Unlike the accounts in GAAP, Non-GAAP accounts provide some space for managers to adjust the estimation of their operations. In the latest two decades, analyzing the Non-GAAP accounts attract lots of researchers' attention in Accounting and Finance. With the adjustment of Non-GAAP accounts, both managers and investors expect that the revealed information will be closer to the real situation of the company. Bhattacharya et al. [1] show that Non-GAAP accounts are more important than GAAP accounts for investors' decision making. Several works have evidenced that the Non-GAAP accounts influence the investors' decisions by comparing the market reactions to the Non-GAAP accounts [4, 10, 11, 22, 23]. Black et al. [2] provide an overview of the current practice of the Non-GAAP accounts. The above-mentioned works show that extracting the GAAP and Non-GAAP accounts is important to real-world usage. Although the account information is important for investors' decision-making, no publicly dataset is available for researchers to construct an automatic method to extract these accounts in earnings conference calls. In this paper, we propose a new dataset for covering this problem.

3 ECNUM DATASET

3.1 Task Design and Annotation Process

We propose three numeral-related tasks, which can be considered as three classification tasks. Given a sentence and a target numeral, we provide three different kinds of labels on the target numeral based on the context information.

- **Numeral Understanding**: Chen et al. [6] define the numeral understanding as numeral disambiguation task. In this task, we need to classify numerals into specific categories by their meanings. We annotate numerals in the proposed data with the categories as listed in Table 1. That is, we define this task as an eleven-class classification task.
- **Account Differentiation**: As we mentioned in Section 2, account information is important for investor's decision-making, and it is always represented by numerals. Based on the survey of Black et al. [2], over 97% of S&P 500 companies use Non-GAAP accounts to adjust their balance sheets, and this kind of adjustment will influence the investors' decisions. In this paper, we attempt to classify a given numeral into one of three classes, including *GAAP*, *Non-GAAP*, and *Other*.

Table 2: Statistics of Account Differentiation and Influence Analysis labels.

Account Differentiation			Influence Analysis		
Class	# labels	%	Class	# labels	%
GAAP	3,675	40.68%	Positive	5,467	60.52%
Non-GAAP	432	4.78%	Negative	669	7.41%
Other	4,927	54.54%	Neutral	2,898	32.08%

- **Influence Analysis**: In addition to understanding whether the given numeral is related to accounts or not, we also want to know the influence on the related account. Thus, we annotate the infleunce with three labels, including *Positive*, *Negative*, and *Neutral*.

Three graduate students in the Department of Finance annotate the numerals in the earnings conference calls. Before starting to annotate the data in the proposed dataset, the annotators take exercises until they achieve high agreement. In each round, the annotators label the same earnings conference call transcript independently, and then discuss the results with one another after they finish the annotation. Note that we use different earnings conference call transcripts in each round. They annotate 167 instances in the first round and 215 instances in the second round. Finally, the annotators achieve 94.88%, 92.06%, and 96.38% Cohen's kappa agreement in Task DNU-Account, DNU-Influence, and GNU, respectively. After practice, annotators start to label different earnings conference calls.

3.2 Statistics

The resulting ECNum dataset contains 9,034 annotated numerals. Table 1 and Table 2 show the statistics of the proposed dataset. In Table 1, we also compare the categories of the numerals in different narratives, including earnings conference calls (proposed ECNum), analysis reports [5], and social media posts [7]. First, we find that managers rarely discuss the company's stock price, and few analysts use technical analysis in their reports. However, social media users regularly tweet about technical analysis results. From this we can differentiate managers from investors and professionals from amateur investors. Second, numerals reveal different habits of market participants. Total 29.80% of numerals in earnings calls are PERCENTAGES (including both relative and absolute subcategories), but they constitute only 12.55% of social media data. When describing company operations, managers focus on comparisons rather than financial statements such as earnings. In contrast, investors, especially social media users, use many MONETARY numerals. Third, analysts use more TEMPORAL information than other market participants.

4 METHOD

Numeral-Aware Model Figure 1 shows the architecture of the proposed Numeral-Aware Model (NAM). There are three parts in the NAM, including sentence encoder (f^S), numeral encoder (f^N), and hierarchical decoder (f^H). Given a sentence and a target numeral, we use f^S to encode the whole sentence for getting a sentence embedding (E_S). We further use f^N to encode the target numeral, and then we will get a numeral embedding (E_N). The concatenation of E_S and E_N becomes the input of f^H for distilling the sentence into a tuple.

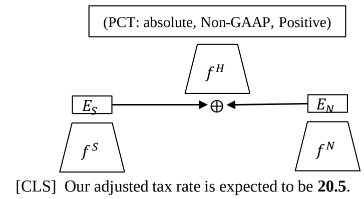

Figure 1: Architecture of the proposed NAM. $\oplus$ denotes concatenate.

