Demystifying Financial Texts Using Natural Language Processing

Sohom Ghosh

August 1, 2024

Demystifying Financial Texts Using Natural Language Processing

Sohom Ghosh*
Jadavpur University
Department of Computer Science and Engineering
Kolkata, India
sohom1ghosh@gmail.com

ABSTRACT

Human beings aspire for a better life. Financial well-being enables this. However, lack of financial literacy, ever-growing wealth inequality, and persuading illicit information floating in social media inhibit one's progress towards a good fortune. In this paper, we discuss four pillars where Natural Language Processing can help improve financial literacy, reduce wealth disparity, ensure a sustainable future, and economic prosperity. These pillars are: Inclusive investing, Improved investing, Impactful (green) investing, and Informed investing. Additionally, we focus to specifically cater to the Indian market (Indic investing) and present several resources to enhance comprehensibility of financial texts. Inclusive investing deals with enhancing the readability and reachability of financial texts. Improved investing addresses the need to simplify investors' journey by providing them with hypernyms and relations between entities. Impactful investing is associated with focusing on sustainable pathways. Improved investing is about eradicating finance related misinformation from social media, like evaluating trustworthiness of posts by executives, detecting in-claim and exaggerated numerals, etc. In most cases, we are able to demonstrate the efficacies of our approaches by benchmarking them with existing state-of-the-art methods.

CCS CONCEPTS

• **Applied computing → Economics**; • **Information systems → Clustering and classification**; **Social networks**; *Information retrieval*; • **Computing methodologies → Information extraction**; *Language resources*; *Lexical semantics*; *Information extraction*; Ensemble methods.

KEYWORDS

Financial Natural Language Processing, Financial Texts, Text comprehension, Green Investing, Inclusive Investing, Indic Texts

1 INTRODUCTION

In today's world, to address the ever-growing rich-poor divide, it is essential to focus on financial literacy. Financial literacy is the art of managing money. The financial literacy rate for developed nations (like the United States, the United Kingdom) stands between 51 to 60%[1] whereas for the developing nations (like India) it is less than 30%.[2] Financial literacy leads to financial well-being, which in turns results in economic prosperity of the nation. To address this, we have been working to improve the investment process and make it more inclusive. We refer to this as **Inclusive** and **Improved investing**. Currently, we have worked on enhancing the readability and reachability of financial texts [26], [27]. Furthermore, we worked to improve the investment journey by detecting hypernyms and relations between entities [11], [14], [28].

Various nations across the globe have pledged to be carbon-neutral by 2050.[3] Thus, investors are looking for sustainable avenues for investing their money. They are keen to understand the Environmental, Social, and Corporate Governance (ESG) aspects of funds. To address this, we use natural language processing (NLP) to analyse financial texts related to ESG and sustainability [21], [39], [40]. We refer to this as **Impactful Investing**.

Persuading posts by financial influencers[4], disrupts the stock market adversely. Executives also try to reap benefits leveraging the social media platforms. Thus, we work on identifying potential miss-information and prevent its spread [17], [16], [20], [18], [38], [19], [4]. This is called **Informed Investing**.

For a nation like India, where the top 10% of citizens controls 65% of the total wealth of the nation,[5] it is of utmost importance to address the needs of the bottom of the economic human pyramid. While most previous research focusses on analysing financial texts in English, we have been working to improve the comprehensibility of financial texts in various Indian languages [37]. Moreover, we have also worked on ESG and numeracy related tasks in Hindi, Bengali, and Telugu. We refer to this as **Indic Investing**.

Additionally, we open-source several tools[6] [7] to analyse financial texts in different languages [23].

2 PROBLEMS

In this section, we present the tasks we have been focussing along with the corresponding publications.

2.1 Inclusive Investing

Task-1: Given a financial text (FT), we want to assess its readability and simplify it. ([27])

*supervised by Prof. (Dr.) Sudip Kumar Naskar

[1] https://blogs.illinois.edu/view/7550/558591870

[2] https://yourstory.com/2023/07/financial-literacy-is-key-to-unlocking-india-economy (all URLS are accessed on 18th Jan 2024)

[3] https://www.un.org/en/climatechange/net-zero-coalition

[4] https://economictimes.indiatimes.com/markets/stocks/news/finfluencer-mess-assessing-the-need-for-sebi-intervention/articleshow/105551155.cms

[5] https://www.livemint.com/economy/india-among-top-countries-with-high-income-wealth-inequality-undp-report-11699284168538.html

