IndicFinNLP: Financial Natural Language Processing for Indian Languages

Sohom Ghosh[1], Arnab Maji[2], Aswartha Narayana[3], Sudip Kumar Naskar[4]
[1,4] Jadavpur University, [2,3] Independent Researcher
India
{sohom1ghosh, arnabmaji09, dkaswartha, sudip.naskar}@gmail.com

Abstract

Applications of Natural Language Processing (NLP) in the finance domain have been very popular of late. For financial NLP (FinNLP), while various datasets exist for widely spoken languages like English and Chinese, datasets are scarce for low resource languages,particularly for Indian languages. In this paper, we address these challenges by presenting IndicFinNLP – a collection of 9 datasets consisting of three tasks related to FinNLP for three Indian languages. These tasks are Exaggerated Numeral Detection, Sustainability Classification, and Environmental, Social, and Governance (ESG) Theme Determination of financial texts in Hindi, Bengali, and Telugu. Moreover, we have released the datasets under the CC BY-NC-SA 4.0 licence for the benefit of the research community.

Keywords: Financial Natural Language Processing, Language Resources, Indian Languages

1. Introduction

In a nation, Financial literacy leads to overall well-being of citizens and economic prosperity. The financial literacy rate in India is only 27%.[1] The cultural diversity and existence of more than 100 major languages in India make it difficult to spread financial knowledge among its citizens. Although most of the researchers working on FinNLP focused on creating datasets for high-resource languages like English and Chinese, to the best of our knowledge such datasets do not exist for Indian languages, even though some of the Indian languages belong to most spoken languages worldwide.[2] To improve the overall financial literacy of the country, it is essential to educate the citizens in their own vernacular languages. As misinformation is a prevalent problem in today's society, it is extremely important to ensure that the financial knowledge being imparted is authentic. Many a time, numbers and figures are misrepresented to allure common people who are novice investors. To address this, we develop a system to detect exaggerated numerals in financial texts in Indian languages. With the ever-growing concern for climate change, investors are increasingly looking for avenues of green investing like sustainable and Environmental, Social, and Governance (ESG) aspect of funds. We propose frameworks to automatically assess the sustainability aspect and detect ESG related themes present in financial texts written in Indian languages. Finally, we evaluate the necessity of having India specific FinNLP models. We summarize these tasks in Figure 1. Our datasets and models can be accessed here[3].

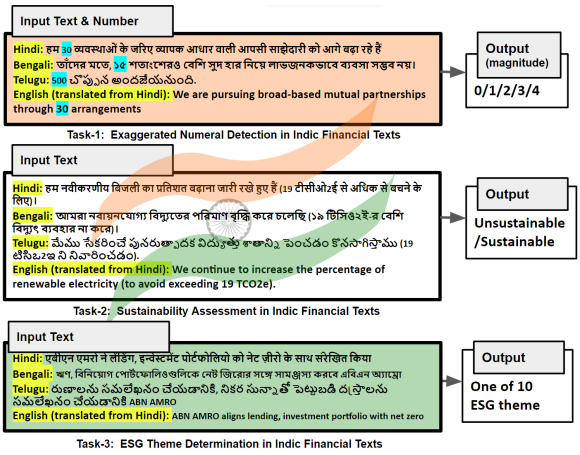

Figure 1: Financial Natural Language Processing for Indian Languages

Our Contributions

In this paper, we present **IndicFinNLP** - a collection of 9 datasets corresponding to three FinNLP tasks in three most spoken[4] Indian Languages (Hindi, Bengali, and Telugu). The tasks are: Exaggerated Numeral Detection, Sustainability Assessment, and ESG Theme Determination in Financial Texts. To the best of our knowledge, we are the first to create open-source Indic FinNLP datasets.

[1] https://yourstory.com/2023/07/financial-literacy-is-key-to-unlocking-india-economy (accessed on 11th September, 2023)

[2] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[3] https://github.com/sohomghosh/IndicFinNLP

[4] https://www.superprof.co.in/blog/indian-languages/ (accessed on 11th September, 2023)

LREC-COLING 2024, pages 9010–9018
20-25 May, 2024.

