FinSim4-ESG Shared Task: Learning Semantic Similarities for the Financial Domain. Extended edition to ESG insights

Juyeon Kang[1] , **Mehdi Kchouk**[1] , **Sandra Bellato**[1] and **Mei Gan**[1] and **Ismail El Maarouf**[2]
[1]Fortia Financial Solutions
[2]Imprevisible
{juyeon.kang, mehdi.kchouk, sandra.bellato, mei.gan}@fortia.fr, ismail.elmaarouf@imprevisible.com

Abstract

This paper describes *FinSim4-ESG*[1] shared task organized in the 4th FinNLP workshopwhich is held in conjunction with the IJCAI-ECAI-2022 conferenceThis year, the FinSim4 is extended to the Environment, Social and Government (ESG) insights and proposes two subtasks, one for ESG Taxonomy Enrichment and the other for Sustainable Sentence Prediction. Among the 28 teams registered to the shared task, a total of 8 teams submitted their systems results and 6 teams also submitted a paper to describe their method. The winner of each subtask shows good performance results of 0.85% and 0.95% in terms of accuracy, respectively.

1 Introduction

The FinSim shared task aims to spark interest from communities in NLP, ML/AI, Knowledge Engineering and Financial document processing. Going beyond the mere representation of words is a key step to industrial applications that make use of Natural Language Processing (NLP). This is typically addressed using either 1) Unsupervised corpus-derived representations like word embeddings, which are typically opaque to human understanding but very useful in NLP applications or 2) Supervised approach to semantic representations learning, which typically requires an important volume of labeled data, but has high coverage for the target domain or 3) Manually labeled resources such as corpora, lexica, taxonomies and ontologies, which typically have low coverage and contain inconsistencies, but provide a deeper understanding of the target domain.

These approaches form a different spectrum which a number of them have attempted to combine, particularly in tasks aiming at expanding the coverage of manual resources using automatic methods.

- The Semeval community has organized several evaluation campaigns to stimulate the development of methods which extract semantic/lexical relations between concepts/words ([Bordea *et al.*, 2015], [Bordea *et al.*, 2016], [Jurgens and Pilehvar, 2016], [Camacho-Collados *et al.*, 2018]).
- A large number of datasets and challenges specifically look at how to automatically populate knowledge bases such as DBpedia or Wikidata (e.g. KBP challenges, https://tac.nist.gov/2020/KBP/SM-KBP/).
- There are also a number of studies on the supervised and unsupervised approaches to the extraction of semantic relations between concepts and terms ([Fauconnier and Kamel, 2015], [Shwartz *et al.*, 2016], [Wang *et al.*, 2017], [Sarkar *et al.*, 2018], [Martel and Zouaq, 2021]).

This new edition of FinSim4-ESG is extended to the "Environment, Social and Governance (ESG)" related issues based on the sustainability reports, ESG reports, Environment report and annual reports periodically published by financial companies. The ESG criteria is a set of standards for a company's behavior used by socially conscious investors to screen potential investments. Environmental criteria consider how a company safeguards the environment, including corporate policies addressing climate change, for example. Social criteria examine how it manages relationships with employees, suppliers, customers, and the communities where it operates. Governance deals with a company's leadership, executive pay, audits, internal controls, and shareholder rights. For example, in financial domain, the ESG criteria are applied to assess the companies risk on these ESG aspects, so that help investors supporting business aligned with green initiatives.

According to the European Commission, from the end of 2022, companies providing investment products that make sustainability or environmental claims will be required to disclose how their portfolios align with the EU taxonomy[2] and ESG regulations for sustainable activities. The objective of this shared task is to elaborate an ESG taxonomy, ESG concepts representations, based on the data like companies' sustainability reports, annual reports, environment reports, etc. and make use of them to analyze how an economic activity complies with the taxonomy. Consequently, it allows us to know how an investment product aligns with ESG regulations.

[1]https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2022/shared-task-finsim4-esg

[2]https://ec.europa.eu/info/business-economy-euro/banking-and-finance/sustainable-finance/eu-taxonomy-sustainable-activities_en

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing
(FinNLP@IJCAI 2022), pages 57-63, Vienna, Austria, July 24th, 2022

2 Related works

The FinSim4-ESG proposes two subtasks: ESG Taxonomy Enrichment and Sustainability Prediction. The subtask1 is similar to the previous tasks of FinSim shared tasks, given a training set of terms and a fixed set of concepts, participants are asked to propose systems allowing to categorize new terms to their most likely concepts. A term-concept pair has a hierarchical and semantic relation if one of them can be conceived as a more generic term (e.g. *Emissions - Greenhouse gas emissions*). And the subtask2 proposes to solve a sentence classification problem in order to classify each sentence extracted from the ESG related reports into sustainable or unsustainable.