In this paper, we adopt BERT-Large [9] as f^S and use CNN [13] as f^N. In f^N, we represent the target numeral by character and magnitude embeddings. The character embedding is a one-hot embedding of 0 to 9 and a decimal point. The magnitude embedding is the position embedding in v. Taking 20.5 in Figure 1 as an example, "2", "0", ".", "5" are in the positions 4, 3, 2, and 1 of v, respectively. The length of the concatenation of character embedding and magnitude embedding is 22. Because the longest numeral in the training set of ECNum has 11 digits, the target numeral is present by an 11×22 tensor.

We use a hierarchical multilayer perceptron structure for decoding the sentence into a tuple. The f^H first makes a prediction on the numeral understanding task. When making a prediction on account differentiation task, the input of the multilayer perceptron is the concatenation of sentence embedding, numeral embedding, and the prediction of the numeral understanding task. In addition to the input used for making a prediction on account differentiation task, when predicting the influence, the prediction of account differentiation task is also given. Thus, the hierarchical order of the tasks in f^H is numeral understanding task, account differentiation task, and influence analysis task.

Because the labels in account differentiation tasks are highly imbalanced, we propose an auxiliary task to improve performance. We merge "GAAP" and "Non-GAAP" labels as a group related to accounting principle and keep "Other" labels as a group in the auxiliary task. That is, the auxiliary task, which is a binary task, is a simplified task of account differentiation task. The f^H makes prediction on the auxiliary task and then predict on other tasks. We will show the usefulness of the auxiliary task based on the ablation analysis.

Volatility Forecasting After transferring all sentences in an earnings conference call into tuples, we use a two-layer transformer [21] to make volatility forecasting with these tuples. The linear activation function is adopted to make predictions, and the mean square error (MSE) is used as a loss function. We use Adam optimizer [14] during training process.

5 EXPERIMENTS

5.1 Experimental Setting

We separate the ECNum dataset by 70%, 10%, and 20% as the training set, development set, and test set. We then adopt vanilla BERT

Table 3: Experimental results under ECNum.

	NU	Account	Influence
BERT-Large	36.83	75.34	62.44
Multi-task BERT	48.61	75.13	72.77
T5-Base (220M)	74.60	85.28	74.38
NAM-Base (110M)	86.45	86.05	87.39
NAM (340M)	**88.42**	**88.15**	**89.87**

Table 4: Ablation analysis.

	NU	Account	Influence
NAM-Base	86.45	86.05	87.39
w/o f^N	48.37	77.50	67.57
w/o f^N & AT	36.98	77.70	64.20

(BERT-Large) [9], multi-task BERT, and T5 model (T5-Base) [18] as baselines. The multi-task BERT encodes the textual data with BERT-Large and make prediction on all tasks at the same time. That is, it does not use the hierarchical structure as that in the proposed model.

We use the transcipts of earnings conference calls provided by Qin and Yang [17], called ECT, to evaluate the performance of volatility forecasting task. We follow the separation criterion of previous works [17, 24] to separate the dataset into training, development, and test sets. We also follow the equations of previous works to calculate the volatility [17, 24]. Note that the time period of the data in ECNum is from 2018 to 2019, which does not overlap with that of ECT. For the volatility forecasting task, we adopt MDRM [17] and HTML [24] as our baselines.

The macro-F1 score is adopted for evaluating the performance of the proposed tasks. We also use the same metric in previous works [17, 24], mean square error (MSE), for evaluating volatility forecasting task.

5.2 Experimental Results

Table 3 shows the experimental results under the proposed dataset. Comparing the performance of BERT-Large with that of Multi-task BERT, we find that jointly learning with the proposed tasks can improve the performance of both numeral understanding (NU) and influence analysis tasks. T5-Base performs better among the baseline models, and the proposed NAM outperforms all baseline models in all tasks. Since the number of parameters of NumBERT-NU is more than that of T5-Base, it may cause an unfair comparison. We replace BERT-large with BERT-base in NAM and find that this model (NAM-Base) still performs better than T5-Base. In Table 4, we further show the ablation analysis by removing numeral embedding (f^N) and auxiliary task (AT) from NAM. The experimental results support the effectiveness of both f^N and the proposed auxiliary task.

Table 5 shows the results of volatility forecasting. Note that previous works focus on constructing fusion models with textual and audio data. In this work, we pay attention to the numerals in textual data. The experimental results show that the proposed model outperforms all baselines in 3- and 7-day volatility forecasting. In many cases, using the larger model can get better performance. However, we find that we only need few layers for making volatility forecasting with the distilled tuples.