2. Related Works

Over the last few years, research in the FinNLP space has evolved rapidly. Researchers have applied FinNLP on various languages, extending beyond English (Chen et al., 2021, 2020a) to encompass Japanese (Kannan and Seki, 2023), Chinese (Tseng et al., 2023) and multiple European languages like Danish, Spanish, Turkish, etc. (Jørgensen et al., 2023). Some of the shared tasks like Financial Narrative Summarization (FNS) have recently moved from monolingual corpus in English (Zmandar et al., 2021) to multilingual corpus in English, Greek and Spanish (El-Haj et al., 2022). Similarly, the Financial Table of Content (FinTOC) (Maarouf et al., 2021) shared task has embraced multilingualism (Kang et al., 2022); the corpus is available in various European languages (English, French, Spanish).

The financial language resources available in English are quite varied, ranging from ESG-related aspects (Kang and El Maarouf, 2022; Chen et al., 2023) to stock market-related sentiments (Chen et al., 2020a). Various FinNLP tasks have been proposed in English that use annual reports (EDGAR-CORPUS) (Loukas et al., 2021), Earning Call Transcripts (Sawhney et al., 2021), Financial News (Chang et al., 2016; Alanyali et al., 2013), Analyst reports (Keith and Stent, 2019), speeches (Shah et al., 2023), and Social Media (Takayanagi et al., 2023). In addition to this, several shared tasks related to FinNLP are regularly organised, including FNS (El-Haj et al., 2022), FinTOC (Maarouf et al., 2021), FinCausal (Mariko et al., 2022), FinNum (Chen et al., 2019, 2020b, 2022), etc. Some of the FinNLP specific language models include FinBERT (Araci, 2019), FlangRoBERTA (Shah et al., 2022), SEC-BERT (Loukas et al., 2022), FinGPT (Yang et al., 2023), BloombergGPT (Wu et al., 2023) etc.

Lately, some researchers have been working on India-centric FinNLP (Narayan et al., 2022; Buggana et al., 2022). However, these works only focus on English texts, ignoring the linguistic diversity present in India. YubiBERT[5] is the only effort towards FinNLP in Indian languages. None of the existing works, including YubiBERT, has released datasets for addressing and benchmarking FinNLP related tasks in Indian languages.

3. Tasks

We focused on the following three FinNLP tasks in Indian languages.

Task-1: Given the position of an unknown numeral N in a financial text, the task is to determine its magnitude x such that $10^x <= N < 10^{x+1}$ where x $\in$ {0,1,2,3,4}. Numerals with magnitude more than 4 are treated as 4. Detecting the magnitude of numerals help in understanding if a certain number in a given context is exaggerated.

Task-2: Given a financial text, the task is to classify it into two classes: sustainable or unsustainable

Task-3: Given a financial text, the task is to determine the ESG theme related to it. The list of ESG themes are mentioned in Table 3.

[5] https://www.go-yubi.com/blog/yubibert-a-tiny-fintech-language-model/ (accessed on 11th September, 2023)

Language	0	1	2	3	4
Hindi	2435	3624	1444	2485	652
Bengali	1574	1886	931	1416	323
Telugu	1737	1800	983	1182	314

Table 1: Task-1 label wise distribution. 0/1/2/3/4 are the magnitudes

4. Datasets

In this section, we describe the datasets. For each of the datasets, we used 80%, 10%, and 10% instances selected randomly for training, validation, and testing respectively.

4.1. For Task-1

For Task-1, we extracted texts from budget speeches delivered by Finance Ministries of different State Governments and the Central Government of India. We focused on Hindi, Bengali, and Telugu-speaking states– Punjab, Uttarakhand, Haryana, West Bengal, Telengana, and Andhra Pradesh. We considered the budget speeches starting from the year 2011 till 2023 since for most of the states we could get these data for the recent few years. Subsequently, we filtered sizeable volumes of texts in Hindi, Bengali, and Telugu which were related to finance from the Samanantar corpus (Ramesh et al., 2022). For filtering, we extracted topics using the topic classification model of Antypas et al. (2022) and retained ones belonging primarily to the 'Business & Entrepreneurship' topic. Subsequently, we added all instances from Task-2 and Task-3. We kept only those texts which had at-least 6 words and one or more numerals in them. For preparing the final dataset, we created separate instances for each numeral in a given text. Statistics about the dataset is presented in Table 1. We present few samples in Table 7 (Appendix §B).