A taxonomy represents a semantic relation of term pairs, which is "*isa*" pairs, largely used in NLP and IE tasks. The taxonomy extraction task from a domain specific corpora as a competition is first proposed by the shared task TExEval [Bordea *et al.*, 2015] and TExEval-2 [Bordea *et al.*, 2016] as part of SemEval-2015 and 2016. Several works introduced methods to learn hypernymy from the corpora and showed how to induce taxonomies from "*isa*" pairs. While those systems largely exploit semantic lexical resources like WordNet, BabelNet, YAGO, Wiki, DBPedia, distributional approach was not meaningfully adopted.
More recently, as part of SemEval 2020, the shared task on Predicting Multilingual and Cross-Lingual (Graded) Lexical Entailment [Glavaš *et al.*, 2020] proposed a challenge for detecting semantic hierarchical relation, hypernym-hyponym, from multilingual and cross-lingual datasets. The Distributional track was newly added in order to evaluate distributional systems. The participating systems make use of rule-based approach by exploiting Wiktionary definitions of concepts [Kovács *et al.*, 2020] or distributional approach combining distributional word vectors, multilingual lexical resources and translated parallel corpora to obtain cross lingual synonyms, then to extract a set of terms which are semantically most similar to a seed term [Hauer *et al.*, 2020] [Wang *et al.*, 2020]. Also, in the previous FinSim shared tasks [Maarouf *et al.*, 2020] [Kang *et al.*, 2021], the authors proposed various approaches for the hypernyms and synonyms ranking of financial terms by using the state-of-art models like BERT and its variants (e.g. FinBERT) with good performance results.

The distributional semantic models are widely explored in different NLP based financial data analysis. A lot of studies make use of the fine-tuned models from various extensions of BERT model like Sentence-BERT [Reimers and Gurevych, 2019], RoBERTa [Liu *et al.*, 2019], DistilBERT [Sanh *et al.*, 2019]. For example, the fine-tuned model on the financial data, FinBERT [Yang *et al.*, 2020], is largely used and ESG-BERT [Mukherjee, 2020] is also trained on sustainability corpus with the growing interest in the ESG data analysis. Recently, we observe that the ESG related data like 10K report, 10Q filling or annual reports has been studied for automating the ESG ratings of the companies. [Balakrishnan *et al.*, 2010] proposes a linear SVM classifier to predict market performance based on narrative disclosures of 10K reports while [Mehra *et al.*, 2022] and [Armbrust *et al.*, 2020] investigate the effect of the environmental performance of a company on the relationship between the company's disclosures and financial performance. [Sokolov *et al.*, 2021] also proposes an approach to automatically convert unstructured data into ESG scores using a pre-trained BERT model. [Matthew Purver and Pollak, 2022] proposes a diachronic analysis of ESG terms in UK annual reports at the First Computing Social Responsibility Workshop-NLP Approaches to Corporate Social Responsibilities (CSR-NLP[3]). [Luccioni *et al.*, 2020] proposes NLP based methods for the analysis of financial reports in order to identify climate-relevant sections based on a question answering approach and [Guo *et al.*, 2020] develops a pipeline of ESG news extraction, news representations, and Bayesian inference of deep learning models.

3 Task Description

The new edition proposes two subtasks: ESG taxonomy enrichment and Sustainability prediction.

3.1 ESG Taxonomy Enrichment

We have created an in-house sustainable finance taxonomy called "Fortia ESG taxonomy". It is based on different financial data provider's taxonomies as well as several sustainability and annual reports where we looked for ESG related criteria. Given a subset of "Fortia ESG taxonomy", participants will be asked to enrich this training set to cover the rest of the terms of the original "Fortia ESG taxonomy". For this purpose, participants are given a set of ESG related reports of financial companies from which they can develop a model allowing to induce semantically related terms to the concepts defined in the training set. For example, given a set of terms related to the concept *Waste management* (e.g. *Hazardous Waste, Waste Reduction Initiatives*), the participating systems need to find the missing ones by the way that you predict a corresponding concept to unlabeled terms.

