Ranking Environment, Social And Governance Related Concepts And Assessing Sustainability Aspect Of Financial Texts

Sohom Ghosh[1,2*] , **Sudip Kumar Naskar**[2]
[1]Fidelity Investments, Bengaluru, India
[2]Jadavpur University, Kolkata, India
{sohom1ghosh, sudip.naskar}@gmail.com

Abstract

Understanding Environmental, Social, and Governance (ESG) factors related to financial products has become extremely important for investors. However, manually screening through the corporate policies and reports to understand their sustainability aspect is extremely tedious. In this paper, we propose solutions to two such problems which were released as shared tasks of the FinNLP workshop of the IJCAI-2022 conference. Firstly, we train a Sentence Transformers based model which automatically ranks ESG related concepts for a given unknown term. Secondly, we fine-tune a RoBERTa model to classify financial texts as sustainable or not. Out of 26 registered teams, our team ranked 4th in sub-task 1 and 3rd in sub-task 2. The source code can be accessed from https://github.com/sohomghosh/Finsim4_ESG.

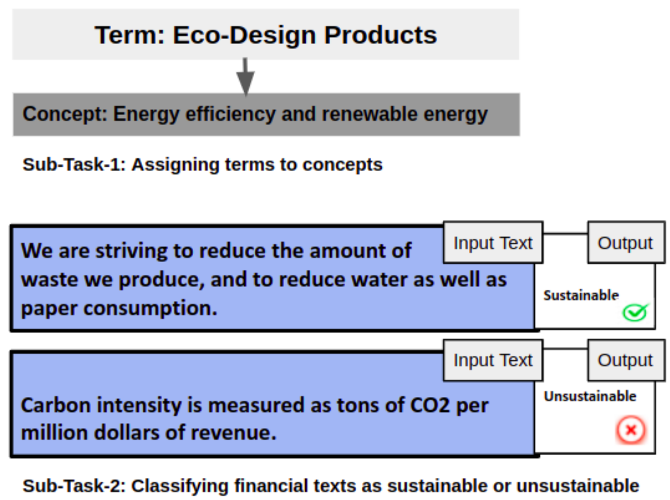

Figure 1: FinSim-4 ESG Sub-Tasks

1 Introduction

These days a lot of investors have become socially responsible and environmentally conscious[1]. They tend to choose stocks and funds which do not harm the environment[2]. Keeping this in mind, organizations are also putting in efforts to increase their Environmental, Social, and Governance (ESG) ratings. They tend to publish reports mentioning the ESG aspect of their policies. However, reading through all such reports is time-consuming and inefficient. This brings in the need for an automated system for mapping terms to ESG concepts and classifying financial texts as sustainable or not. FinNLP workshop of IJCAI-2022 conference hosted a shared task with these problems. We present an example of this in Figure 1. Our team LIPI participated in the shared task and ranked 4th and 3rd in sub-tasks 1 and 2 respectively. In this paper, we describe our solutions.

2 Related Works

The sub-task of mapping terms with high level concepts is similar to hypernym detection. For the Natural Language Processing (NLP) community, Hypernym detection has been an active area of research. Several SemEval tasks ([Bordea *et al.*, 2015], [Bordea *et al.*, 2016], [Augenstein *et al.*, 2017], [Camacho-Collados *et al.*, 2018]) were organized on this topic. Subsequently, three editions of FinSim ([Maarouf *et al.*, 2020], [Mansar *et al.*, 2021], [Kang *et al.*, 2021]) shared task were held which adapted the task of hypernym detection for the financial domain. This year while organizing FinSim-4, this was extended to ESG insights.

With the rising popularity of green investing, understanding the sustainability aspect of financial texts has become extremely important. Smeuninx et al. [Smeuninx *et al.*, 2020] studied the readability of annual reports of several organizations. They highlighted how formula-based readability scores classified these texts as complex documents. They also mentioned the need for NLP based techniques to comprehend the readability of such documents. Luccioni et al. [Luccioni *et al.*, 2020] fine-tuned RoBERTa-base [Liu *et al.*, 2019] model to develop a question-answering based tool, ClimateQA for extracting sections related to climate from financial reports.