4.2. For Task-2

Since we were unable to find resources related to sustainability in the context of India, we translated the existing dataset proposed by Kang and

Language	BS(F1)	Sim.	Class	#
Hindi	>= 0.90	>=0.75	S	1212
			US	1026
Bengali	>=0.88	>=0.68	S	1203
			US	1025
Telugu	>=0.88	>=0.80	S	1119
			US	953

Table 2: Task-2 data distribution & thresholds. S=Sustainable, U=Unsustainable. BS=BERTScore, Sim.=Cosine Similarity

El Maarouf (2022) from English to Indian languages (Hindi, Bengali, and Telugu) using AI for Bharat Machine Translation System (Ramesh et al., 2022). To assess the quality of translation, we back-translated the texts in Indian languages to English using the same system. We calculated BERTScore (Zhang* et al., 2020) between the original and back-translated sentences in English. Subsequently, we calculated LaBSe (Feng et al., 2022) based similarity score for the original texts in English and translated texts in Indian languages. We manually looked at the instances and empirically decided the thresholds for BERTScore and cosine similarity to ensure that we retain only high quality instances after translation. For details regarding the dataset, thresholds are presented in Table 2. In Table 8 (Appendix §B) we show few instances from the dataset.

4.3. For Task-3

Although the ESG domain has been increasingly become popular, we could not find any dataset related to the Indian domain. Thus, we translated existing resources (Chen et al., 2023) from English to Indic languages (Hindi, Bengali, and Telugu) using AI for Bharat Machine Translation System (IndicTrans) (Ramesh et al., 2022). We manually checked each of the instances and made corrections wherever necessary. As the number of instances per ESG label was very low, (Refer: Appendix §A), we opted for coarse-grained classification. Thus, we mapped the issues to the corresponding themes using the mappings provided by MSCI ESG Research LLC.[6] More details regarding the dataset are presented in Table 3. Few instances are shown in Table 9 (Appendix §B).

5. Experiments and Results

We present our results in this section. The results of all three tasks are presented in Table 4.

[6] https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology+-+Access+to+Health+Care+Key+Issue.pdf/683e8c43-7c81-ada7-307d-d9356ec84efb?t=1666182590869 (accessed on 26th Spetember, 2023)

ESG Theme	#
climate change	92
corporate governance	91
environmental opportunities	72
product liability	68
natural capital	50
pollution waste	44
human capital	37
corporate behavior	30
social opportunities	27
stake holder opposition	21

Table 3: Task-3 label wise distribution for Hindi, Bengali, and Telugu.

5.1. Task-1

For the task of detecting exaggerated numeral, we extracted multilingual BERT (M-BERT) (Devlin et al., 2019) and IndicBERT (Kakwani et al., 2020) based embeddings of the numeral based on a context window of 512 tokens around it. We froze the underlying BERT models and trained LightGBM (Ke et al., 2017), XG-Boost (Chen and Guestrin, 2016), and Support Vector Machine (SVM) (Cortes and Vapnik, 1995) models over it. We observed that, the SVM models perform the best in every case. For the three languages, M-BERT outperforms IndicBERT.

5.2. Task-2

For the sustainability assessment task from financial texts, we fine-tuned M-BERT and IndicBERT for classification. Unlike Bengali and Hindi, for Telugu, MBERT outperformed IndicBERT. We pre-fine-tuned the best performing model in each case using Masked Language Modeling (MLM). We observed a notable improvement in the performance only for the Bengali dataset only. Finally, we translated the Indic sentences into English using the AI For Bharat Machine Translation system (Ramesh et al., 2022) and evaluated the RoBERTa based model (E-Ro) (Ghosh and Naskar, 2022) which was trained using the original texts in English. The E-Ro model out-performed all other models. This reveals the fact that for sustainability assessment, it is better to translate Indic texts to English, and score the models trained on English texts rather than developing separate models for each language.

5.3. Task-3

For the task of the ESG theme determination, we fine-tuned IndicBERT, M-BERT (MB) and MLM with M-BERT (MLM-MB) with 426 instances. We also trained an M-BERT model with the original English texts (E-MB) and evaluated it using English texts obtained by translating Indic sentences

into English using AI For Bharat Machine Translation System. For the three languages, the performance was less than 30% for each of the models. This is because the number of instances per ESG theme was very less (<100 for each label). To address this, we paraphrased the original English texts using a paraphraser (Vladimir Vorobev, 2023) and expanded the training and validation set to 4774 and 539 instances respectively (Table 6 in Appendix §A). We translated these instances into Indic languages using AI For Bharat Machine Translation System. We re-trained all the models with the paraphrased data (referred to as *P in Table 4). Paraphrase based models provide significant improvements in performance over baseline models for all languages.

