Overview of the NTCIR-16 FinNum-3 Task: Investor's and Manager's Fine-grained Claim Detection

Chung-Chi Chen
Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Japan
c.c.chen@acm.org

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

Hiroya Takamura
Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Japan
takamura.hiroya@aist.go.jp

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

In the FinNum task series, we proposed numeral category understanding (FinNum-1) and numeral attachment (FinNum-2) tasks for better comprehending the numerals in financial narratives. In FinNum-3, we present a novel task, fine-grained claim detection, and further integrate the new task with the previous ones. There are two subtasks in FinNum-3: (1) Investor's claim detection (Chinese) and (2) Manager's claim detection (English). This round of FinNum attracted ten research teams, and received 10 submissions for Chinese subtask and 15 submissions for English subtask. This paper provides an overview of FinNum-3, including task definition, data annotation, and participants' results.

CCS CONCEPTS

• **Information systems** → **Information extraction**.

KEYWORDS

Claim detection, argument mining, numeral corpus

1 INTRODUCTION

The importance of numerals in financial narratives is pointed out in Chen et al. (2018) [6], which is also the foundation of FinNum-1 [7]. The following-up works further demonstrate the usefulness of understanding numerals in downstream application scenarios, such as stock movement prediction [3, 8] and volatility forecasting [17]. The explorations on different languages are also done [14]. By investigating previous results, we notice that the notion of numeral understanding is closely related to the arguments in financial narratives. Thus, we propose a fine-grained claim detection task [4] for detecting investor's and manager's claim.

Argument mining has attracted much attention recently, and some related datasets are released for different topics. However, few previous studies analyze argumentation in the financial domain. In FinNum-3, we explore the task that detects the claims in the reports written by professional stock analysts (investor's claim) and the claims in the transcriptions of companies' earnings conference calls (manager's claim). Inspired by previous works, which found that numerals provide crucial information in financial narratives, we discover an interesting characteristic of financial claims: both investors and managers make claims with **estimation**. For example, when investors make a claim on sales growth rate, they will provide a fine-grained estimation such as "the sales growth rate may exceed **40**%". This kind of estimation provides more fine-grained opinions than only a stance (bullish or bearish). Comparing with this claim, the claim, "the sales growth rate may exceed **80**%", is stronger. This phenomenon evidences the importance of the numerals in the financial argumentation. Therefore, in FinNum-3, we aim at detecting such claims, i.e., identifying whether the given numeral is an in-claim numeral.

In addition to the claim labels, we also annotate the category information for enhancing the numeral understanding ability of machine learning models. That is, we propose a dataset with both numeral category understanding (a task in FinNum-1) and claim detection labels. We expect that the performance of downstream tasks will be improved with the annotated information. Thus, the task definition of FinNum-3 is:

> Given a sentence and the target numeral, we formulate the problem as a binary classification task to tell whether the given numeral is an in-claim numeral or not. Additionally, the category of target numeral is also included in the dataset. Participants can use this information to design joint learning models. For both claim detection and numeral category classification tasks, the micro-F1 and macro-F1 scores are adopted for evaluating the experimental results.

Table 1: Statistics of analyst report annotations.

Category	Train		Devlopment		Test		Total
	In-Claim	Out-of-Claim	In-Claim	Out-of-Claim	In-Claim	Out-of-Claim	
Monetary_money	428	311	78	57	413	-	1,287
Monetary_change	3	3	-	12	362	-	380
Monetary_price	34	32	8	1	30	28	
Percentage_relative	326	335	82	67	351	452	1,613
Percentage_absolute	171	394	37	106	169	572	1,449
Temporal_date	-	1,775	-	359	-	1,847	3,981
Temporal_time	-	3	-	-	-	1	4
Quantity_absolute	36	183	19	36	40	165	479
Quantity_relative	-	4	-	-	3	16	23
Product Number	1	100	-	35	1	145	282
Ranking	-	-	-	3	-	6	9
Other	-	80	-	25	-	90	195
Total	999	3,220	224	701	1,369	3,322	9,835

Table 2: Statistics of earnings conference call annotations.