3.2 Sustainability Prediction

Participants are asked to design a system which can automatically classify sentences into sustainable or unsustainable sentences making use of the enriched taxonomy. For this purpose, participants are given a list of carefully selected labeled sentences from the sustainability reports and other documents. In this shared task, we consider a sentence as sustainable if a sentence semantically mentions the Environmental or Social or Governance related factors as defined in our ESG taxonomy.

Performance is measured according to the accuracy with which label is assigned, and according to recall (based on the total number of predictions).

This year, we propose a subset of our in-house ESG taxonomy and a dataset composed of financial and non-financial reports. And we are interested in systems which make use of contextual word embeddings such as BERT ([Devlin *et al.*, 2019]), as well as systems which make use of resources related to the ESG (Environmental, Social and Governance) and sustainability including EU taxonomy.

[3]https://csr-nlp.github.io/CSR-NLP-2022/

3.3 ESG Dataset

ESG related Reports Corpus The main topic of FinSim4 is the ESG taxonomy based sustainable activities analysis of the financial companies. For this purpose, we built a corpus composed of 190 sustainability reports, environment reports, annual reports and ESG reports, where the companies periodically publish the results of their activities showing its social or environmental impact.

ESG Taxonomy and Concepts-Terms Data Preparation We elaborated a first version of ESG taxonomy where the Environment, Social and Government topics are organized into groups of concepts and each concept is composed of semantically related terms. The Environment topic contains 9 concepts: Carbon factor, Emissions, Energy efficiency and renewable energy, Waste management, Water & waste-water management, Biodiversity, Sustainable Transport, Sustainable Food & Agriculture and Circular economy while the Social topic has 9 concepts: Employee development, Recruiting and retaining employees (incl. work-life balance), Future of work, Employee engagement, Injury frequency rate, Injury frequency rate for subcontracted labour, Community, Human rights and Product Responsibility, the Government topic with 6 concepts: Board Independence, Board Make-Up, Audit Oversight, Shareholder rights, Executive compensation and Share Capital. And we carefully selected the terms for each of these concepts as described in Table 1 taking into account the principals of EU taxonomy for sustainable activitiesand manually validated them based on the criteria used by ESG data providers[4].

Sustainable and Unsustainable Sentences Annotation For the subtask2, we first collected the candidate sentences using the dataset elaborated for the subtask1, a total of 792 terms. These terms based sentences extraction allowed creating a dataset of 2265 sustainable and unsustainable sentences from the corpus composed of the ESG related reports as above mentioned (See Table 2). Then, we manually annotated them reading the whole context from where the candidate sentence is extracted, otherwise, this information was not included in the dataset provided by the shared task. For this task, two experienced annotators cross-validated the annotated sentences.

Concepts	Training	Test
Energy efficiency and renewable energy	59	12
Sustainable Food & Agriculture	54	10
Product Responsibility	51	10
Circular economy	47	8
Sustainable Transport	46	7
Emissions	39	9
Shareholder rights	38	10
Board Make-Up	37	6
Injury frequency rate for subcontracted labour	35	5
Executive compensation	32	7
Biodiversity	29	10
Community	27	7
Employee engagement	23	5
Employee development	22	5
Water & waste-water management	21	4
Carbon factor	19	6
Future of work	18	5
Waste management	16	4
Recruiting and retaining employees	11	4
Human Rights	10	4
Audit Oversight	7	3
Share Capital	2	1
Board Independence	2	2
Injury frequency rate	2	1
Total	647	145

Table 1: ESG terms-concepts data for the subtask1

Label	Training	Test
Sustainable	1223	103
Unsustainable	1042	102
Total	2265	205

Table 2: Sustainability sentences data for the subtask2

4 Evaluation Setup

4.1 Baselines

We prepared two simple baselines in order to help the participants get started. Both baselines are based on a custom Word2Vec model that was trained on a corpus composed of ESG reports, Sustainability reports, environment reports and annual reports. The vector representation for each term is computed as the average of the word embeddings of their tokens. In the case of the subtask1, for each test sample, the first baseline ranks all the possible hypernyms using the hyponym-hypernym similarity in the embedding space. The second baseline trains a logistic regression model that classifies each test sample into different classes where each class represents one possible hypernym. In the case of the subtask2, for each test sample, the first baseline classifies a list of sentences into sustainable or unsustainable based on the sentence similarity and the second trains a classic classifier, both using the custom Word2Vec trained on the ESG dataset.