6. Conclusion

In this paper, we narrated the datasets we created to solve three FinNLP tasks in three Indian languages (Hindi, Bengali, and Telugu). These tasks are: exaggerated numeral detection, sustainability assessment, and ESG theme determination in financial texts. Subsequently, we released baselines for each of the tasks. We observed that for Task-2 and Task-3 when we have texts in Hindi, Bengali, or Telugu instead of developing separate models for each language, we can simply translate these languages to English and score the existing model's trained English corpus. This improves the overall performance and saves the time and effort needed for training new models. Exploring other tasks in FinNLP and applying them to other low-resource languages are directions for future research. We would also want to work more on the datasets we created and further improve the baselines.

7. Limitations

In this paper, we cover only three of the main Indian languages. To avoid issues related to copyright infringement, we used openly available public India specific financial datasets like budget speeches. Unlike Bengali, the numerals in Hindi and Telugu were presented using English fonts in the original texts. As of now, we restricted ourselves to three tasks only. However, more tasks can be created from the dataset which we have proposed. Due to the non-availability of India-specific ESG and sustainability related data, for Task-2 and Task-3, we translated the corpora from English to Indic languages. As the texts in English were not specific to the Indian context, there may be some data drift while applying them in the Indian context. We acknowledge that translation and paraphrasing may lead to minor loss in semantics.

Ts	L	Model	Test			
			Pr	Re	F1	Acc
1	H	MB+LGB	0.63	0.64	0.63	0.64
1	H	IB+LGB	0.44	0.49	0.45	0.49
1	H	MB+XGB	0.63	0.64	0.63	0.64
1	H	IB+XGB	0.46	0.49	0.46	0.49
1	H	MB+SVM	**0.69**	**0.68**	**0.68**	**0.68**
1	B	MB+LGB	0.64	0.64	0.63	0.64
1	B	IB+LGB	0.51	0.51	0.50	0.51
1	B	MB+XGB	0.62	0.62	0.61	0.62
1	B	IB+XGB	0.51	0.50	0.48	0.50
1	B	MB+SVM	**0.66**	**0.65**	**0.65**	**0.65**
1	T	MB+LGB	0.59	0.61	0.59	0.61
1	T	IB+LGB	0.44	0.46	0.44	0.46
1	T	MB+XGB	0.59	0.60	0.59	0.60
1	T	IB+XGB	0.41	0.43	0.41	0.43
1	T	IB+XGB	**0.69**	**0.68**	**0.68**	**0.68**
2	H	IB	0.86	0.86	0.86	0.86
2	H	MB	0.77	0.77	0.77	0.77
2	H	MLM-IB	0.29	0.54	0.38	0.54
2	H	E-Ro	**0.95**	**0.95**	**0.95**	**0.95**
2	B	IB	0.80	0.80	0.80	0.80
2	B	MB	0.76	0.76	0.76	0.76
2	B	MLM-IB	0.81	0.81	0.81	0.81
2	B	E-Ro	**0.92**	**0.92**	**0.92**	**0.92**
2	T	IB	0.79	0.79	0.79	0.78
2	T	MB	0.90	0.89	0.89	0.89
2	T	MLM-IB	0.90	0.90	0.90	0.90
2	T	E-Ro	**0.92**	**0.92**	**0.92**	**0.92**
3	H	IB	0.03	0.17	0.05	0.17
3	H	MB	0.20	0.20	0.08	0.20
3	H	MLM-MB	0.20	0.20	0.11	0.20
3	H	E-MB	0.11	0.30	0.16	0.30
3	H	IB-P	0.12	0.26	0.16	0.26
3	H	MB-P	0.45	0.48	0.44	0.48
3	H	MLM-MB-P	0.43	0.46	0.44	0.46
3	H	E-MB-P	**0.56**	**0.63**	**0.59**	**0.63**
3	B	IB	0.03	0.17	0.05	0.17
3	B	MB	0.03	0.17	0.05	0.17
3	B	MLM-MB	0.11	0.20	0.10	0.20
3	B	E-MB	0.11	0.26	0.14	0.26
3	B	IB-P	0.20	0.30	0.23	0.30
3	B	MB-P	0.40	0.37	0.35	0.37
3	B	MLM-IB-P	0.32	0.37	0.33	0.37
3	B	E-MB-P	**0.55**	**0.59**	**0.55**	**0.59**
3	T	IB	0.03	0.17	0.05	0.17
3	T	MB	0.09	0.24	0.12	0.24
3	T	MLM-MB	0.07	0.22	0.11	0.22
3	T	E-MB	0.07	0.22	0.11	0.22
3	T	IB-P	0.27	0.31	0.22	0.31
3	T	MB-P	0.44	0.46	0.42	0.46
3	T	MLM-MB-P	0.36	0.41	0.37	0.41
3	T	E-MB-P	**0.56**	**0.63**	**0.58**	**0.63**

