NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001323

Overview of the NTCIR-17 FinArg-1 Task: Fine-Grained Argument Understanding in Financial Analysis

Chung-Chi Chen
AIST, Japan
c.c.chen@acm.org

Chin-Yi Lin, Chr-Jr Chiu
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
{cyline,ccchiu}@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
Institute of Information Science, Academia Sinica, Taiwan
hhhuang@iis.sinica.edu.tw

Alaa Alhamzeh,
Universität Passau, Germany
alaa.alhamzeh@uni-passau.de

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
AIST, Japan
takamura.hiroya@aist.go.jp

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
hhchen@ntu.edu.tw

ABSTRACT

This paper provides an overview of FinArg-1 shared tasks in NTCIR-17. We propose six subtasks with three different resources, including company manager presentations, professional analyst reports, and social media posts. 19 research teams registered for FinArg-1, and 11 teams submitted their system output for official evaluation. Participants explored several state-of-the-art language models such as BERT, T5, ELECTRA, and GPT-3.5, and leveraged techniques such as fine-tuning, ensemble learning, and prompt-based approaches.

CCS CONCEPTS

• **Information systems → Information extraction**.

KEYWORDS

argument mining, argument unit detection, argument relation, sentiment analysis

1 INTRODUCTION

In the series of FinNum tasks [5–7], we focused on a crucial feature of financial narratives—numerals. As these tasks were engineered to comprehend the numeric elements, we assert that the holistic view of entire financial documents hasn't been fully encapsulated in prior tasks. Consequently, we introduce a novel shared task series emphasizing the fine-grained argument information in financial narratives.

The aim of the FinArg task series is to understand the arguments present in investor-generated text, encompassing both professional and amateur textual data. Table 1 presents an overview of our blueprint for the FinArg task series. We plan to annually propose two intriguing tasks, employing both Chinese and English data, thereby expanding the participant group and expediting financial argument mining's development.

In FinArg-1, we introduce two tasks: (1) Argument-based Sentiment Analysis, and (2) Argumentative Relation Identification within discussion threads. Subsequent tasks, FinArg-2 (Argument Validity Period Assessment, Temporal Reference Detection [2]), and FinArg-3 (Argument Forecasting Skill Estimation, Argument Quality Assessment [1]), focus on temporal information assessment—a unique phenomenon in financial opinions—and leveraging all features and findings from FinNum-1 to FinArg-2 to discern opinions with high forecasting skills. We believe that through the exploration of FinNum and FinArg, numerous innovative ideas will emerge, and the model's competence in understanding financial documents will be enhanced.

Despite argument mining being a topic of discussion for several years [18, 31], financial argument mining is still nascent. Table 2 provides an overview of the task in FinArg-1. In FinNum-3, we broached the concept of identifying arguments in financial narratives. In a bid to conduct a more nuanced analysis, we introduced an argument-based sentiment analysis task in FinArg-1, rooted in the notion that positive news does not always lead to a bullish claim. In this task, we bifurcate the analyst report into two sections: premise and claim, and further label the sentiment directed towards the argument. The premise is labeled with positive/neutral/negative sentiment, while the claim is labeled as bullish/neutral/bearish. This approach allows us to better comprehend the argumentation structure in professional reports. Additionally, we also adopt the transcripts of earnings conference calls as a resource for traditional

Table 1: Overview of FinArg task series.

Short Name	Language	Source	Task
FinArg-1	English	Analyst Report	Argument-based Sentiment Analysis
	English	Earnings Call	Argument Unit/Relation Identification
	Chinese	Social Media	Identifying Attack and Support Argumentative Relations in Social Media Discussion Thread
FinArg-2	English	Analyst Report	Premise's Influence Period Assessment
	English	Earnings Calls	Argument Temporal Reference Detection
	Chinese	Social Media	Claim's Validity Period Assessment
FinArg-3	English	Analyst Report	High Forecasting Skill Report Retrieval
	English	Earnings Calls	Argument Quality Assessment
	Chinese	Social Media	High Forecasting Skill Opinion Retrieval

Table 2: Overview of FinArg-1.

Task	Subtask
1. Argument-based Sentiment Analysis	1. Argument Classification
	2. Premise Sentiment Analysis
	3. Claim Sentiment Analysis
2. Argument Identification	1. Argument Unit Identification
	2. Argument Relation Identification
3. Identifying Attack and Support Argumentative Relations in Social Media Discussion Thread	-

Table 3: Data statistics of argument-based sentiment analysis.