[4]Among others, we can refer to https://numeum.fr/societe/vigeo-eiris and https://www.refinitiv.com/fr/sustainable-finance/esg-scores

4.2 Evaluation metrics

We use the same metrics as the previous edition of FinSim, Accuracy for the subtasks 1 and 2, and Mean Rank for the subtask1. For each term x_i with a label y_i, the expected prediction is a top 3 list of labels ranked from most to least likely to be equal to the ground truth by the predictive system $\hat{y}_i^l$. We note by $rank_i$ the rank of the correct label in the top-3 prediction list, if the ground truth does not appear in the top-3 then $rank_i$ is equal to 4. Given those notation the accuracy can be expressed as:

$$Accuracy = \frac{1}{n} * \sum_{i=1}^{n} I(y_i = \hat{y}_i^l[0])$$

And the Mean Rank as:

$$Mean\_Rank = \frac{1}{n} * \sum_{i=1}^{n} rank_i$$

4.3 Submissions

Among 28 teams registered to the shared task, a total of 8 teams submitted their systems results and 6 teams also submitted a paper to describe their method. The extended version of the shared task to ESG has gained more attention from private institutions including Rakuten, Trading Central, Tata Consultancy Services, Fidelity Investments, Fidelity Brokerage Services LLC. (See Table 3 for more details).

Team name	Institutions
FORMICA	Jozef Stefan Institute & Queen Mary University of London
JETSONS	Fidelity Brokerage Services LLC.
KAKA	Rakuten Group
LIPI	Fidelity Investments
TCSTWIM	Tata Consultancy Services
Trading Central Labs	Trading Central Labs - La Rochelle

Table 3: Participant teams

FORMICA The FORMICA team proposes a system for the subtask2. The authors make use of knowledge background approach for the prediction of sustainable sentences, especially the embeddings model based on the knowledge derived from taxonomies, Tax2Vec, and an extended BERT model introducing the background knowledge, LinkBERT, to capture dependencies and knowledge that span across documents. The authors led experiments, first, using contextual or non contextual word features on BERT representations and LSA representations, second, using knowledge graph or taxonomy based features on Tax2Vec, TransE, DisMult and RotatE representations, finally using the joint representations of all the generated representations. The two submitted runs were generated based on the joint latent representations (first run) and using the result of the ensemble modeling methods from multiple models, LinkBERT, FinBERT and the joint SVD (second run). The second run slightly outperforms the first with 0.89% of accuracy in the testset while the fine-tuned LinkBERT achieved 0.96% of F1-score on the internal data split.

JETSONS The JETSONS team tackles both of the subtasks proposed by FinSim4-ESG. For the first subtask, the final submission was generated from the approach using the fine-tuned Sentence-BERT representations as encoder and the logistic regression classifier as decoder. We observe that the result of the classification varies from 0.89% to 0.61% on the ten-fold cross validation on the train set and on the testset, respectively. Their experiments show that the submitted approach outperforms the results of the similarity measuring either based on the pre-trained DistilBERT without fine-tuning or fine-tuned DistilBERT on the financial reports or the pre-trained Sentence-BERT without fine-tuning. For the second subtask, the fine-tuned RoBERTa shows the best performance with 93% of accuracy comparing to the results from BERT and T5[Raffel *et al.*, 2020] (Text-to-Text Transfer Transformer) models.

KAKA team The KAKA team tackles two subtasks leading several experiments based on the state-of-the art algorithms. For the subtask1, the authors propose two approaches: one with a classical Machine Learning model as our first baseline proposes but combining the tf-idf vectors with the custom Word2Vec and the other with a deep attention model using the custom Word2Vec model. They trained the word embeddings on an augmented data by adding the term-definition pair in the provided corpus by the organizer and also in the training and test data. The second approach slightly outperforms the classical approach on the test data while the first outperforms the deep learning approach on the validation data. For the subtask2, the authors led experiments based on different pre-trained Language models like BERT, RoBERTa, ALBERT, DistillBert and XLNet by fine-tuning them for the sustainable sentence classification. The fine-tuned RoBERTa model outperforms all the submitted systems with 94.63% of accuracy.