[6] https://huggingface.co/spaces/sohomghosh/FENCE_Financial_Exaggerated_Numeral_ClassifiEr

[7] https://huggingface.co/spaces/sohomghosh/FiNCAT_Financial_Numeral_Claim_Analysis_Tool

Task-2: Given two FTs, we want to assess which one would to reach more people. ([13])

2.2 Improved Investing

Task-3: Given a financial jargon in a FT, we would like to retrieve its hypernym. ([11], [14])
Task-4: Given two entities in a FT, we would like to determine the relationship between them. ([28])

2.3 Impactful Investing

Task-5: Classify a FT as Sustainable or Unsustainable ([21])
Task-6: Detect ESG Issues from FTs in English. ([39])
Task-7: Identify ESG impact type & duration from FTs. ([40], [3])

2.4 Informed Investing

Task-8: Detect exaggerated and in-claim numerals from FTs. ([16], [20], [22])
Task-9: Evaluate the Rationals of Amateur Investors. [19]
Task-10: Evaluate the trustworthiness of Social Media Posts by Executives on Stock Prices ([38])
Task-11: Fine-grained Argument Understanding in FTs ([4])

2.5 Indic Investing

Task-12: Financial Argument Analysis in Bengali ([37])
Task-13: Extract ESG Issues, Assess Sustainability, and Detect exaggerated numerals from FTs in Hindi, Bengali, & Telugu ([15])

2.6 Tools for FinNLP

Task-14: Develop tools for processing FTs ([23], [18], [17], [26])

3 METHODOLOGY AND RESULTS

Firstly, we explored the existing works [24], participated in various FiNLP shared tasks and presented our preliminary findings [25].

For **Task-1**, we proposed a new dataset Financial Readability Assessment Dataset (**FinRAD**) comprising 13,000+ definitions of financial terms for measuring readability. Subsequently, we released a fine-tuned version of FinBERT [2] and a tool, FinRead [26]. We added a paraphraser on top of it and created the Financial Language Simplifier (FinLanSer) tool.[8] For **Task-2**, we collected a set of tweet pairs that were related to finance, and each tweet in a pair was similar to the other. We leveraged the reasoning power of Large Language Models (LLMs) with the discriminative power of pre-trained encoder models (RoBERTa [32]) to determine which of the tweets in a give pair will receive more re-tweets.

Task-3 deals with hypernym detection. It is the third shared task on learning semantic similarities for the financial domain [30]. We ranked third in this shared task [11]. Subsequently, we refined our approach by ensembling results from two sentence transformer [35] models to achieve state-of-the-art (SOTA) results [14]. For **Task 4**, we used the REFinD dataset [31]. It deals with extracting relationship between financial entities. We proposed the Mask One At a Time (MOAT) framework and benchmarked its performance with that of LLMs [28] (Falcon, Dolly, MPT and llama-2).

[8]https://youtu.be/YcHJliaSyuY (accessed on 24$^{\text{th}}$ Jan 2024)

We participated in several shared tasks related to ESG and sustainability. The datasets of **Tasks 5,6, and 7** are obtained by participating in shared tasks [29], [8], [9] & [10]respectively. For **Task-5, 6**, we fine-tuned a RoBERTa [32] model and a SEC-BERT [33] model respectively for classification [21], [39]. For **Task-7**, we extracted ESG impact types & predicted duration of impacts from FTs in English, French, Japanese, Chinese, & Korean. We outperformed others for Japanese, Chinese [40] & French datasets [3]. We enriched the datasets through translation & paraphrasing.

For detecting in-claim numerals **Task-8**, we ensemble outputs from models [16] created by fine-tuning FinBERT [2] and BERT [12]. This dataset and task was first proposed by [5]. To determine exaggerated numerals, we propose the Financial Exaggerated Numeral ClassifiEr (FENCE) [22] tool. **Task-9** is about estimating if one financial opinion will lead to more profit or loss than the other [6]. We ensemble two systems [19] created using FinBERT [2] and SBERT Chinese.For **Task-10**, we used a Gated Recurrent Unit model to investigate whether tweets by executives have more influence than that of the public on the closing price of various stocks [38]. **Task-11** is one of the shared tasks of NTCIR-17 [7]. We fine-tuned FinBERT [2] and SEC-BERT [33] using the cross encoder architecture [35] to determine the relationship between argumentative FTs in English and Chinese, respectively.

