Term-Driven Forward-Looking Claim Synthesis in Earnings Calls

Chung-Chi Chen, Hiroya Takamura
AIST, Japan
c.c.chen@acm.org, takamura.hiroya@aist.go.jp

Abstract

Argument synthesis aims to generate rational claims, representing a fundamental objective in this field. While existing models excel in summarizing arguments and engaging in debates, we observe a critical gap in their ability to generate accurate arguments that incorporate forward-looking perspectives. In light of this observation, this paper introduces a novel task called "forward-looking claim planning." We delve into this task by exploring the efficacy of well-performing classification and generation models. Furthermore, we propose several customized preprocessing methods that yield substantial performance improvements. Through comprehensive discussion and analysis, we also outline a future research agenda for the forward-looking claim planning task.

Keywords: Argument synthesis, generation, forward-looking

1. Introduction

Arguments can be divided into claims and premises, where claims represent subjective viewpoints and premises provide supporting evidence. Upon closer examination, claims can be categorized into two types: past/present claims and forward-looking claims. For instance, given the evidence "Due to our main competitors' production capacity constraints," a past/present claim could be "our market share increased by 5% this quarter," explaining a past event. Conversely, a forward-looking claim, such as "we expect that our market share will continue to increase by 5%," offers a foresight statement regarding a possible future event. Forward-looking claims hold significant importance and find widespread use in various domains, including the Centers for Disease Control and Prevention (CDC) predicting epidemic spread and financial analysts forecasting market movements. To delve into this subject, this study introduces a forward-looking claim dataset and analyzes the characteristics of such claims.

Motivated by prior research in the financial and accounting fields (Li, 2010; Muslu et al., 2015; Bozanic et al., 2018), we observed the prevalence of forward-looking arguments in companies' formal reports and managers' discussions during earnings conference calls. These conference calls serve as quarterly meetings for companies to share operational details and address analysts' inquiries, making the information divulged during these calls valuable firsthand knowledge for investors. Recent studies have also highlighted the significance of understanding earnings conference calls (Qin and Yang, 2019; Chen et al., 2021b; Yang et al., 2022). Hence, we utilize the transcriptions of earnings conference calls as a resource in our experiments.

Foreseeing potential future events and their impacts based on existing facts is a crucial task for professionals such as managers and analysts. Table 1 presents an example of how managers, specifically Chief Financial Officers (CFOs), make forward-looking claims based on provided premises. Additionally, we showcase the generated forward-looking claims by the T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) models. Notably, the models achieve high BERTScores (Zhang et al., 2019) by successfully generating sentences containing keywords like "expect." This leads to high similarity scores between the generated results and the manager's forward-looking claim. However, we contend that the intended meanings of the manager's claim and the generated claims differ since they make claims on distinct financial terms, such as revenue, adjusted EBITDA margin, and CAPEX. Building on this observation, we propose a novel task, termed "forward-looking claim planning," aimed at learning which financial term(s) should be incorporated into a forward-looking claim given the premises.

Given Premise	
We have not changed our outlook on our expected 2021 **CAPEX** spend of $170 million or our restructuring spend of between $40 million and $45 million.	
Forward-Looking Claim	**Score**
CFO: In terms of Q2, we would expect **revenue** in Q2 to be flat to slightly up compared with Q2 of 2020, and we expect **adjusted EBITDA margin** to be approximately 11%.	-
T5: We expect our 2021 **CAPEX** spend to be between $70 million and $90 million, which is a significant increase from our prior guidance.	0.8594
BART: For the full year 2021, we now expect total **CAPEX** to be between $170 million and $180 million, up from our prior guidance of between $175 million to $200 million.	0.8555

Table 1: Examples of forward-looking claim and generation results. Score indicates BERTScore.

In this paper, we explore several customized preprocessing methods for financial terms and numerals in the input premises to address the proposed task. We conduct experiments using both classi-

LREC-COLING 2024, pages 15752–15760
20-25 May, 2024.

fication models and generation models, and our findings demonstrate significant performance improvements achievable through appropriate preprocessing techniques.

2. Related Work

Argument synthesis has emerged as a captivating research topic within our community (Hua et al., 2019; Gretz et al., 2020; Bar-Haim et al., 2020; Schiller et al., 2021). However, most prior studies have predominantly focused on data from online debate platforms or fake news clarification platforms, which primarily revolve around arguments pertaining to the past and present. Consequently, there has been limited analysis of forward-looking claims within the argument synthesis domain.