*Contact Author

[1]https://news.gallup.com/poll/389780/investors-stand-esg-investing.aspx (accessed on 10 June 2022)

[2]http://bwdisrupt.businessworld.in/article/Sustainable-Investing-To-Surge-To-125-B-In-India-By-2026-Report/09-06-2022-432078/ (accessed on 10 June 2022)

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing
(FinNLP@IJCAI 2022), pages 87-92, Vienna, Austria, July 24th, 2022

Guo et al. [Guo *et al.*, 2020] proposed a framework ESG2Risk for predicting stock prices by analyzing ESG related events from financial news. They specifically used sentiments from these events.

Nugent et al. [Nugent *et al.*, 2020] pre-trained a BERT [Devlin *et al.*, 2019] model with financial news articles from Reuters News Archive for predicting ESG related controversies. Furthermore, they used it for mapping financial news into one of the United Nations Sustainable Development Goals.

3 Problem Statements

The fourth edition of FinSim presented two sub-tasks. They are as follows:

3.1 Sub-Task 1:

Given a set J consisting of n tuples of terms and their high level concepts i.e. $J = \{(t_1, c_1), (t_2, c_2), ...(t_n, c_n)\}$ where c_i represents the high level concept corresponding to the i^{th} term t_i and $c_i \epsilon$ set of concepts mentioned in Table 1. For a given unknown term, the task was to develop a system to rank these concepts.

The evaluation metrics for this sub-task were accuracy and mean rank. As per the evaluation script shared by the organizers, the rank of an instance was calculated by checking the presence of the true value in the first three elements of the predicted ranked list.

3.2 Sub-Task 2:

Given a set F consisting of n tuples of financial texts and their sustainability labels i.e. $F = \{(f_1, l_1), (f_2, l_2), ...(f_n, l_n)\}$ where l_i represents the sustainability label corresponding to the i^{th} financial text f_i and $l_i \epsilon$ {sustainable, unsustainable}. We need to develop a system to classify an unknown financial text as sustainable or not.

The evaluation metric for this sub-task was accuracy.

4 Data

The data sets provided by the organizers consist of a set of 190 documents in PDF format, 651 terms mapped to 24 concepts and 2265 financial texts labelled as sustainable or unsustainable. We provide more details about the data set in the following sections.

4.1 Data Description

For sub-task 1, the number of instances for each concept has been mentioned in Table 1. For sub-task 2, out of 2,265 financial texts 1,223 were sustainable whereas 1,042 were unsustainable. We maintained a training to validation split of 80% to 20% for both the sub-tasks.

4.2 Data Augmentation

Firstly for sub-task 1, we started by using 80% of 651 instances for training. To bring in more context, we collected the definitions for each of the 24 concepts from various websites. For each term (t_i, concept c_i) pair, we obtained the corresponding concept definition d_i. Since, each term t_i present here were mapped to a concept definition d_i, we had only positive instances i.e. similarity score of 1.0 corresponding to the (t_i, d_i) pair. Subsequently, we thought of adding negative samples in the training process as well. For each term, concept definition pair (t_i, d_i), we experimented by randomly paring t_i with 1, 5 or 15 concepts definitions. Later, we grouped the concepts manually. This is presented in Table 2. We could group 20 out of 24 concepts. The remaining four were singleton sets. For randomly selecting concept definitions for term t_i, we tried out the following sampling methods:

Concept	Count
Energy efficiency and renewable energy	59
Sustainable Food & Agriculture	54
Product Responsibility	51
Circular economy	47
Sustainable Transport	46
Emissions	39
Shareholder rights	38
Board Make-Up	37
Injury frequency rate for subcontracted labour	35
Executive compensation	32
Biodiversity	29
Community	27
Employee engagement	23
Employee development	22
Water & waste-water management	21
Carbon factor	19
Future of work	18
Waste management	16
Recruiting and retaining employees	11
Human Rights	10
Audit Oversight	7
Injury frequency rate	2
Board Independence	2
Share Capital	2

Table 1: Distribution of concepts

- Select any concept definition d_j such that concept $c_j \neq$ concept c_i, and assign a similarity score of 0.0 to the (t_i, d_j) pair.
- Select any concept definition d_j such that concept $c_j \notin$ the group where concept c_i is present, and assign a similarity score of 0.0 to the (t_i, d_j) pair.
- Select any concept definition d_j, if concept $c_j \notin$ the group where concept c_i is present assign a similarity score of 0.0 to the (t_i, d_j) pair, else assign a similarity score of 0.5 to the (t_i, d_j) pair.