Table 4: Tasks (Ts) 1, 2, 3 results for Languages (L) Hindi (H), Bengali (B), Telugu (T). E=English, -P= -Paraphrased, IB=IndicBERT, MB=MBERT, XGB=XGBoost, LGB=LightGBM, Pr=Precision, Re=Recall, Acc=Accuracy. **Bold** means the best.

Acknowledgements

We would like to thank Ms. Dhariya Suman and Ms. Garima Verma for help us to prepare the datasets in Hindi.

References

Merve Alanyali, Helen Susannah Moat, and Tobias Preis. 2013. Quantifying the relationship between financial news and the stock market. Scientific reports, 3(1):3578.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In Proceedings of the 29th International Conference on Computational Linguistics, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Sathvik Sanjeev Buggana, Deepti Saravanan, Shravya Kanchi, Ujwal Narayan, Shivam Mangale, Lini T. Thomas, Kamalakar Karlapalem, and Natraj Raman. 2022. Sebi regulation biography. In Companion Proceedings of the Web Conference 2022, WWW '22, page 598–603, New York, NY, USA. Association for Computing Machinery.

Ching-Yun Chang, Yue Zhang, Zhiyang Teng, Zahn Bozanic, and Bin Ke. 2016. Measuring the information content of financial news. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3216–3225, Osaka, Japan. The COLING 2016 Organizing Committee.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020a. Issues and perspectives from 10,000 annotated financial social media data. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 6106–6110, Marseille, France. European Language Resources Association.

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).

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, pages 19–27.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2020b. Overview of the ntcir-15 finnum-2 task: Numeral attachment in financial tweets. Development, 850(194):1–044.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Overview of the FinNLP-2023 ML-ESG task: Multi-lingual esg issue identification. In Proceedings of the Joint Workshop of the 5th Financial Technology and Natural Language Processing (FinNLP) and 2nd Multimodal AI For Financial Forecasting (Muffin).

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. Machine learning, 20(3):273–297.

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.

Mahmoud El-Haj, Nadhem Zmandar, Paul Rayson, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, George Giannakopoulos, Aris Kosmopoulos, Blanca Carbajo-Coronado, and Antonio Moreno-Sandoval. 2022. The financial narrative summarisation shared task

(FNS 2022). In Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022, pages 43–52, Marseille, France. European Language Resources Association.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

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), pages 243–249, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A dataset for multilingual financial NLP. In Findings of the Association for Computational Linguistics: EACL 2023, pages 894–909, Dubrovnik, Croatia. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In Findings of EMNLP.

Juyeon Kang, Abderrahim Ait Azzi, Sandra Bellato, Blanca Carbajo Coronado, Mahmoud El-Haj, Ismail El Maarouf, Mei Gan, Ana Gisbert, and Antonio Moreno Sandoval. 2022. The financial document structure extraction shared task (FinTOC 2022). In Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022, pages 83–88, Marseille, France. European Language Resources Association.

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), pages 211–217, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for esg scores. In Proceedings of the Joint Workshop of the 5th Financial Technology and Natural Language Processing (FinNLP) and 2nd Multimodal AI For Financial Forecasting (Muffin).