In finance, forward-looking claims by companies are heavily scrutinized by both scholars and investors. Li et al. (Li, 2010) analyzed these claims in 10-Q and 10-K reports, noting that top-performing firms often make more optimistic projections. Muslu and Ormazabal (Muslu et al., 2015) and Bozanic and Furnham (Bozanic et al., 2018) made similar observations in other report sections, emphasizing the relationship between claims and stock prices or earnings uncertainty. Yet, there's a gap in research on the planning and generation of such claims. This study initiates an exploration into training models for accurate forward-looking claim generation.

3. Dataset

Given the goal of the forward-looking claim planning task is to identify or generate the correct financial term(s) that should be included in the claim, we must separate a narrative into two components: the forward-looking claim and the premise. Subsequently, it is necessary to isolate the financial term(s) within the forward-looking claim. This section introduces the methodology employed to construct the dataset. The dataset is published under CC BY-SA 4.0 license.[1]

3.1. Dataset Construction

To construct the dataset, we initially divide the sentences within a paragraph into two components: (1) forward-looking claims and (2) premises. We observed a recurring pattern at the beginning of earnings conference calls where the host explicitly mentions the inclusion of forward-looking statements as per the regulations outlined in the securities act of 1933 and the securities exchange act of 1934. These forward-looking statements, often characterized by terms like anticipate, expect, intend, may, will, should, or their equivalents, involve inherent risks and uncertainties (Muslu et al., 2015).

We leverage this insight, along with human evaluation results, to identify forward-looking sentences effectively. By utilizing a few keywords, we can successfully distinguish forward-looking statements from others. This aligns with our intuition since companies' reports and managers' speeches are carefully prepared and regulated under the securities act. Managers are cautious with their choice of words, recognizing the potential impact on company valuation. Therefore, we employ ten seed words (future, anticipate, believe, estimate, expect, intend, predict, will, would, could) to filter out forward-looking claims.

The next step involves identifying the financial terms mentioned within the forward-looking claims. We utilize a domain-specific dictionary, specifically the Investopedia Dictionary. This dictionary encompasses a total of 6,261 financial terms. Notably, we employ the longest matching strategy, ensuring there are no mismatches between compound words and individual terms. For instance, there would be no issue in matching both "earnings" and "earnings before Interest and taxes" accurately.

3.2. Dataset Statistics

To validate the human evaluation results presented in (Muslu et al., 2015) and assess the effectiveness of the proposed annotation method, we employ the same keywords and methodology utilized in the FinNum-3 dataset (Chen et al., 2022). The FinNum-3 dataset includes 11,911 sentences annotated with binary labels indicating whether the sentence is forward-looking or not. The keyword-based method used in our evaluation achieves a precision of 74.64%. Since our objective is to identify forward-looking claims and leverage the financial terms within them, precision is a crucial metric. This indicates that the proposed method demonstrates strong performance in accurately identifying forward-looking claims.

To gather data for our experiments, we collected transcripts of earnings conference calls from SeekingAlpha. A total of 6,026 transcripts published between February 2021 and August 2021 were obtained. In these transcripts, we identified 2,037 unique financial terms. To address sparsity concerns, we selected the top-50 financial terms as labels. These top-50 financial terms appear more than 500 times within the dataset. We focused on paragraphs containing forward-looking claims that include one or more of the top-50 financial terms. Consequently, we obtained 34,933 paragraphs that encompass forward-looking claims and associated premises. To create training and testing sets, we allocated 80% of the paragraphs for training and reserved the remaining 20% for testing purposes.