5 System Description

As per the rules, for every team, the number of submissions for each sub-task was restricted to two. We describe each of our submissions here. We pictorially depict our methodology in Figure 2.

Group-1	Group-2	Group-3	Group-4
Carbon factor	Employee development	Injury frequency rate	Audit Oversight
Emissions	Recruiting and retaining employees	Injury frequency rate for subcontracted labour	Shareholder rights
Energy efficiency and renewable energy	Future of work	Human Rights	Executive compensation
	Employee engagement		Share Capital

Group-5	Group-6	Group-7
Waste management	Sustainable Transport	Board Independence
Water waste-water management	Sustainable Food Agriculture	Board Make-Up

Table 2: Concepts divided into groups

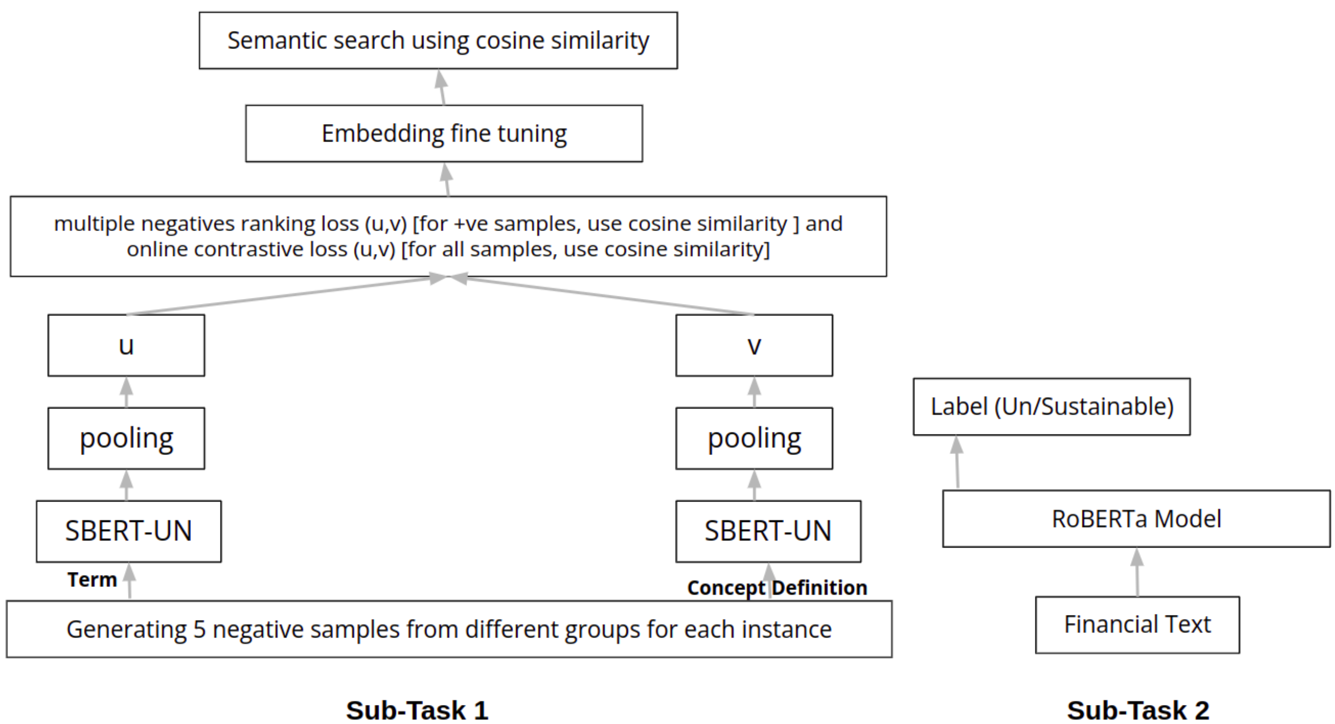

Figure 2: Methodology Sub-Task 1 and 2

5.1 Sub-Task 1, System -1

We fine-tuned a sentence transformer [Reimers and Gurevych, 2019] model[3] (SBERT-UN) which was pre-trained with United Nations (UN) sustainable development goals. For each of the terms in the training set, we randomly picked five concept definitions from different groups as mentioned in section 4.2. Our objective was to minimize the Multiple Negatives Ranking Loss as well as the Online Contrastive Loss. This was trained for 15 epochs with a batch size of 20.[4]. For sub-task 1, among all our submissions, this performed the best in terms of both accuracy and mean rank. This is similar to the solution [Chopra and Ghosh, 2021] presented at FinSim-3.