Table 5: Experimental results of volatility 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
Proposed	**0.745**	**0.300**	0.232	0.187

Table 6: Performances in account differentiation task.

	Precision	Recall	F1-Score
GAAP	87.50	93.33	90.32
Non-GAAP	90.48	66.28	76.51
Other	94.99	92.49	93.72

5.3 Performances of Extracting Non-GAAP Accounts

Although Non-GAAP accounts only occupy 4.78% of numerals in earnings conference calls, many previous works [4, 10, 11, 22, 23] show that Non-GAAP account is a matter of investor's decision-making process. We report the details of the performances of NAM in Table 6. The performance on detecting Non-GAAP accounts is the worst due to the imbalanced distribution. In view of the importance of the Non-GAAP account, we plan to design a tailor-made model for extracting this information in the future.

6 CONCLUSION

In this paper, we propose a dataset and a model for distilling the numeral information from earnings conference calls, and show that the proposed model performs well in understanding the numeral information in the narrative. With the distilled information, the proposed method outperforms previous models in short-term volatility forecasting tasks. The experimental results support our postulation that understanding the numerals in the textual data is important. We release the annotations for academic usage under the CC BY-NC-SA 4.0 license.[1]

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 109-2218-E-009-014, MOST 110-2634-F-002-028, and MOST 110-2221-E-002 -128 -MY3. It was also financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 110L900202) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. 110-2634-F-002-045.

[1] http://ecnum.nlpfin.com

REFERENCES

[1] Nilabhra Bhattacharya, Ervin L Black, Theodore E Christensen, and Chad R Larson. 2003. Assessing the relative informativeness and permanence of pro forma earnings and GAAP operating earnings. *Journal of Accounting and Economics* 36, 1-3 (2003), 285–319.

[2] Dirk E Black, Theodore E Christensen, Jack T Ciesielski, and Benjamin C Whipple. 2018. Non-GAAP reporting: Evidence from academia and current practice. *Journal of Business Finance & Accounting* 45, 3-4 (2018), 259–294.

[3] Tim Bollerslev. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* 31, 3 (1986), 307–327.

[4] Mark T Bradshaw, Theodore E Christensen, Kurt H Gee, and Benjamin C Whipple. 2018. Analysts' GAAP earnings forecasts and their implications for accounting research. *Journal of Accounting and Economics* 66, 1 (2018), 46–66.

[5] 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*. 1973–1976.

[6] Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, and Hsin-Hsi Chen. 2018. Numeral understanding in financial tweets for fine-grained crowd-based forecasting. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 136–143.

[7] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*. 19–27.

[8] Neşat Dereli and Murat Saraclar. 2019. Convolutional Neural Networks for Financial Text Regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Florence, Italy, 331–337. https://doi.org/10.18653/v1/P19-2046

[9] 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*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[10] W Brooke Elliott. 2006. Are investors influenced by pro forma emphasis and reconciliations in earnings announcements? *The Accounting Review* 81, 1 (2006), 113–133.

[11] James R Frederickson and Jeffrey S Miller. 2004. The effects of pro forma earnings disclosures on analysts' and nonprofessional investors' equity valuation judgments. *The Accounting Review* 79, 3 (2004), 667–686.

[12] Katherine Keith and Amanda Stent. 2019. Modeling Financial Analysts' Decision Making via the Pragmatics and Semantics of Earnings Calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 493–503. https://doi.org/10.18653/v1/P19-1047

[13] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky, and Percy Liang. 2018. Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 82–92. https://doi.org/10.18653/v1/D18-1008

[16] Daniel B Nelson. 1991. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society* (1991), 347–370.

[17] 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*. Association for Computational Linguistics, Florence, Italy, 390–401. https://doi.org/10.18653/v1/P19-1038

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[19] Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Alexander Dür, Linda Andersson, and Allan Hanbury. 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 (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1712–1721. https://doi.org/10.18653/v1/P17-1157

[20] Georgios Spithourakis and Sebastian Riedel. 2018. Numeracy for Language Models: Evaluating and Improving their Ability to Predict Numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2104–2115. https://doi.org/10.18653/v1/P18-1196

[21] 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*. 5998–6008.

[22] Elmar R Venter, David Emanuel, and Steven F Cahan. 2014. The value relevance of mandatory non-GAAP earnings. *Abacus* 50, 1 (2014), 1–24.

[23] Matthew M Wieland, Mark C Dawkins, and Michael T Dugan. 2013. The differential value relevance of S&P's core earnings versus GAAP earnings: the role of stock option expense. *Journal of Business Finance & Accounting* 40, 1-2 (2013), 55–81.

[24] 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* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 441–451.