[1] Dataset: http://flcp.nlpfin.com/

Preprocessing	Example	BERT		LinkBERT		FinBERT	
		Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Original	Tesla received 14,335 orders in Europe	13.33%	6.67%	14.09%	7.15%	17.31%	10.73%
FinTerm Only	orders	16.67%	9.84%	14.79%	8.34%	16.55%	11.19%
Emph FinTerm	Tesla received 14,335 [FIN] orders [FIN] in Europe	13.71%	8.77%	14.33%	6.89%	17.33%	10.80%
Num Tag	Tesla received [NUM] orders in Europe	12.41%	6.34%	10.36%	4.99%	16.59%	10.32%
Emph Num	Tesla received [NUM] 14,335 [NUM] orders in Europe	14.24%	7.10%	14.08%	7.38%	17.98%	10.98%
Sci-1	Tesla received 1.4 [EXP] 4 orders in Europe	15.43%	8.70%	13.18%	6.65%	17.15%	10.36%
Attach	orders | 14,335	14.49%	7.91%	13.06%	7.83%	16.91%	11.97%
Emph FinTerm + Emph Num	Tesla received [NUM] 14,335 [NUM] [FIN] orders [FIN] in Europe	13.44%	8.18%	14.56%	8.28%	15.35%	9.27%
FinTerm Only + Emph Num	Tesla received [NUM] 14,335 [NUM] orders in Europe [SEP] orders	**22.31%**	**13.83%**	18.06%	11.05%	20.83%	13.56%
FinTerm Only + Sci-1	Tesla received 1.4 [EXP] 4 orders in Europe [SEP] orders	19.67%	12.44%	17.04%	11.17%	22.28%	14.79%
Emph Num + Attach	Tesla received [NUM] 14,335 [NUM] orders in Europe [SEP] orders | 14,335	19.93%	11.88%	**18.19%**	**11.46%**	20.53%	13.52%
Sci-1 + Attach	Tesla received 1.4 [EXP] 4 orders in Europe [SEP] orders | 14,335	20.37%	11.78%	16.31%	9.72%	**22.94%**	**15.60%**

Table 2: Experimental results of classification models. The bold results are the best-performance among all preprocessing methods.

4. Experiment

4.1. Preprocessing Methods

In this paper, our focus lies in investigating the impact of proper preprocessing on improving performance in the proposed task. Table 2 presents a compilation of preprocessing methods and corresponding examples. The baseline method (*Original*) serves as a reference point, representing the absence of any preprocessing.

Recognizing the significance of financial terms in financial causality, we introduce two preprocessing methods specifically tailored to handle these terms within the input premise. Firstly, we employ the *FinTerm Only* method, which retains only the financial terms in the premise as input, disregarding the remaining content. Additionally, we propose the *Emph FinTerm* method, wherein we emphasize financial terms by incorporating [FIN] tags before and after them.

Drawing inspiration from the work of Yang et al. (2022), who demonstrated the importance of numerals in financial decision making within earnings conference calls, we devise preprocessing methods for numerals. The first approach (*Num Tag*) involves replacing numerals with a commonly used tag, [NUM]. The second method (*Emph Num*) emphasizes numerals by incorporating [NUM] tags before and after each numeral. The third technique (*Sci-1*) converts numerals to scientific notation, rounded to one decimal place.

To consolidate the representation of financial terms and numerals within the same sentences, we adopt the linearized sequence format proposed by Yin et al. (2020) (*Attach*). This format presents the financial terms and linked numerals in the "[Financial Term 1] | [Linked Numeral 1] | [Linked Numeral 2]" structure. In cases where multiple financial terms exist, we follow Yin et al. (2020) in separating them using the "[SEP]" marker.

4.2. Experimental Results

We conduct experiments using both classification models (BERT (Devlin et al., 2019), LinkBERT (Yasunaga et al., 2022), and FinBERT (Huang et al., 2020)) and generation models (T5 (Raffel et al., 2020) and BART (Lewis et al., 2020)). Given that forward-looking claims may involve multiple financial terms, our task is formulated as a multi-label classification setting for classification models. We evaluate the classification results using Micro-F1 and Macro-F1 metrics. For the generation models, we assess the results based on two criteria: (1) the ratio of all target financial terms appearing in the generated claim (ALL), and (2) the ratio of at least one financial term appearing in the generated claim (LEAST). Notably, LinkBERT represents the state-of-the-art language model incorporating multi-hop knowledge during pre-training, while FinBERT is a domain-specific language model trained on financial documents, specifically earnings conference calls published from 2003 to 2020. It is important to highlight that FinBERT's training data does not overlap with the proposed dataset, which utilizes data from 2021.