Category	Train		Devlopment		Test		Total
	In-Claim	Out-of-Claim	In-Claim	Out-of-Claim	In-Claim	Out-of-Claim	
Monetary_money	352	1,144	45	221	24	338	2,124
Monetary_change	100	298	32	52	11	237	730
Percentage_relative	223	1,866	9	227	75	407	2,807
Percentage_absolute	193	490	18	67	36	119	923
Temporal_date	-	1,616	-	221	-	465	2,302
Temporal_time	-	8	-	-	-	-	8
Quantity_absolute	143	1,050	8	114	36	364	1,715
Quantity_relative	17	161	1	44	5	74	302
Product Number	9	226	1	25	-	14	275
Ranking	-	35	-	-	-	8	43
Other	2	404	-	106	-	170	682
Total	1,039	7,298	114	1,077	187	2,196	11,911

2 DATASET

Table 1 shows the statistics of our annotation results on professional analysts' reports written in Chinese, and Table 2 shows the statistics of our annotation results on the transcriptions of earnings conference calls written in English. For the details of category definitions, please refer to the FinNum-1 overview paper [7] and our previous work [6].

Based on these statistics, we find that in-claim numerals only occupy about 26.35% and 11.25% in investors' analysis and managers' talks, respectively. It indicates the proposed dataset is imbalanced. Since claims contain investors' and managers' opinions, we think that the in-claim numerals are more important than the out-of-claim ones. Thus, we will discuss the performance on in-claim labels of different models in the later section. Additionally, we also find the relationship between the numeral categories and the claim annotations. The numeral in some categories may never appear in the investor's and manager's claims. That shows the reason why we combine both the category understanding task and the claim detection task in FinNum-3.

3 METHODS IN OFFICIAL RUNS

For both subtasks, we adopt the CapsNet-based model with both numeral encoding and category auxiliary task [4] as the baseline. We summarize the methods proposed by participants in Table 3. The pre-trained language models and discussed methods are shown separately.

Participants solved FinNum-3 from 4 directions, including (1) data augmentation for dealing with the data imbalance problem, [11, 16] (2) numerical representation [13, 15], (3) knowledge-based approach [9], and (4) traditional machine learning techniques [1, 10]. IMNTPU and CYUT apply the translation and GPT-2 generation data augmentation methods, respectively. WUST represents target numerals with an additional encoder. JRIRD proposes an interesting exploration with several language models and several kinds of numerical representation for models' input. TMUNLP team adopts a knowledge base in their method. LIPI and Passau21 probe several traditional techniques. JRIRD, TMUNLP, and WUST explore joint learning approaches given the relationships between claim detection and category classification tasks.

Table 3: Summary of participants' methods.

Team	Subtask	Pre-Trained Language Model	Method
IMNTPU [16]	Chinese & English	XLM-RoBERTa	Data Augmentation (Translation)
CYUT [11]	Chinese & English	MacBERT, RoBERTa, and GPT-2	Data Augmentation (GPT-2), and AWD-LSTM
WUST [13]	Chinese & English	RoBERTa	Numeral Encoder, and Position Representation
JRIRD [15]	English	BERT, RoBERTa, FinBERT (News), and T5	Numerical Representation
LIPI [10]	English	FinBERT (News), and BERT-base	Ensemble
Passau21 [1]	English	BERT	Decision Tree, SVM, Naive Bayes, and CNN
TMUNLP [9]	Chinese	BERT, and RoBERTa	Knowledge-Based Approach

Table 4: Experimental results on analyst report.

Submission	Claim Detection		Numeral Category	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
CapsNet [4]	80.32%	69.19%	62.59%	20.99%
WUST_1	84.89%	75.70%	56.13%	17.35%
CYUT_2	91.73%	86.76%	-	-
TMUNLP_2	91.11%	87.76%	94.03%	72.99%
CYUT_3	92.16%	88.20%	-	-
CYUT_1	92.11%	88.80%	-	-
TMUNLP_1	92.82%	89.56%	94.31%	73.68%
TMUNLP_3	92.75%	89.68%	94.67%	73.89%
IMNTPU_2	94.14%	91.64%	-	-
IMNTPU_3	95.20%	92.91%	-	-
IMNTPU_1	95.31%	93.18%	-	-

Table 5: Experimental results on earnings conference call.

Submission	Claim Detection		Numeral Category	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
CapsNet [4]	89.97%	57.36%	49.64%	26.50%
BERFIN_2	85.10%	68.26%	-	-
WUST_1	93.37%	71.72%	48.76%	24.02%
BERFIN_1	94.67%	80.26%	-	-
LIPI_2	95.17%	81.33%	-	-
LIPI_1	95.09%	82.82%	-	-
LIPI_3	95.59%	84.73%	-	-
CYUT_1	94.67%	85.53%	-	-
Passau21_1	96.01%	87.12%	-	-
CYUT_2	95.64%	87.49%	-	-
CYUT_3	96.43%	87.88%	-	-
IMNTPU_1	96.18%	88.39%	-	-
JRIRD_2	96.73%	89.55%	89.76%	72.84%
IMNTPU_2	96.73%	89.86%	-	-
JRIRD_1	97.15%	90.80%	89.68%	72.94%
JRIRD_3	97.27%	91.03%	89.26%	69.11%

4 EXPERIMENTAL RESULTS

We use micro-F1 and macro-F1 scores[1] to evaluate the experimental results. Table 4 and Table 5 report the evaluations on baseline method and participants' methods.