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 3149–3157, Red Hook, NY, USA. Curran Associates Inc.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. EDGAR-CORPUS: Billions of tokens make the world go round. In Proceedings of the Third Workshop on Economics and Natural Language Processing, pages 13–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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), pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Ismail El Maarouf, Juyeon Kang, Abderrahim Ait Azzi, Sandra Bellato, Mei Gan, and Mahmoud El-Haj. 2021. The financial document structure extraction shared task (FinTOC2021). In Proceedings of the 3rd Financial Narrative Processing Workshop, pages 111–119, Lancaster, United Kingdom. Association for Computational Linguistics.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (FinCausal 2022). In Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022, pages 105–107, Marseille, France. European Language Resources Association.

Ujwal Narayan, Pulkit Parikh, Kamalakar Karlapalem, and Natraj Raman. 2022. Detecting regulation violations for an indian regulatory body

through multi label classification. In Companion Proceedings of the Web Conference 2022, WWW '22, page 610–614, New York, NY, USA. Association for Computing Machinery.

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, 10:145–162.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Ratn Shah. 2021. Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines. 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), pages 6751–6762, Online. Association for Computational Linguistics.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion dollar words: A new financial dataset, task & market analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6664–6679, Toronto, Canada. Association for Computational Linguistics.

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, pages 2322–2335, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Takehiro Takayanagi, Chung-Chi Chen, and Kiyoshi Izumi. 2023. Personalized dynamic recommender system for investors. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, page 2246–2250, New York, NY, USA. Association for Computing Machinery.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In Proceedings of the 32nd ACM International Conference on Information & Knowledge Management, CIKM '23, New York, NY, USA. Association for Computing Machinery.

Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on chatgpt paraphrases.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations.

Nadhem Zmandar, Mahmoud El-Haj, Paul Rayson, Ahmed Abura'Ed, Marina Litvak, Geroge Giannakopoulos, and Nikiforos Pittaras. 2021. The financial narrative summarisation shared task FNS 2021. In Proceedings of the 3rd Financial Narrative Processing Workshop, pages 120–125, Lancaster, United Kingdom. Association for Computational Linguistics.

A. Appendix: ESG Issue Distribution

The labelwise ESG issue distribution is presented in Table 5. The distribution of ESG themes for Hindi, Bengali, and Telugu after augmenting the datasets with paraphrases are presented in Table 6.

B. Appendix: Sample Datasets

Sample datasets for each of the tasks are presented in Table 7, 8, and 9 respectively.

C. Appendix: Assessing Quality of Translation

For assessing the quality of translations generated by AI For Bharat Machine Translation System (Ramesh et al., 2022), we calculated COMET (Rei

issue	#
Board	55
Product Carbon Footprint	30
Human Capital Development	28
Opportunities in Renewable Energy	27
Opportunities in Clean Tech	25
Packaging Material & Waste	25
Responsible Investment	24
Financing Environmental Impact	23
Biodiversity & Land Use	23
Carbon Emissions	23
Ownership & Control	23
Opportunities in Green Building	20
Business Ethics	19
Community Relations	17
Consumer Financial Protection	16
Water Stress	16
Climate Change Vulnerability	16
Pay	13
Toxic Emissions & Waste	13
Health & Demographic Risk	12
Accounting	11
Raw Material Sourcing	11
Access to Finance	9
Opportunities in Nutrition & Health	9
Privacy & Data Security	7
Chemical Safety	6
Electronic Waste	6
Supply Chain Labor Standards	6
Access to Health Care	5
Access to Communications	4
Controversial Sourcing	4
Product Safety & Quality	3
Labor Management	3

Table 5: ESG issue distribution for Hindi, Bengali, and Telugu.

	Train	Validation
climate change	825	99
corporate governance	825	88
environmental opportunities	649	77
product liability	605	66
natural capital	440	55
pollution waste	396	44
human capital	330	33
corporate behaviour	275	33
social opportunities	242	22
stake-holder opposition	187	22

Table 6: ESG theme distribution for Hindi, Bengali, & Telugu after augmenting the datasets with paraphrases.

et al., 2020) score for the dataset used in Task-3 which was manually verified and corrected. The COMET score for Hindi, Bengali, and Telugu were 0.825, 0.896, and 0.826 respectively.

D. Appendix: Hyper-parameters

We present the hyperparameters of the models we developed in Table 10. We use the default values for the rest of the hyperparameters & models, which are not mentioned in this table.