Preprocessing	T5		BART	
	ALL	LEAST	ALL	LEAST
Original	18.12%	26.16%	19.16%	27.87%
FinTerm Only + Emph Num	**19.73%**	**28.45%**	**20.01%**	**28.68%**
Emph Num + Attach	19.59%	28.37%	19.52%	28.25%
Sci-1 + Attach	19.31%	27.61%	19.40%	27.95%

Table 3: Experimental results of generation models.

Table 2 presents the results obtained from the classification models. Firstly, in the vanilla setting, FinBERT exhibits superior performance compared to LinkBERT and BERT, while LinkBERT outperforms BERT. Secondly, the outcomes of the *FinTerm Only* experiment underscore the significance of financial terms in the proposed task. Thirdly, the results of the *Num Tag* experiment suggest that numerals indeed contribute valuable information to financial document understanding, as all models perform worse when numerals are replaced with tags. Lastly, by combining the aforementioned methods, we aim to identify the optimal approach for the proposed task. Notably, almost all models demonstrate improved performance compared to the baseline. Moreover, with appropriate preprocessing methods, the general language model (BERT) achieves performance comparable to the

Manager (GT)	We expect investment spend to pick up in the second half to circa $700 million to $750 million.
T5	We expect to continue to see a significant impact on our D&A and investment spend in the second half of '21.
BART	For the full year '21, we expect our technology and investment spend to be in the range of $40 million to $50 million, which is up from our previous guidance of $35 million to 45 million.

Table 4: Case Study. GT denotes ground truth.

	Appear in Input		Not Appear in Input	
Generation	ALL	LEAST	ALL	LEAST
T5	36.50%	52.54%	6.39%	9.30%
BART	33.99%	49.11%	8.89%	12.44%
Classification	Micro-F1	Macro-F1	Micro-F1	Macro-F1
BERT	22.67%	13.55%	22.01%	13.96%
LinkBERT	18.67%	11.86%	17.79%	11.08%
FinBERT	23.43%	16.22%	22.55%	15.06%

Table 5: Analysis of the experimental results.

	GPT-4	Human
Original	25.0%	22.5%
FinTerm Only + Emph Num	42.5%	37.5%
Same	7.5%	12.5%

Table 6: Human and GPT-4 evaluation.

domain-specific language model (FinBERT).

In Table 3, we present the results obtained from the generation models. We leverage the best representation methods identified from the classification results and observe that *FinTerm Only + Emph Num*, which demonstrated superior performance with BERT, also yields the best results with T5 and BART. This observation aligns with our interpretation that since the encoders of T5 and BART are transformer-based, akin to BERT, the optimal method for BERT is expected to be effective for these models as well.

5. Discussion

5.1. Case Study

Table 4 provides an example to discuss the generated forward-looking claims. First, models succeed in learning the template for making forward-looking claims. The generated sentences are often included "believe" or "expect". Second, the generated sentences are fluent and seem correct. However, when taken a close look, there exists some hallucination. For example, in the generated claim of BART, there are some incorrect descriptions like (1) the wrong time inference (For the full year '21), (2) wrong monetary range, and (3) hallucination on previous guidance. These are all interesting directions in the forward-looking claim generation task. Additionally, this analysis also points out that numeral information such as time and monetary terms are important in financial narratives.

5.2. Copy vs. Inference

When examining the generation results, we observed a tendency of models to predominantly replicate the financial terms present in the input, resulting in lower performance when the target financial term is absent from the input. Within the test set, we found 3,095 instances where at least one of the target financial terms appeared in the given premise, and 3,892 instances where the target financial terms were not present in the given premise. Remarkably, this implies that over 50% of forward-looking claims cannot be directly copied from the financial terms provided in the premises. The analysis of this phenomenon is presented in Table 5. These findings indicate that while current-generation models excel in summarization, there is still room for improvement in generating forward-looking claims based on the given premises. In contrast, the classification models exhibit similar performance in both scenarios.