Since most participants explore their methods with different pre-trained language models (LM), it is a bit hard to compare the methods given the same condition on LMs. However, based on these results, we get the following findings. Firstly, JRIRD gets the best performance in the English subtask. That shows numerical representation [15] is a worth exploring issue in the numerical-rich documents, such as the documents and narratives in the financial domain. Secondly, data augmentation [11, 16] can address the data imbalance issue in the proposed tasks, and it works in both English and Chinese subtasks. Thirdly, based on the results of IMNTPU, we find that there are only a few performance gaps between the Chinese and English subtasks, but the best-performing methods given different languages are not the same. Fourthly, Passau21 shows that some keywords, like "expect" in earnings conference calls, provide an important clue for the decision tree model. A similar analysis of analyst reports is shown in our previous work [4].

5 DISCUSSION

5.1 Error Analysis

In this section, we analyze the instances that get the false predictions based on each team's best-performing model. We categorize the error instances into the following cases.

[1]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

(1) **Older Claim**: Reached our forecast 4Q18 after-tax profit of **69%**.
(2) **Near Claim Sentence**: Researchers estimate that the cash dividends distributed by TSMC in 2019 will increase slightly by 0.5 yuan to 8.5 yuan compared with 2018. Based on the 1/17 closing price of **220.5** yuan, the cash dividend yield is 3.9%.
(3) **Seldom Narrative Pattern**: The theoretical revenue of the second-stage production capacity of health food products is about **2.2** billion yuan.
(4) **Further Explanation**: We anticipate producing 200000 barrels per day this year as I communicated at the Analyst Meeting. And that's – of course that's up more than **10%** from 2017.

The 69% in the first instance is the old claim of the analyst, and it should not be considered a claim in the current report. In the second instance, the 220.5 is the stock's closing price, and models may be misleading due to the previous sentence being a claim. In the third case, "theoretical revenue" is not commonly used in the instances in training data. Thus, models cannot identify it as an in-claim case. The fourth example shows that the further explanation sentence is not identified as claim-related among all best-performing models. These instances indicate some possible direction for future work,

Table 6: Evaluation on In-claim cases in analyst report.

	P	R	F1
WUST_1	68.63%	54.52%	60.76%
CYUT_1	78.11%	87.88%	82.71%
TMUNLP_3	79.64%	88.98%	84.05%
IMNTPU_1	**87.09%**	**91.76%**	**89.36%**

Table 7: Evaluation on In-claim cases in earnings conference call.

	P	R	F1
WUST_1	63.06%	37.43%	46.98%
LIPI_3	72.04%	71.66%	71.85%
Passau21_1	71.30%	82.35%	76.43%
CYUT_3	76.29%	79.14%	77.69%
IMNTPU_2	73.19%	**91.98%**	81.52%
JRIRD_3	**79.33%**	88.24%	**83.54%**

e.g., to propose datasets or to develop models in the claim detection task of financial narratives.

5.2 Evaluation on In-Claim Cases

Because the distribution of the labels in the proposed dataset is highly imbalanced, we further analyze the performance of in-claim instances. As we mentioned, the claims of both analysts and managers are always the foci of market participants. That's why we pay attention to in-claim cases. Table 7 and Table 6 show the results of both subtasks. We notice that the method proposed by team IMNTPU achieves the highest recall in both subtasks. We guess it may be contributed by words changed during the data augmentation process. The evaluation results also show that there is still room for performance on in-claim numeral detection.

5.3 Future Directions

In this section, we point out some possible directions for future work and further analysis.

- **Combination of the Findings in FinNum-3**: Future work could adopt the numeral representation proposed by team JRIRD [15] with the data augmentation method proposed by team IMNTPU [16]. It is expected to improve performance because both teams solve the FinNum-3 task from different aspects and get great results. Additionally, the cross-analysis of the generated instances between the translation-based augmentation method [16] and the GPT2-based augmentation method [11] would also be an interesting direction.
- **Adopting Document-Specific Language Model**: We notice that some participants adopt FinBERT (News) [2] in their model. Although FinBERT (News) is trained with financial news articles and titles, the topics and narrative style may still be different from those in analyst reports and earnings conference calls. We suggest future studies adopt the FinBERT (Earnings Conference Call) [12] in their experiment. That could also improve the performance.
- **Adopting the Extracted Claims in Downstream Tasks**: Since the length of analyst reports and earnings conference calls always exceeds the maximum length limitation of neural network models, it makes the models hard to get complete pictures of the report/talk. Given the importance of the analyst's and manager's claims, we think that extracting such claims could be a way to distill the long documents, and the information embedded in the claims may provide enough information for investors to do further market information forecasting.
- **In-Depth Argumentation Analysis**: FinNum-3 proposes the first exploration of argument mining topics in financial documents. As the research agenda proposed in our previous work [5], there is much fine-grained information that could be explored in the future, such as (1) the argumentation structure in the report/talk and (2) the argumentation relations among financial opinions. Thus, we plan to propose a series of shared tasks, named financial argument mining (FinArg) in the future.