E. Appendix: Experimental Setup

We conducted the experiments in Google Colab (free tier, T4 GPU) and Kaggle (T4 GPU) platforms.

indic text	num	start	end	L	M
एनपीसीआई द्वारा 24 घंटे संदाय प्रणाली	24	16	18	H	1
एफएसआई के 4 क्षेत्रीय कार्यालय हैं–	4	10	11	H	0
মাতিল ম্যানুকান (১৯১৪-২০০১), রিয়েল এস্টেট বিনিয়োগকারী।	২০০১	22	26	B	3
যা টাকায় ১০ কোটি টাকারও বেশি।	১০	9	11	B	1
24 నుండి 30 లక్షల మధ్య ఉండవచ్చు	24	0	2	T	1
బజాజ్ పల్సర్ ఆర్ఎస్200 విడుదల: ధరల వివరాలు	200	19	22	T	2

Table 7: Task -1 samples. num = numeral, start = start position, end = end position, L = Language, M = Magnitude. H,B,T represents Hindi, Bengali, & Telugu respectively.

indic text	L	Sustainability
आपके संगठन का सकल वैश्विक स्कोप 1 उत्सर्जन मीट्रिक टन में कितना था?	H	unsustainable
फिश में रीसाइक्लिंग का सबसे अधिक महत्व है।	H	sustainable
প্রতি মিলিয়ন ডলারে কার্বন ডাই অক্সাইডের পরিমাণ হিসাবে কার্বনের তীব্রতা পরিমাপ করা হয়।	B	unsustainable
আরও সার্কুলার, দক্ষ উ□পাদন এবং পরিবহন পদ্ধতি টেন্ডার প্রক্রিয়ায় সুবিধা প্রদান করে।	B	unsustainable
2018 నుండి సంస్థ తన కార్బన్ పాదముద్రను 50% తగ్గించింది, అదే కాలంలో కార్బన్ తీవ్రతను 46% తగ్గించింది.	T	sustainable
2015 జనవరిలో విడుదల చేసిన జిహెచ్జి ప్రోటోకాల్ యొక్క స్కోప్ 2 గైడెన్స్ కార్పొరేట్ ప్రమాణాలకు సవరణ ప్రకారం మా జిహెచ్జి ఉద్గారాలను నివేదిస్తున్నాం.	T	unsustainable

Table 8: Task -2 samples. L = Language. H,B,T represents Hindi, Bengali, & Telugu respectively.

indic text	L	ESG-Theme
इएनजीआइ और टोटल ग्रीन हाइड्रोजन प्रोजेक्ट पर एकजुट हुए	H	environmental opportunities
पुटनाम ने नए ईएसजी फंडों के साथ सक्रिय ईटीएफ सुइट का विस्तार किया	H	product liability
ঋণ, বিনিয়োগ পোর্টফোলিওগুলিকে নেট জিরোর সঙ্গে সামঞ্জস্য করবে এবিএন অ্যাম্রো	B	climate change
অন্টারিও শিক্ষক পেনশন পরিকল্পনা বোর্ডগুলিতে 40% মহিলাদের প্রতিনিধিত্বের প্রত্যাশা সেট করে	B	corporate governance
ఫీనిక్స్ గ్రూప్ క్లైర్ హాకిన్స్‌ను ఎగ్జిక్యూటివ్ కమిటీకి నియమిస్తుంది, సస్టైనబిలిటీ స్ట్రాటజీకి బాధ్యత వహిస్తుంది	T	corporate governance
ప్రిన్సిపాల్ ఫైనాన్షియల్ గ్రూప్ వైవిధ్యం, చేరిక మరియు పర్యావరణ సుస్థిరత కార్యక్రమాలను ప్రారంభిస్తుంది	T	product liability

Table 9: Task -3 samples. L = Language. H,B,T represents Hindi, Bengali, & Telugu respectively.

Model	Hyper-parameters
IndicBERT and MBERT (for Task-2,3)	evaluation_strategy = 'epoch', save_strategy = 'epoch', learning_rate=2e-5, per_device_train_batch_size=32, per_device_eval_batch_size=32, num_train_epochs=2.5, weight_decay=0.01, load_best_model_at_end=True, metric_for_best_model='accuracy'
SVM	Regularization parameter (C) = 4, kernel = 'rbf', Tolerance=1e-6

Table 10: Hyperparameters of various models