LIPI The LIPI team proposes the solutions to both of the subtasks. For the subtask1, the authors first propose an augmented terms dataset by adding definitions of each concept to make use of more contextual information. Then the pre-trained Sentence-BERT model was fine-tuned on the United Nations (UN)'s sustainable development goals[5] for the first run and the RoBERTa model for the second run while the Sentence-BERT fine-tuned on UN reports results the best performing score with 0.76% of accuracy. For the subtask2, the pre-trained FinBERT was fine-tuned for the first run and the pre-trained RoBERTa for the second run. We observe that the latter outperforms for the sentence classification task too on the testset with 0.93% of accuracy.

TCSWITM The TCSWITM team submitted the results for both of the subtasks. For the subtask1, the authors explore semantic similarity features inside BERT architecture by the way that they augment the obtained embeddings from the fine-tuned BERT model on the ESG related reports with Word2Vec, Cosine and Jaccard similarity features. Then they trained a logistic regression classifier on top of these representations also using PCA to handle the dimensionality issue. The experiments show that it improves the ESG terms prediction results with 0.82% of accuracy comparing to the result of the generic BERT model (0.76%) on their internal data split. For the subtask2, they introduce various lexical features for the sustainable sentence classification task like sentiment polarity, POS tags, NER tags, etc. Then they led various experiments based on different word and sentence embeddings including Word2Vec, GloVe, FastText, ELMo, InferSent, BERT, and ESG BERT and also trained several widely used classification methods including Logistic Regression, Gradient Boosting and XGBoost Classifier, gradually augmenting the models by adding the features one by one. The results show that the logistic regression classifier trained on top of the ESG BERT along with all the NLP features performs better than other setups with 0.87% of accuracy.

Trading Central Labs-LaRochelle The Trading Central Labs team tackles two subtasks of the shared task FinSim4-ESG. For the first one, the authors use a pre-trained Sentence-

[5]https://www.undp.org/sustainable-development-goals

BERT to embed the terms and then train a classifier on the train set to reach the top 1 of the subtask1 with 0.85% of accuracy. The authors consider all the terms of a same concept as paraphrases having similar semantic information so the trained model returns a high score in terms of similarity on two paraphrases. They propose a simple but effective way to combine Sentence-BERT and a logistic regression to classify terms without concepts. In the second subtask, for the final submission, based on the results of experiments on DistillBERT, BERT and RoBERTa, they use a pre-trained RoBERTa model with a feed forward layer to classify the sentences to reach the fourth best performing system with 0.93% of accuracy on the testset.

5 Results and Analysis

In Table 4, we ranked the results of the 12 system runs submitted by 6 teams and in Table 5 the results of the 14 system runs by 8 teams according to the metric described in the section 4.2, both including those of our baselines. The overall results of the subtask1 were obtained by combining those of Mean Rank and Accuracy and the Trading Central Labs-La Rochelle team's runs won first and second places for both metrics. For the subtask2, the KAKA team's second run won first place and CompLx team came second with the accuracy of 0.95 and 0.94%, respectively.

Team	Accuracy (%)	Mean Rank
Baseline_1	0.46	2.28
Baseline_2	0.74	1.52
JETSONS_1	0.61	1.97
KAKA_1	0.74	1.44
KAKA_2	0.75	1.54
LIPI_1	0.71	1.52
LIPI_2	0.70	1.67
TCSWITM_1	0.77	1.46
TCSWITM_2	0.78	1.45
TradingCentralLabs_1	0.83	1.26
TradingCentralLabs_2	**0.85**	**1.26**
vishleshak_1	0.68	1.61

Table 4: Mean Rank and Accuracy (listed alphabetically) for the subtask1: ESG Taxonomy Enrichment

All the participating teams commonly explored BERT models along with its variants to measure the semantic relatedness between terms and concepts. The fine-tuned models on the ESG corpus on a basis of BERT [Devlin *et al.*, 2019], Sentence-BERT [Reimers and Gurevych, 2019], DistilBERT [Sanh *et al.*, 2019], RoBERTa [Liu *et al.*, 2019], LinkBERT [Yasunaga *et al.*, 2022], FinBERT [Yang *et al.*, 2020], ALBERT [Lan *et al.*, 2019] are proposed by most of the participating systems either for the word representations in vector space or for the term/sentence classification task and the classical logistic regression model is trained for the classification task giving the most performing results.