5.2 Sub-Task 1, System -2

This is a RoBERTa-base [Liu *et al.*, 2019] based classifier. We fine-tune the pre-trained RoBERTa-base model so that its [CLS] token learns how to classify terms into 24 pre-defined concepts or classes. It's hyper-parameters are as follows: maximum length = 16, batch size = 256, epochs = 60, learning rate = 0.00002. We use the checkpoint created at 57^{th} epoch as this was the best performing one.

5.3 Sub-Task 2, System -1

This system consists of the pre-trained FinBERT [Araci, 2019] fine-tuned for classifying financial texts as sustainable or unsustainable. It's hyper-parameters are as follows: maximum length = 128, batch size = 256, epochs = 60, learning rate = 0.00002. We use the checkpoint created at the 8^{th} epoch as this performed the best on the validation set.

5.4 Sub-Task 2, System -2

It consists of the pre-trained RoBERTa-base [Liu *et al.*, 2019] fine-tuned for the task of classifying financial texts as sustainable or not. It's hyper-parameters are as follows: maximum length = 128, batch size = 256, epochs = 60, learning rate = 0.00002. We use the checkpoint created at the 12^{th} epoch as this performed the best on the validation set. Among all our submissions, this performed the best on the test set.

6 Experiments and Results

We initiated by fine-tuning the all-mpnet-base-v2 model [Song *et al.*, 2020] using sentence transformer architecture. Our objective was to reduce the Multiple Negatives Ranking Loss as well as the Online Contrastive Loss for the task of Information Retrieval[4]. We also studied the effect of changing this model with the SBERT-UN model, adding negative samples and concepts as it is. We further experimented with different sampling methods as mentioned in section 4.2. Furthermore, we fine-tuned a RoBERTa-base [Liu *et al.*, 2019] based model to classify terms into 24 pre-defined concepts or classes.

Subsequently, we extracted texts from the documents provided in PDF format and fine-tuned a SBERT-UN model using Masked Language Modeling. However, this did not improve the performance. We also tried adding the definitions of 73 terms obtained from DBpedia [Auer *et al.*, 2007]. However, this did not yield any substantial improvement in the results. We present the result of sub-task 1 in Table 3. The SBERT-UN model trained with negative samples (SL. No. 8) performed the best in the validation as well as the test set.

For sub-task-2, we fine-tuned four models for classifying financial texts into two classes sustainable and unsustainable. These models are: RoBERTa-base [Liu *et al.*, 2019], FinBERT [Araci, 2019], SBERT-UN and SBERT-UN fine-tuned for sub-task 1. We present the results in Table 4. FinBERT [Araci, 2019] performed the best in the validation set whereas RoBERTa-base [Liu *et al.*, 2019] performed the best in the test set. Each of these models was trained for a maximum of 128 input tokens with a batch size of 256, a learning rate of 0.00002 and for 60 epochs.

We present the test set results in Table 5.

7 Conclusion and Future Work

In this paper, we elaborate on our team LIPI's approach toward solving the FinSim-4-ESG sub-tasks. As per the official report, out of 28 registered teams, 6 and 8 teams participated in sub-task 1 and 2 respectively. For sub-task 1, our team ranked 4^{th} whereas for sub-task 2, our team ranked 3^{rd}.

In future, we would like to collect more data and work towards improving the model performance. Developing a user-friendly tool for assigning terms to concepts and automatically evaluating the sustainable aspect of financial texts are other directions of future work.

Disclaimer

The opinions expressed in this paper are of the authors'. They do not reflect the opinions of their affiliations.