6 CONCLUSION

This paper summarizes the methods in FinNum-3, and provides a cross-method analysis of participants' results. The discussion on state-of-the-art performance in financial claim detection is also included. Additionally, we point out several possible directions for future work.

Since 2018, the FinNum share task series have explored several tasks for understanding the numerals in the financial narrative, including category classification, numeral attachment, and claim detection. Both formal (earnings conference call and analyst report) and informal (social media data) are explored, and two languages (English/Chinese) are included. The participants proposed several models and features for solving the proposed tasks, and we learned a lot from their results. For example, representing the numeral features are important in the numeral category classification task, and the numeral attachment may still be a challenging task for vanilla neural network models. In FinNum-3, we learn the methods of numerical representation and data augmentation for financial narratives. We believe that future work related to financial narrative analysis or other numeral-rich document understanding could benefit from the exploration of FinNum participants. The findings of FinNum shared tasks could also be used in the next share task series of financial document understanding.

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 110-2634-F-002-028, MOST 110-2634-F-002-016, MOST 110-2221-E-002-128-MY3, and MOST 110-2634-F-002-050. 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.

REFERENCES

[1] Alaa Alhamzeh, M. Kürsad Lacin, and Előd Egyed-Zsigmond. 2022. Passau21 at the NTCIR-16 FinNum-3 Task: Prediction Of Numerical Claims in the Earnings Calls with Transfer Learning. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.*

[2] Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063* (2019).

[3] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Crowd View: Converting Investors' Opinions into Indicators.. In *IJCAI.* 6500–6502.

[4] 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.

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

[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] Chung-Chi Chen, Hen-Hsen Huang, Chia-Wen Tsai, and Hsin-Hsi Chen. 2019. Crowdpt: Summarizing crowd opinions as professional analyst. In *The World Wide Web Conference.* 3498–3502.

[9] Tzu-Ying Chen, Yu-Wen Chiu Chiu, Hui-Lun Lin, Chia-Tzu Lin, Yung-Chung Chang, and Chun-Wei Tung. 2022. TMUNLP at the NTCIR-16 FinNum-3 Task: Multi-task Learning on BERT for Claim Detection and Numeral Category Classification. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.*

[10] Sohom Ghosh and Sudip Kumar Naskar. 2022. LIPI at the NTCIR-16 FinNum-3 Task: Ensembling transformer based models to detect in-claim numerals in Financial Conversations. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.*

[11] Xie-Sheng Hong, Jia-Jun Lee, Shih-Hung Wu, and Tian-Jian Jiang. 2022. CYUT at the NTCIR-16 FinNum-3 Task: Data Resampling and Data Augmentation by Generation. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.*

[12] Allen Huang, Hui Wang, and Yi Yang. 2020. FinBERT—A Deep Learning Approach to Extracting Textual Information. *Available at SSRN 3910214* (2020).

[13] Liu Maofu Liu, Yuxuan and Mengjie Wu. 2022. WUST at NTCIR-16 FinNum-3 Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.*

[14] Maofu Liu, Xinxin Xia, and Wei Wang. 2021. A Chinese Dataset for Exploring Financial Numeral Attributes. In *Companion Proceedings of the Web Conference 2021.* 255–259.

[15] Shunsuke Onuma and Kazuma Kadowaki. 2022. JRIRD at the NTCIR-16 FinNum-3 Task: Investigating the Effect of Numerical Representations in Manager's Claim Detection. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.*

[16] Yung-Wei Teng, Pei-Tz Chiu, Ting-Yun Hsiao, Mike Tian-Jian Jiang, and Min-Yuh Day. 2022. IMNTPU at the NTCIR-16 FinNum-3 Task: Data augmentation for financial Numclaim classification. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies.*

[17] Linyi Yang, Jiazheng Li, Ruihai Dong, Yue Zhang, and Barry Smyth. 2022. NumHTML: Numeric-Oriented Hierarchical Transformer Model for Multi-task Financial Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence.*