5.3. Human Evaluation

We further conducted human evaluations and utilized GPT-4 for assessing and comparing the generated texts. The question posed to human annotators was which text is closer to the ground truth or, if difficult to discern, which presents a better claim based on the given premise. Similarly, we asked GPT-4 to determine which text constitutes a better claim based on the premise. Our comparison focused on BART with *Original* preprocessing versus BART with *FinTerm Only + Emph Num* preprocessing. We annotated 40 instances, and the statistics are presented in Table 6. In certain instances, both options were similarly incorrect in various aspects, leading to the "Same" row indicating these occurrences. Overall, both human evaluators and GPT-4 showed a preference for the forward-looking claims generated with *FinTerm Only + Emph Num* preprocessing. This preprocessing approach enables BART to produce a more detailed narrative, significantly influencing both human and GPT-4 preferences towards this text. Furthermore, we observed that models tend to segment the given

Preprocessing	T5			BART		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Original	0.2476	0.0590	0.1866	0.2516	0.0613	0.1892
FinTerm Only + Emph Num	**0.2518**	**0.0619**	**0.1901**	0.2526	0.0649	0.1895
Emph Num + Attach	**0.2518**	0.0617	0.1892	**0.2527**	**0.0653**	**0.1899**
Sci-1 + Attach	0.2514	0.0618	0.1892	0.2510	**0.0653**	0.1892

Table 7: ROUGE Metric.

premises, which may explain the observed increase in performance when using copying, as discussed previously.

5.4. Evaluating Generation Results with ROUGE Metric

ROUGE score is often used for evaluating the generation results. Table 7 provides the evaluation results. We find that the generated sentences after fine-tuning T5 and BART are only slightly different. Because the proposed task aims at generating correct financial terms in the forward-looking claims, we focus on the proposed ALL and LEAST metrics in the discussions. However, the results in Table 7 still show that the preprocessing methods are also helpful in improving performances under the ROUGE metric.

5.5. Application Scenario

Our goal is to develop a system that could help professionals in speech preparation. For example, one objective of this research is to assist managers in drafting their speeches. Given the impact of these speeches on the market (Qin and Yang, 2019; Koval et al., 2023; Mukherjee et al., 2022), meticulous planning is essential. This study represents an initial step in speech preparation. Future work will aim to control the tone and professionalism, enhancing the alignment of the generated speech scripts with real-world professional scenarios. A further research direction is whether the manager's forward-looking claim will raise professional analysts' doubts or questions. Although there exists a work attempt to generate questions based on the earning call transcript (Juan et al., 2023), the discussion of leveraging generated questions to refine the speech transcript has not yet been discussed. Future studies can simulate the interaction between the speaker and the audience to assist the speaker in speech preparation. We believe that the forward-looking claim planning in this paper would be one of the important points for this application scenario.

On the other hand, the generation of forward-looking claims plays a crucial role in the task of stock research report generation (Chen et al., 2021a; Yan, 2022). Future research could employ the preprocessing methodologies delineated in this study to forge more logically coherent causal connections between the provided premises and forward-looking claims.

5.6. Ethical Note

Argument synthesis and generation raise concerns about their potential misuse (Solaiman et al., 2019). Given the significant impact of managers' forward-looking claims on financial decision-making (Qin and Yang, 2019; Chen et al., 2021b; Yang et al., 2022), the generation of claims that mix false and genuine information can influence the financial market. We have observed that current generation models are capable of generating forward-looking claims with fluency and seemingly plausible narratives. The following example illustrates this:

- T5 (Generated): We anticipate a continued sequential decrease in gross margin for the second half of the year.
- Manager (Actual): We expect this action to have a cost offsetting impact and thus **benefit** gross margins in future quarters.

This example highlights that existing models can produce plausible yet incorrect forward-looking claims. Our findings show both classification and generation models need refinement. This amplifies future challenges for fact-checking. To mitigate these challenges, understanding the strengths and weaknesses of current models is key. We aim to encourage deeper exploration into forward-looking claim generation, potentially aiding false inference detection and fact verification by scrutinizing claim consistency with given premises.

6. Conclusion

This paper introduces a novel task, namely forward-looking claim planning, and conducts a preliminary investigation employing both classification and generation models. While preprocessing may be considered a well-established topic, our findings demonstrate its continued relevance in effectively providing prompts to models. Additionally, we highlight the limitations of existing high-performing models when applied to the proposed task. By establishing this new research direction, we aim to stimulate future investigations into forward-looking argument generation tasks.

7. Acknowledgement

This work 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).

8. Bibliographical References

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Zahn Bozanic, Darren T Roulstone, and Andrew Van Buskirk. 2018. Management earnings forecasts and other forward-looking statements. *Journal of Accounting and Economics*, 65(1):1–20.