To specifically focus on Indian languages, we started by analysing argumentative texts in Bengali (**Task-12**) [37]. The first task was to classify a FT in Bengali as ‘Premise’ or ‘Claim’. The second task was to classify the relationship between two FTs in Bengali as ‘Support’, ‘Attack’, or ‘No Relation’. For both tasks, the fine-tuning of multilingual BERT (MBERT) [36] gave us the best performance. Furthermore, we analyse the financial budget speeches of different states of India (**Task-13**). As of now, we proposed three tasks: exaggerated numeral, ESG issue, and sustainability detection in Hindi, Bengali, and Telugu. We fine-tuned MBERT and leveraged the AIforBharat machine translation system [34].

To summarise, we have worked on financial texts in seven different languages (English, French, Japanese, Chinese, Hindi, Bengali, and Telugu) and proposed seven new FinNLP datasets. Of these, three datasets have been published ([27], [38], [37]) and four of them are under review. Finally, to improve the usability of the proposed solutions, we created several tools to analyse financial texts (**Task-14**) using gradio [1]. These include: Financial Language Understandability Enhancement Toolkit (FLUEnT) [23], FinRead [26] , Financial Numeral Claim Analysis Tool (FinCAT) [18], FinCAT-2 [17], and Financial Argument Analysis in Bengali (FAAB) [37]. In Table 1, we present all our contributions. Furthermore, for each task, we mention the metric for evaluation, briefly describe our approach, benchmark our approach with that of SOTA, and present the performance numbers. Additionally, we state if the dataset being used for a task is new and the language of this dataset. Lastly, we mention whether we have created any new tool for the given task and showcase our publications.

4 CONCLUSION AND FUTURE WORK

In this paper, we presented our contributions in the field of FinNLP till date. We developed several models, created and released different datasets and open-sourced several tools for mining financial texts.

Table 1: Approaches and results for different tasks.
AU-ROC = Area under the ROC curve, Acc. = Accuracy, MPP = Maximum Possible Profit, ML = Maximum Loss, MAPE = Mean Absolute Percentage Error, NA = Not Applicable, SOTA = State of the Art, LLM = Large Language Model, PLM = Pre-trained Language Model, Trans-Prp = Translate Paraphrase, IT = Impact Type, ID = Impact Duration

Task #	Metric	Approach Summary	SOTA	Performance	New Data	Language	New Tool	Publication(s)
1	AU-ROC	FinBERT finetune	Yes	0.993	Yes	English	Yes	[26], [26]
2	F1	RoBERTa + Claude (LLM)	Yes	0.731	Yes	English	No	[13]
3	Acc.	SBERT finetune	Yes	0.967	No	English	No	[14], [11]
4	F1	SEC-BERT + Neural Network	No	0.736	No	English	No	[28]
5	Acc.	RoBERTa finetune	No	0.932	No	English	No	[21]
6	F1	SEC-BERT finetune	No	0.715	No	English	Yes	[39]
7	F1	FinBERT finetune	No	0.929 (IT)	No	English	No	[40]
7	F1	Trans-Prp + FinBERT finetune	No	0.756 (IT)	No	French	No	[40]
7	F1	Trans-Prp + FinBERT finetune	Yes	0.679 (IT)	No	Japanese	No	[40]
7	F1	Trans-Prp + FinBERT finetune	Yes	0.677 (IT)	No	Chinese	No	[40]
7	F1	Trans-Prp + PLM finetune	No	0.5882 (ID)	No	English	No	[3]
7	F1	Trans-Prp + PLM finetune	Yes	0.5616 (ID)	No	French	No	[3]
8	F1	Ensemble (FinBERT, BERT + Logistic Regression)	No	0.948	No	English	Yes	[16], [20], [18], [17]
9	MPP, ML	SBERT Chinese + Classifier, FinBERT	No	0.575 (MPP), 0.598 (ML)	No	Chinese	No	[19]
10	MAPE	Gated Recurrent Unit	Yes	0.382	Yes	English	Yes	[38]
11	F1	Cross Encoder (FinBERT Finetuned)	No	0.789	No	English	No	[4]
11	F1	Translate + Cross Encoder (SEC-BERT)	No	0.641	No	Chinese	No	[4]
12	F1	MBERT, Cross Encoder (MBERT)	No	0.721 (1st task), 0.755 (2nd Task)	Yes	Bengali	Yes	[37]
13	F1	MBERT+Classifier, Translate + RoBERTa, Translate+MBERT	Yes	0.680 (1st task), 0.950 (2nd task), 0.590 (3rd task)	Yes	Hindi	No	[15]
13	F1	MBERT+Classifier, Translate + RoBERTa, Translate+MBERT	Yes	0.650 (1st task), 0.920 (2nd task), 0.550 (3rd task)	Yes	Bengali	No	[15]
13	F1	MBERT+Classifier, Translate + RoBERTa, Translate+MBERT	Yes	0.680 (1st task), 0.920 (2nd task), 0.580 (3rd task)	Yes	Telugu	No	[15]
14	NA	Gradio (frontend)	NA	NA	NA	Various	Yes	[23], [18], [17], [26]

Furthermore, we share our findings from participating in various FinNLP shared tasks. While we could achieve SOTA performance in most cases, there is further scope for improvement specifically for low resource languages.