5.1 Subtask1: ESG Taxonomy Enrichment

For the ESG taxonomy enrichment task, the data augmentation methods was introduced by KAKA and LIPI teams not only to enrich the data size but also to add more contextual information for each term using the definition related to the term and its concept. The LIPI team also used the ESG related UN's reports data, in addition to the data provided by the shared task, which is not yet widely explored by showing that it helps improve the result of ESG terms prediction. The fine-tuned Sentence-BERT model representations and a classical linear classification model like logistic regression commonly shows a high performance to predict the ESG term-concept.

Team name	Accuracy (%)
Baseline_1	0.50
Baseline_2	0.82
CompLx_1	0.94
FORMICA_1	0.88
FORMICA_2	0.89
JETSONS_1	0.93
KAKA_1	0.93
KAKA_2	**0.95**
LIPI_1	0.92
LIPI_2	0.93
TCSTWIM_1	0.87
TradingCentralLabs-LaRochelle_1	0.91
TradingCentralLabs-LaRochelle_2	0.93
vishleshak_1	0.91

Table 5: Accuracy (listed alphabetically) for the subtask2: Sustainability Prediction

5.2 Subtask2: Sustainability Prediction

The teams JETSONS, KAKA, LIPI and Trading Central show that the fine-tuned RoBERTa outperforms other models like the fine-tuned Sentence-BERT or FinBERT on the sustainable sentence classification task.

We observe that the evaluation results on the training set by the participant's internal data split tend to show an important gap comparing to the results on the testset even though the training and test sets have a high level of similarity. We also observe this between the sustainable and unsustainable sentences:

- Unsustainable: ***By transitioning*** *the gas network to bring hydrogen (and other gases such as biomethane) to homes and industries, it* ***can reduce*** *its carbon footprint.*
- Sustainable: *Together, these initiatives further* ***reduced*** *the carbon footprint of the Autolease portfolio.*

Some sentences in both classes require more contexts for a clear understanding in terms of sustainability. This issue was already taken into account at the sustainability data preparation, consequently, the sentences were selected by the way that it can be easily classifiable for the participants but the results analysis show that there still remain difficulties to classify into sustainable or unsustainable even by human.

6 Conclusions and Perspectives

The FinSim4-ESG proposed two subtasks: ESG Taxonomy Enrichment and Sustainability Prediction. Among the 8 participating teams, 6 teams submitted the systems runs to the

subtask1 and all submitted to the subtask2. All of the system runs showed very promising results using state-of-art NLP and ML techniques and features. As the first edition about ESG Taxonomy and Sustainability prediction, the systems with the best performance achieved a good accuracy of 83%∼85% for the subtask1 and a high accuracy of 94%∼95% for the subtask2. All the participating systems largely exploited distributional methods for the similarity measures between terms and sentences, and for the classification task, and the results showed that using distributed and contextual features improve the performance of their systems. Especially, several experiments from the participating systems show that the fine-tuned RoBERTa on the ESG data outperforms other models like BERT, Sentence-BERT, FinBERT, LinkBERT and the linear classifier performs better than non linear classifiers for both subtasks. And the results confirm that the data augmentation helps improve the overall results as also shown in the results of the previous editions of the FinSim shared task.

The impact of AI technologies grows more and more in ESG related domains like ESG ratings for analyzing the sustainable activities of the companies, Green investments supporting activities aligned with environmentally friendly business and helping investors to hold green bonds, green ETFs, green funds or stock of the companies supporting green initiatives, ESG risk assessment, ESG databases, etc. and this requires a study on how to exploit a large scale of ESG related concepts and build a knowledge representation of those concepts. The EU Taxonomy was already released in the objectives of European Green Deal[6] but they needs to extend the scope of the concepts toward Social and Government topics. In this shared task, we elaborated and provided a first version of ESG Taxonomy taking into account the EU taxonomy along with the ESG criteria proposed by the well known ESG data providers like Refinitiv and Moody's. It will be possible to improve FinSim-ESG task by proposing to increase the coverage of ESG concepts and its terms as the proposed concepts are still limited to those observed in the corpus composed of a limited number of reports. Also, the current task is focused on a monolingual data processing. Knowing that ESG data analysis is gaining increasing global attention and has become an increasingly important part of the investment, every year, more companies publish their activities related to ESG topics in non financial reports in different languages from different countries. The majority of the ESG concepts are language-independent, so it will be interesting to extend the task to a multilingual data processing.