References

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

[Auer *et al.*, 2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

[Augenstein *et al.*, 2017] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[Bordea *et al.*, 2015] Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. SemEval-2015 task 17:

[3]https://huggingface.co/Rodion/sbert_uno_sustainable_development_goals

[4]The details are available at https://www.sbert.net/examples/training/quora_duplicate_questions/README.html

Sl. No.	Base Model	Data Augmentation	Mean Rank	Accuracy
1	all-mpnet-base-v2	No (only positives)	1.4692	0.6923
2	all-mpnet-base-v2	Yes (1 negative per positive)	1.5769	0.7000
3	sbert_un	No (only positives)	1.5308	0.6769
4	sbert_un	Yes (1 negative per positive)	1.4769	0.7308
5	sbert_un	Yes (1 negative per positive) + concepts	1.4615	0.7154
6	sbert_un	Yes (1 negative per positive) - concept definitions + concepts	1.4846	0.7462
7	sbert_un	Yes (1 negative per positive) [out of group sampling]	1.4385	0.7462
8	sbert_un	Yes (5 negative per positive) [out of group sampling]	**1.4308**	**0.7615**
9	sbert_un	Yes (15 negative per positive) [out of group sampling]	1.5308	0.7000
10	sbert_un	Yes (5 negative per positive) [out of group sampling] {batch size = 40, epoch = 30}	1.4154	0.7462
11	sbert_un	Yes (5 negative per positive) [out of group sampling] {batch size = 40, epoch = 20}	1.4615	0.7462
12	roberta classifier	-	1.4846	0.7538
13	sbert_un	Yes (1 negative per positive) [same group & out of group sampling]	1.4615	0.7462
14	sbert_un	Yes (5 negative per positive) [same group & out of group sampling]	1.5000	0.7385
15	baseline-1	-	2.5308	0.3769
16	baseline-2	-	1.6846	0.7154

Table 3: Results of Sub-Task 1 on the validation set.
NOTE: Where not mentioned, definitions of concepts were used with batch size of 20 for 15 epochs.

Sl. No.	Model	Accuracy
1	roberta-base	0.9338
2	finbert	**0.9426**
3	sbert_un	0.8653
4	sub-task1 finetune	0.8543

Table 4: Results of Sub-Task 2 on the validation set.

Sub-Task	Submission	Accuracy	Mean Rank
1	1	**0.7103**	**1.5172**
1	2	0.7034	1.6689
2	1	0.9219	-
2	2	**0.9317**	-

Table 5: Test set results for sub-tasks 1 and 2.

Taxonomy extraction evaluation (TExEval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 902–910, Denver, Colorado, June 2015. Association for Computational Linguistics.

[Bordea *et al.*, 2016] Georgeta Bordea, Els Lefever, and Paul Buitelaar. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091, San Diego, California, June 2016. Association for Computational Linguistics.

[Camacho-Collados *et al.*, 2018] Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 712–724, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Chopra and Ghosh, 2021] Ankush Chopra and Sohom Ghosh. 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*, pages 46–51, Online, 19 August 2021. -.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics.

[Guo *et al.*, 2020] Tian Guo, Nicolas Jamet, Valentin Betrix, Louis-Alexandre Piquet, and Emmanuel Hauptmann.

Esg2risk: A deep learning framework from esg news to stock volatility prediction, 2020.

[Kang *et al.*, 2021] Juyeon Kang, Ismail El Maarouf, Sandra Bellato, and Mei Gan. 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*, pages 31–35, Online, 19 August 2021. -.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[Luccioni *et al.*, 2020] Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. Analyzing sustainability reports using natural language processing, 2020.

[Maarouf *et al.*, 2020] Ismail El Maarouf, Youness Mansar, Virginie Mouilleron, and Dialekti Valsamou-Stanislawski. The FinSim 2020 shared task: Learning semantic representations for the financial domain. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 81–86, Kyoto, Japan, 5 January 2020. -.

[Mansar *et al.*, 2021] Youness Mansar, Juyeon Kang, and Ismail El Maarouf. *The FinSim-2 2021 Shared Task: Learning Semantic Similarities for the Financial Domain*, page 288–292. Association for Computing Machinery, New York, NY, USA, 2021.

[Nugent *et al.*, 2020] Tim Nugent, Nicole Stelea, and Jochen L. Leidner. Detecting esg topics using domain-specific language models and data augmentation approaches, 2020.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. 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)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Smeuninx *et al.*, 2020] Nils Smeuninx, Bernard De Clerck, and Walter Aerts. Measuring the readability of sustainability reports: A corpus-based analysis through standard formulae and nlp. *International Journal of Business Communication*, 57(1):52–85, 2020.

[Song *et al.*, 2020] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020.