Argument	Sentiment	Train	Dev	Test	Whole
Premise	Positive	4,441	189	479	5,109
	Negative	3,712	224	592	4,528
	Neutral	984	42	114	1,140
Claim	Bullish	2,374	106	324	2,804
	Bearish	2,013	105	380	2,498
	Neutral	977	73	85	1,135
Total		14,501	739	1,974	17,214

Table 4: Data statistics of argument unit identification.

	Train	Dev	Test	Whole
Preminse	4,062	508	508	5,078
Claim	3,691	461	461	4,613
Total	7,753	969	969	9,691

argument mining tasks [3], argument unit identification and argument relation identification.

Simultaneously, another task aims to identify the attack and support argumentative relationships within the social media discussion thread. Instead of analyzing individual social media posts, we examine the entire discussion thread. We strive to link the posts with attack and support labels, enhancing our understanding of the argumentation structure among opinions. We posit that the features extracted in the FinArg-1 tasks are linked to forecasting skills, a topic we will delve into in FinArg-3. FinArg-1 is expected to spur further discussions within our community regarding more granular information embedded within financial documents.

Table 5: Data statistics of argument relation identification.

	Train	Dev	Test	Whole
Support	3,859	482	482	4,823
Attack	62	8	8	78
Other	1,600	200	200	2,000
Total	5,521	690	690	6,901

Table 6: Data statistics of social media data.

	Train	Dev	Test	Whole
Support	3,676	460	460	4,596
Attach	2,158	270	270	2,698
Other	684	85	85	854
Total	6,518	815	815	8,148

2 TASK DESIGN

Table 2 shows an overview of FinArg-1. There are three subtasks in the argument-based sentiment analysis task: (1) argument classification, (2) premise sentiment analysis, and (3) claim sentiment analysis. In the argument classification subtask, participants are asked to classify the given sentence into claim or premise. In the premise sentiment analysis subtask, participants need to classify the given premise into positive, neutral, or negative. In the claim sentiment analysis subtask, participants will classify the given claim into bullish, neutral, or bearish.

For argument identification within earnings conference calls, participants are confronted with two subtasks: (1) argument unit identification, and (2) argument relation identification. The first subtask requires participants' systems to distinguish whether the given sentence functions as a claim or a premise. The second subtask, on

Table 7: Methods for the social media subtask.

Team	Language Model	Approach	Feature
TMUNLP [20]	ChatGPT, MacBERT [11], BARD [25]	Data Augmentation, Emsemble	ChatGPT Keywords
SCUNLP-2 [16]	Chatgpt-detector-roberta-chinese [13], BERT [12], XLM-RoBERTa [10]	Prompting	ChatGPT Generated Information
Quack [21]	BERT [12]	Two-Step Forecasting, Masked-LM Fine-tuning	
LIPI [4]	BERT-SEC [24], FLANG-RoBERTa [28], SBERT[1], DistilRoBERTa [27]		Translate
CYUT [33]	ChatGPT	Prompting	
IMNTPU [30]	RoBERTa [22], ALBERT [17]		
WUST [32]	BERT [12]		

the other hand, necessitates the identification of the relationship, specifically discerning whether it's one of support, attack, or other.

There is only one goal in the third task — identifying attack and support argumentative relations in social media discussion threads. Participants are asked to identify the argumentative relations (attack, support, or irrelevant) between two given social media posts.

3 DATASET

Table 3 provides the statistics of the dataset for argument-based sentiment analysis tasks. We found that professional analysts listed more positive premises than negative ones, and they have more bullish claims than bearish ones.

Tables 4 and 5 show the statistics of the dataset for argument unit and relation identification tasks, respectively. We found that managers seldom attack their own statements in the presentation. They try to support their claims in most of their speeches. This data covers earnings conference calls in the period of 2015-2019 for four tech companies [3].

Table 6 shows the statistic of the attack and support argumentative relation between social media posts. We found that social media users support others' opinions more than attack others' opinions.