Acknowledgments

This shared task is fully supported by Fortia Financial Solutions. We thank all the participants for their motivations and constructive exchange.

References

[Armbrust *et al.*, 2020] Felix Armbrust, Henry Schäfer, and Roman Klinger. A computational analysis of financial and environmental narratives within financial reports and its value for investors. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 181–194, Barcelona, Spain (Online), December 2020. COLING.

[Balakrishnan *et al.*, 2010] Ramji Balakrishnan, Xin Ying Qiu, and Padmini Srinivasan. On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, 202(3):789–801, 2010.

[Bordea *et al.*, 2015] Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. SemEval-2015 task 17: 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 (SemEval-2018)*, New Orleans, LA, United States, 2018.

[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.

[Fauconnier and Kamel, 2015] Jean-Philippe Fauconnier and Mouna Kamel. Discovering hypernymy relations using text layout. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 249–258, Denver, Colorado, June 2015. Association for Computational Linguistics.

[Glavaš *et al.*, 2020] Goran Glavaš, Ivan Vulić, Anna Korhonen, and Simone Paolo Ponzetto. SemEval-2020 task 2: Predicting multilingual and cross-lingual (graded) lexical entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online), December 2020.

[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.

[Hauer *et al.*, 2020] Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. UAlberta at SemEval-2020 task 2: Using translations to predict cross-lingual entailment. In *Proceedings of the Four-*

[6] https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en

teenth Workshop on Semantic Evaluation, pages 263–269, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[Jurgens and Pilehvar, 2016] David Jurgens and Mohammad Taher Pilehvar. SemEval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1092–1102, San Diego, California, June 2016. Association for Computational Linguistics.

[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.

[Kovács *et al.*, 2020] Ádám Kovács, Kinga Gémes, Andras Kornai, and Gábor Recski. BMEAUT at SemEval-2020 task 2: Lexical entailment with semantic graphs. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 135–141, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[Lan *et al.*, 2019] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2019.

[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. *CoRR*, abs/2011.08073, 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.

[Martel and Zouaq, 2021] Félix Martel and Amal Zouaq. Taxonomy extraction using knowledge graph embeddings and hierarchical clustering. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 836–844, New York, NY, USA, 2021. Association for Computing Machinery.

[Matthew Purver and Pollak, 2022] Riste Ichev Igor Lončarski Katarina Sitar Šuštar Aljoša Valentinčič Matthew Purver, Matej Martinc and Senja Pollak. Tracking changes in esg representation: Initial investigations in uk annual reports. In *Proceedings of the LREC Workshop on Corporate Social Responsibility*, 2022.

[Mehra *et al.*, 2022] Srishti Mehra, Robert Louka, and Yixun Zhang. ESGBERT: Language model to help with classification tasks related to companies' environmental, social, and governance practices. In *Embedded Systems and Applications*. Academy and Industry Research Collaboration Center (AIRCC), mar 2022.

[Mukherjee, 2020] Mukut Mukherjee. Esg-bert: Nlp meets sustainable investing. https://towardsdatascience.com/nlp-meets-sustainable-investing-d0542b3c264b, 2020.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[Sarkar *et al.*, 2018] Rajdeep Sarkar, John P. McCrae, and Paul Buitelaar. A supervised approach to taxonomy extraction using word embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[Shwartz *et al.*, 2016] Vered Shwartz, Yoav Goldberg, and Ido Dagan. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Sokolov *et al.*, 2021] Alik Sokolov, Jonathan Mostovoy, Jack Jie Ding, and Luis A. Seco. Building machine learning systems for automated esg scoring. 2021.

[Wang *et al.*, 2017] Chengyu Wang, Xiaofeng He, and Aoying Zhou. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1190–1203, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[Wang *et al.*, 2020] Shike Wang, Yuchen Fan, Xiangying Luo, and Dong Yu. SHIKEBLCU at SemEval-2020 task 2: An external knowledge-enhanced matrix for multilingual and cross-lingual lexical entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 255–262, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[Yang *et al.*, 2020] Yi Yang, Mark Christopher Siy UY, and Allen Huang. FinBERT: A Pretrained Language Model for Financial Communications. *arXiv e-prints*, page arXiv:2006.08097, June 2020.

[Yasunaga *et al.*, 2022] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links, 2022.