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 & Knowledge Management*, pages 2920–2924.

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 (NTCIR-16)*.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. The workweek is the best time to start a family–a study of GPT-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 528–544.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Allen Huang, Hui Wang, and Yi Yang. 2020. FinBERT—a deep learning approach to extracting textual information. *Available at SSRN 3910214*.

Yining Juan, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Generating multiple questions from presentation transcripts: A pilot study on earnings conference calls. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 449–454, Prague, Czechia. Association for Computational Linguistics.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2023. Forecasting earnings surprises from conference call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8197–8209, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Feng Li. 2010. The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Volkan Muslu, Suresh Radhakrishnan, KR Subramanyam, and Dongkuk Lim. 2015. Forward-looking MD&A disclosures and the information environment. *Management Science*, 61(5):931–948.

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.

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*, 21:1–67.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Sixing Yan. 2022. Disentangled variational topic inference for topic-accurate financial report generation. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 18–24, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Linyi Yang, Jiazheng Li, Ruihai Dong, Yue Zhang, and Barry Smyth. 2022. NumHTML: Numeric-oriented hierarchical transformer model for multi-task financial forecasting.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

9. Appendix

9.1. Implement Detail

We experiment with the transformers package proposed by Hugging Face. [2] The hardware for experiments is Intel Xeon Gold CPU and Nvidia Tesla V100 w/32GB. Table 8 sorts out the details of the models used in our experiments and the links to Hugging Face model pages.

9.2. Top-50 Financial Terms

Table 9 provides the statistics of the top-50 financial terms in the proposed dataset.

[2] https://huggingface.co/docs/transformers/index

	URL
BERT (Devlin et al., 2019)	https://huggingface.co/bert-base-uncased
LinkBERT (Yasunaga et al., 2022)	https://huggingface.co/michiyasunaga/LinkBERT-base
FinBERT (Huang et al., 2020)	https://huggingface.co/yiyanghkust/finbert-tone
T5 (Raffel et al., 2020)	https://huggingface.co/t5-base
BART (Lewis et al., 2020)	https://huggingface.co/facebook/bart-base

Table 8: Reference for the models in our experiments.

Rank	Financial Term	Proportion	Accumulation
1	business	9.81%	9.81%
2	revenue	7.80%	17.61%
3	market	5.69%	23.29%
4	range	4.18%	27.47%
5	demand	3.70%	31.17%
6	guidance	3.22%	34.39%
7	customer	3.14%	37.53%
8	capital	3.00%	40.53%
9	support	2.99%	43.52%
10	gaap	2.90%	46.42%
11	basis	2.61%	49.03%
12	value	2.27%	51.30%
13	investment	2.26%	53.56%
14	industry	2.10%	55.65%
15	return	2.03%	57.69%
16	earnings	2.00%	59.68%
17	adjusted ebitda	1.92%	61.60%
18	acquisition	1.83%	63.43%
19	risk	1.69%	65.12%
20	year-over-year	1.62%	66.74%
21	expansion	1.60%	68.34%
22	commercial	1.43%	69.77%
23	margin	1.40%	71.18%
24	balance sheet	1.38%	72.56%
25	leverage	1.37%	73.93%
26	free cash flow	1.35%	75.28%
27	sec	1.30%	76.58%
28	debt	1.24%	77.83%
29	momentum	1.22%	79.05%
30	volume	1.17%	80.22%
31	fiscal year	1.15%	81.37%
32	trend	1.14%	82.51%
33	transaction	1.11%	83.61%
34	expense	1.10%	84.72%
35	supply	1.10%	85.82%
36	brand	1.05%	86.86%
37	gross margin	1.02%	87.89%
38	ebitda	1.01%	88.89%
39	infrastructure	1.01%	89.90%
40	inventory	1.00%	90.90%
41	supply chain	1.00%	91.90%
42	order	0.98%	92.88%
43	marketing	0.98%	93.86%
44	offset	0.97%	94.83%
45	liquidity	0.90%	95.73%
46	cash flow	0.88%	96.61%
47	interest	0.88%	97.49%
48	manufacturing	0.87%	98.36%
49	loan	0.82%	99.19%
50	asset	0.81%	100.00%

Table 9: Statistics of top-50 financial terms.