In future, we would like to embrace multi-modality and focus on low resource Indic languages. We aspire to create India specific multilingual knowledge graphs and work on improving the comprehensibility of financial texts in Indic languages. Subsequently, we would like to simplify and summarize the different financial documents (like economic reviews, budgets, etc.) which are released to cater the interests of masses. We would like to extensively explore if NLP can be leveraged to predict the outcome of Initial Public Offerings.

REFERENCES

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. https://doi.org/10.48550/arXiv.1906.02569

[2] Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063 [cs.CL]

[3] Neelabha Banerjee, Anubhav Sarkar, Swagata Chakraborty, Sohom Ghosh, and Sudip Kumar Naskar. 2024. Fine-tuning Language Models for Predicting the Impact of Events Associated to Financial News Articles. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing @ LREC-COLING 2024*. Torino, Italia.

[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*. NII, Tokyo, Japan, 29–36. 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*. NII, Tokyo, Japan, 87–91.

[6] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2022. Overview of the FinNLP-2022 ERAI Task: Evaluating the Rationales of Amateur Investors. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. ACL, Abu Dhabi, UAE (Hybrid).

[7] Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2023. Overview of the NTCIR-17 FinArg-1 Task: Fine-Grained Argument Understanding in Financial

Analysis. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan*. NII, Tokyo, Japan, 16–20.

[8] Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-Lingual ESG Issue Identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, Chung-Chi Chen, Hiroya Takamura, Puneet Mathur, Remit Sawhney, Hen-Hsen Huang, and Hsin-Hsi Chen (Eds.). -, Macao.

[9] Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-Lingual ESG Impact Type Identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

[10] Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anais Lhuissier, Yohei Seki, Hanwool Lee, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-Lingual ESG Impact Duration Inference. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*. ELRA and ICCL, Torino, Italia, 219–227.

[11] Ankush Chopra and Sohom Ghosh. 2021. Term Expansion and FinBERT fine-tuning for Hypernym and Synonym Ranking of Financial Terms. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*. -, Online, 46–51. https://aclanthology.org/2021.finnlp-1.8

[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)*. ACL, Minneapolis, Minnesota, 4171–4186.

[13] Sohom Ghosh, Chung-Chi Chen, and Sudip Kumar Naskar. 2024. Generator-Guided Crowd Reaction Assessment. In *Companion Proceedings of the ACM on Web Conference 2024* (Singapore) *(WWW '24)*. New York, NY, USA, 597–600.

[14] Sohom Ghosh, Ankush Chopra, and Sudip Kumar Naskar. 2023. Learning to Rank Hypernyms of Financial Terms Using Semantic Textual Similarity. *SN Computer Science* 4, 5 (2023), 610. https://doi.org/10.1007/s42979-023-02134-z

[15] Sohom Ghosh, Arnab Maji, Aswartha Narayana, and Sudip Kumar Naskar. 2024. IndicFinNLP: Financial Natural Language Processing for Indian Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia.

[16] Sohom Ghosh and Sudip Kumar Naskar. 2022. Detecting context-based in-claim numerals in Financial Earnings Conference Calls. *International Journal of Information Technology* 14 (2022), 2559–2566.

[17] Sohom Ghosh and Sudip Kumar Naskar. 2022. FiNCAT-2: An enhanced Financial Numeral Claim Analysis Tool. *Software Impacts* 12 (2022), 100288.

[18] Sohom Ghosh and Sudip Kumar Naskar. 2022. FiNCAT: Financial Numeral Claim Analysis Tool. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion)* (Virtual Event, Lyon, France). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3487553.3524635

[19] Sohom Ghosh and Sudip Kumar Naskar. 2022. LIPI at the FinNLP-2022 ERAI Task: Ensembling Sentence Transformers for Assessing Maximum Possible Profit and Loss from Online Financial Posts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. ACL, Abu Dhabi, UAE (Hybrid).