4 PARTICIPANTS' METHODS

4.1 Argument Unit Identification in ECCs

TMUNLP [20] ranked as the best team with 76.55% macro-F1 score as seen in Table 8. Their submitted runs depend on different combinations of pairs of language models using the concept of voting. IDEA [29] used the last_hidden_state embedding generated by BERT[12] as the initial state of a convolutional neural network. TUA1 [34] adopted the prompt-based learning and instruction finetuning on the T5 model [26]. They experimented various sorts of prompts, and achieved their best performance using a short simple one "Choose premise or claim:". IMNTPU [30] explored the potential of GPT 3.5 Turbo. However, a Roberta base solution still overcome it in their conducted experiments. GPT 3.5 Turbo was also used by the team of Monetech [15] in a zero and ten shots learning strategies. They also used it to generate more data similar to the one provided by the task. The generated rephrased sentences are then passed into a data filtering based on its length. Their best submitted run was using a Bert model fine-tuned (with a freezed embedding layer) on the training dataset that has the shortest 25% of the data removed. LIPI [4] fine-tuned the model of Bert-SEC [24], and similarly, WUST [32] applied simply Bert. Thus, we consider it as our baseline for this sub-task.

4.2 Argument Relation Detection and Classification

4.2.1 *Earnings Conference Calls.*

Participant teams have examined and explored different language models like Bert [12], DistilBert [27], Bert-SEC [24], Bart[19], and DeBERTa [14], as well as different approaches like ELECTRA [9], and data augmentation.

Among others, TUA1-1 [34] scores the best results by fine-tuning T5-large model [26] on the Financial Phrasebank dataset. They follow the prompt-based learning and instruction fine-tuning. Similarly, IDEA [29] classified the sentence-pairs based on prompting. LIPI [4] and IMNTPU [30] achieved their best results by tuning FinBert [23], while TMUNLP [20] adopted both Bart and Deberta, with different sampling strategies. They also used the LLR (Log-Likelihood Ratio) method as a measure of word relationships between both sentences. Finally, SCUNLP [8] utilized both the original data along with the generated answers to ten proposed questions by ChatGPT as additional supporting features to fine-tune Distilbert model.

4.2.2 *Social Media Threads.*

Table 7 provides an overview of the techniques suggested by participants for the social media subtask. A variety of language models were explored, such as ChatGPT, MacBERT [11], BARD [25], and others including Chatgpt-detector-roberta-chinese [13] and SBERT. TMUNLP [20] and SCUNLP-2 [16] utilized generated text from ChatGPT as supplementary indicators for predictions. Quack [21] introduced a bifurcated strategy: initially filtering unrelated pairs from the "support/attack" category, followed by predicting their stance in the subsequent step. Conversely, CYUT [33] engaged ChatGPT without any fine-tuning adjustments.

5 EXPERIMENTAL RESULTS

Tables 8, 9, and 10 show the experimental results of argument unit identification, argument relation identification in ECCs, and attack support argument relation identification in social media, respectively.

In terms of argument unit identification in ECCs, different large language models were examined either by prompting or finetuning, with no huge difference in the outcome. We consider WUST [32] who fine-tuned Bert as the task baseline (74.41% macro F1-score). TMUNLP-1 [20] achieved the best performance (76.55% macro F1-score) by assembling the outputs of ELECTRA and Roberta using a voting mechanism. This sheds the light on the added value of ensemble learning techniques. By merging the collective predictions, we can significantly enhance the predictive accuracy.

However, the relation classification in this type of conversational text shows more complexity, especially with the unbalance nature

Table 8: Results of argument unit identification.

Team	Micro-F1	Macro-F1	Weight-F1
TMUNLP-1	76.57%	76.55%	76.59%
IDEA-1	76.47%	76.46%	76.48%
TUA1-1	76.37%	76.36%	76.38%
IMNTPU-2	76.06%	76.05%	76.07%
TMUNLP-3	76.06%	76.04%	76.07%
TMUNLP-2	75.95%	75.94%	75.97%
MONETECH-3	75.54%	75.53%	75.56%
IMNTPU-1	75.44%	75.31%	75.40%
MONETECH-1	75.13%	75.13%	75.12%
MONETECH-2	75.03%	75.02%	75.04%
TUA1-0	74.61%	74.56%	74.62%
WUST-1	74.41%	74.41%	74.41%
LIPI-3	73.89%	73.86%	73.90%
IDEA-3 (Late)	73.68%	73.68%	73.69%
LIPI-1	73.48%	73.47%	73.49%
LIPI-2	73.27%	73.27%	73.28%
SCUNLP-1-2	71.10%	71.07%	71.02%
SCUNLP-1-3	71.10%	70.53%	70.73%
SCUNLP-1-1	68.73%	68.62%	68.53%
WUST-2	69.04%	67.76%	68.07%
IMNTPU-3	56.97%	56.82%	56.70%

Table 9: Results of argument relation identification in ECCs.