[20] 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*. NII, Tokyo, Japan.

[21] Sohom Ghosh and Sudip Kumar Naskar. 2022. Ranking Environment, Social And Governance Related Concepts And Assessing Sustainability Aspect of Financial Texts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. ACL, Abu Dhabi, UAE (Hybrid).

[22] Sohom Ghosh and Sudip Kumar Naskar. 2023. FENCE: Financial Exaggerated Numeral ClassifiEr. https://easychair.org/publications/preprint_download/s9Ds

[23] Sohom Ghosh and Sudip Kumar Naskar. 2023. FLUEnT: Financial Language Understandability Enhancement Toolkit. In *6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD) (CODS-COMAD 2023)* (Mumbai, India). Association for Computing Machinery, New York, NY, USA, In press. https://doi.org/10.1145/3570991.3571067

[24] Sohom Ghosh and Sudip Kumar Naskar. 2023. Recent trends in financial natural language processing research. *Science Talks* 8 (2023), 100270.

[25] Sohom Ghosh and Sudip Kumar Naskar. 2023. Using Natural Language Processing to Enhance Understandability of Financial Texts. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)* (Mumbai, India) *(CODS-COMAD '23)*. Association for Computing Machinery, New York, NY, USA, 301–302.

[26] Sohom Ghosh, Shovon Sengupta, Sudip Naskar, and Sunny Kumar Singh. 2021. FinRead: A Transfer Learning Based Tool to Assess Readability of Definitions of Financial Terms. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*. NLP Association of India (NLPAI), Silchar, India.

[27] Sohom Ghosh, Shovon Sengupta, Sudip Kumar Naskar, and Sunny Kumar Singh. 2022. FinRAD: Financial Readability Assessment Dataset - 13,000+ Definitions of Financial Terms for Measuring Readability. In *Proceedings of the The 4th Financial Narrative Processing Workshop (FNP@LREC2022)*. European Language Resources Association, Marseille, France, 1–9.

[28] Sohom Ghosh, Sachin Umrao, Chung-Chi Chen, and Sudip Kumar Naskar. 2024. The Mask One At a Time Framework for Detecting the Relationship between Financial Entities. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation* (Panjim, India) *(FIRE '23)*. Association for Computing Machinery, New York, NY, USA, 40–43. https://doi.org/10.1145/3632754.3632756

[29] Juyeon Kang and Ismail El Maarouf. 2022. FinSim4-ESG Shared Task: Learning Semantic Similarities for the Financial Domain. Extended edition to ESG insights. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. ACL, Abu Dhabi, United Arab Emirates (Hybrid).

[30] Juyeon Kang, Ismail El Maarouf, Sandra Bellato, and Mei Gan. 2021. FinSim-3: The 3rd Shared Task on Learning Semantic Similarities for the Financial Domain. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*. -, Online, 31–35. https://aclanthology.org/2021.finnlp-1.5

[31] Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. REFinD: Relation Extraction Financial Dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 3054–3063. https://doi.org/10.1145/3539618.3591911

[32] 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.

[33] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial Numeric Entity Recognition for XBRL Tagging. ACL, Dublin, Ireland.

[34] Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics* 10 (02 2022), 145–162.

[35] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). ACL, Hong Kong, China, 3982–3992.

[36] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). ACL, Online, 4512–4525.

[37] Rima Roy, Sohom Ghosh, and Sudip Kumar Naskar. 2024. Financial Argument Analysis in Bengali. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation* (Panjim, India) *(FIRE '23)*. Association for Computing Machinery, New York, NY, USA, 88–92.

[38] Anubhav Sarkar, Swagata Chakraborty, Sohom Ghosh, and Sudip Kumar Naskar. 2023. Evaluating Impact of Social Media Posts by Executives on Stock Prices. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation* (Kolkata, India) *(FIRE '22)*. Association for Computing Machinery, New York, NY, USA, 74–82. https://doi.org/10.1145/3574318.3574339

[39] Priyank Soni, Sohom Ghosh, and Sudip Kumar Naskar. [n. d.]. Detecting Issues Related to Environmental, Social, and Corporate Governance using SEC-BERT. In *Proceedings of International Conference on Data Science and Applications: ICDSA 2023*. Springer, "Springer", "Singapore".

[40] Harsha Vardhan, Sohom Ghosh, Ponnurangam Kumaraguru, and Sudip Kumar Naskar. 2023. A low resource framework for Multi-lingual ESG Impact Type Identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. ACL, Bali, Indonesia.