Team	Micro-F1	Macro-F1	Weight-F1
TUA1-1	85.65%	61.50%	84.86%
LIPI-3	79.42%	60.22%	78.90%
TMUNLP-2	82.03%	57.90%	81.57%
TMUNLP-1	81.88%	57.36%	81.45%
TMUNLP-3	81.88%	56.72%	81.52%
TUA1-2	81.30%	56.26%	80.76%
TUA1-0	85.94%	55.36%	85.13%
SCUNLP-1-3	72.17%	54.06%	72.35%
WUST-1	78.70%	53.97%	77.93%
IMNTPU-2	82.61%	52.97%	82.14%
IDEA-3 (Late)	81.74%	51.85%	80.88%
LIPI-1	80.72%	51.35%	80.09%
IDEA-1	80.58%	51.12%	79.89%
LIPI-2	80.29%	51.08%	79.79%
IMNTPU-3	80.72%	50.73%	79.67%
SCUNLP-1-2	68.55%	49.00%	68.57%
IMNTPU-1	78.99%	47.36%	76.54%
SCUNLP-1-1	68.70%	45.68%	68.05%
IDEA-2	57.10%	29.18%	59.39%

of the data. That's because company representatives tend to support their claims more than discussing the opponent point of view, which leads to an attack relation between the premise and the claim. Hence, the best classification is delivered by TUA1-1 [34] who employed T5 (fine-tuned using the financial Phrasebank dataset) with a weighted random sampler to increase the probability of sampling minority labels.

Table 10: Results on the social media dataset.

Team	Micro-F1	Macro-F1	Weight-F1
Quack-2	71.66%	73.94%	71.35%
WUST-1	70.55%	70.64%	70.30%
Quack-1	67.85%	70.28%	67.30%
LIPI-3	64.79%	69.45%	64.09%
Quack-3	65.52%	66.88%	63.76%
SCUNLP-2-3	62.58%	66.39%	63.37%
SCUNLP-2-1	56.81%	59.76%	57.08%
SCUNLP-2-2	56.56%	59.61%	57.21%
LIPI-2	56.81%	58.28%	56.89%
LIPI-1	59.14%	57.30%	59.62%
CYUT-2	68.22%	49.62%	68.22%
TMUNLP-1	46.38%	35.37%	45.84%
IMNTPU-1	52.88%	34.77%	48.73%
TMUNLP-3	45.28%	32.48%	43.45%
TMUNLP-2	41.96%	31.69%	41.99%
IMNTPU-2	48.71%	24.64%	40.50%
CYUT-3	29.20%	23.45%	30.56%
CYUT-1	24.54%	20.94%	25.54%

The results presented in Table 10 indicate that Quack's method [21] is the most effective. They adapted BERT using data sourced from another Taiwanese social media platform and employed this fine-tuned BERT for predictions. A comparison with WUST's outcomes [32] sheds light on the distinctions between the fine-tuned and original BERT. Meanwhile, the findings from CYUT [33] underscore how the choice of prompts can markedly influence performance.

6 CONCLUSION

This paper summarizes the dataset and methods in FinArg-1. Participants present a comprehensive exploration of various methodologies adopted by different teams for FinArg-1 tasks. While various models and strategies have shown promise in specific subtasks, there remains ample room for innovation in this field. The nuanced differences in outcomes across tasks underscore the importance of tailoring approaches to the unique characteristics of each dataset. Future research might delve deeper into ensemble techniques, targeted data augmentation, and more refined tuning strategies to further elevate performance in argumentative text analysis.

After understanding and exploring the basic elements of arguments in different financial documents. We plan to propose FinArg-2, which is related to argument temporal inference. We will continue to use research reports, the transcripts of earnings conference calls, and social media posts.

ACKNOWLEDGMENTS

This research is supported by National Science and Technology Council, Taiwan, under grants 110-2221-E-002-128-MY3, 110-2634-F-002-050-, and 111-2634-F-002-023-. The work of Chung-Chi Chen and Hiroya Takamura 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).

REFERENCES

[1] Alaa Alhamzeh. 2023. Financial Argument Quality Assessment in Earnings Conference Calls. In *Database and Expert Systems Applications*, Christine Strauss, Toshiyuki Amagasa, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil (Eds.). Springer Nature Switzerland, Cham, 65–81.

[2] Alaa Alhamzeh. 2023. *Language Reasoning by means of Argument Mining and Argument Quality*. Ph.D. Dissertation. Universität Passau.

[3] Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. It's Time to Reason: Annotating Argumentation Structures in Financial Earnings Calls: The FinArg Dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. 163–169.

[4] Swagata Chakraborty, Anubhav Sarkar, Dhairya Suman, Sohom Ghosh, and Sudip Kumar Naskar. 2023. LIPI at the NTCIR-17 FinArg-1 Task: Using Pre-trained Language Models for Comprehending Financial Arguments. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001281

[5] Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2022. Overview of the ntcir-16 finnum-3 task: investor's and manager's fine-grained claim detection. In *Proceedings of the 16th NTCIR conference on evaluation of information access technologies, Tokyo, Japan (forthcoming)*.

[6] 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.

[7] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2020. Overview of the NTCIR-15 FinNum-2 Task: Numeral attachment in financial tweets. In *Proceedings of the 15th NTCIR conference on evaluation of information access technologies, Tokyo, Japan*.

[8] Ya-Mien Cheng and Jheng-Long Wu. 2023. SCUNLP-1 at the NTCIR-17 FinArg-1 Task:Enhancing Classification Prediction through Feature Generation Based on ChatGPT. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001311

[9] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).

[10] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[11] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3504–3514.

[12] 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 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[13] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597* (2023).

[14] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* (2020).

[15] Supawich Jiarakul, Takenobu Tokunaga, and Hiroaki Yamada. 2023. MONETECH at the NTCIR-17 FinArg-1 Task: Layer Freezing, Data Augmentation, and Data Filtering for Argument Unit Identification. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001314

[16] Han-Chiang Kao, Hsin-Yun Hsu, and Jheng-Long Wu. 2023. SCUNLP-2 at the NTCIR-17 FinArg-1 Task: Enhancing Argumentative Relationship Recognition in the Classification Model with Language Generation Model Prompts. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001312

[17] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

[18] John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics* 45, 4 (2020), 765–818.

[19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[20] Hen-You Lin, Eugene Sy, Tzu-Cheng Peng, Shih-Hsuan Huang, and Yung-Chun Chang. 2023. TMUNLP at the NTCIR-17 FinArg-1 Task. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001286

[21] Zih An Lin, Hsiao Min Li, Adam Lin, Yun Ching Kao, Chia Shen Hsu, and Yao Chung Fan. 2023. Quack at the NTCIR-17 FinArg-1 Task : Boosting and MLM Enhanced Financial Knowledge Sequence Classification. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001305

[22] 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* (2019).

[23] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 4513–4519.

[24] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial Numeric Entity Recognition for XBRL Tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 4419–4431. https://doi.org/10.18653/v1/2022.acl-long.303

[25] James Manyika. 2023. *An overview of Bard: an early experiment with generative AI*. Technical Report. Tech. rep., Technical report, Google AI.

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[27] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[28] Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When FLUE Meets FLANG: Benchmarks and Large Pretrained Language Model for Financial Domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2322–2335. https://doi.org/10.18653/v1/2022.emnlp-main.148

[29] Shaopeng Tang and Lin Li. 2023. IDEA at the NTCIR-17 FinArg-1 Task: Argument-based Sentiment Analysis. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001276

[30] Chia-Tung Tsai, Wen-Hsuan Lian, Hsiao-Chuan Liu, Tzu-Yu Liu, Vidhya Nataraj, Mike Tian-Jian Jiang, and Min-Yuh Day. 2023. IMNTPU at the NTCIR-17 FinArg-1: Financial Argument-Based Sentiment Analysis and Argumentative Relations Identification in Social Media. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001297

[31] Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1338–1352.

[32] Mengjie Wu, Maofu Liu, and Tian Zhang. 2023. WUST at the NTCIR-17 FinArg-1 Task. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001277

[33] Shih-Hung Wu and Tsung Hsun Tsai. 2023. CYUT at the NTCIR-17 FinArg-1 Task2: A Quantitative Prompt Engineering Approach for Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001306

[34] Daichi Yamane, Fei Ding, and Xin Kang. 2023. TUA1 at NTCIR-17 FinArg-1 Task. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies*. https://doi.org/10.20736/0002001288