Using contextual sentence analysis models to recognize ESG concepts

Elvys Linhares Pontes and **Mohamed Benjannet**
Trading Central Labs, Trading Central, Paris, France
{elvys.linharespontes,mohamed.benjannet}@tradingcentral.com

Jose G. Moreno
IRIT, UMR 5505 CNRS, University of Toulouse
Toulouse, France
jose.moreno@irit.fr

Antoine Doucet
L3i, La Rochelle Université
La Rochelle, France
antoine.doucet@univ-lr.fr

Abstract

This paper summarizes the joint participation of the Trading Central Labs and the L3i laboratory of the University of La Rochelle on both sub-tasks of the *Shared Task FinSim-4* evaluation campaign. The first sub-task aims to enrich the 'Fortia ESG taxonomy' with new lexicon entries while the second one aims to classify sentences to either 'sustainable' or 'unsustainable' with respect to ESG (Environment, Social and Governance) related factors. For the first sub-task, we proposed a model based on pre-trained Sentence-BERT models to project sentences and concepts in a common space in order to better represent ESG concepts. The official task results show that our system yields a significant performance improvement compared to the baseline and outperforms all other submissions on the first sub-task. For the second sub-task, we combine the RoBERTa model with a feed-forward multi-layer perceptron in order to extract the context of sentences and classify them. Our model achieved high accuracy scores (over 92%) and was ranked among the top 5 systems.

1 Introduction

Financial markets and investors can support the transition to a more sustainable economy by promoting investments in companies complying to ESG (Environment, Social and Governance) rules. Today there is growing interest among investors in the performances of firms in terms of sustainability. Therefore, the automatic identification and extraction of relevant information regarding companies' strategy in terms of ESG is important. The use of NLP (Natural Language Processing) methods adapted to the field of finance and ESG could help identify and process related information.

Taxonomies are important NLP resources, especially for semantic analysis tasks and similarity measures(Vijaymeena and Kavitha, 2016; Bordea et al., 2016). In this context, the FinSim4-ESG Shared Task proposed the tasks of enrichment of ESG taxonomy and sentences classification. FinSim-4 is the fourth edition of a set of evaluation campaigns that aggregate efforts on text-based needs for the Financial domain (Maarouf et al., 2020; Mansar et al., 2021; Kang et al., 2021). This latest edition is particularly challenging due to the continuously evolving nature of terminology in the domain-specific language of the ESG which leads to a poor generalization of pre-trained word and sentence embeddings.

Several studies addressed the problem of taxonomy generation for different domains (Shen et al., 2020a; Karamanolakis et al., 2020). Deep learning based embedding networks, such as BERT (Devlin et al., 2018) have proven to be efficient for many NLP tasks. Malaviya *et al.* (2020) used BERT for knowledge base completion and showed that BERT performs well for this task. Liu *et al.* (2020) used BERT to complete an ontology by inserting a new concept with the right relation. Kalyan and Sangeetha (2021) used sentence BERT (Reimers and Gurevych, 2019) to measure semantic relatedness in biomedical concepts and showed that sentence BERT outperforms corresponding BERT models. Shen *et al.* (2020b) used sentence BERT to build a knowledge graph for the biomedical domain and showed that it obtains the best results.

For the Shared Task FinSim-4, we proposed several strategies based on BERT language models. For the first sub-task, we proposed a model based on pre-trained Sentence-BERT models to project sentences and concepts in a common space in order to better represent ESG concepts. For the second sub-task, we combined the RoBERTa model with a feed-forward multi-layer perceptron to extract the context of sentences and classify them. Official results of our participation show the effectiveness of our models over the Shared Task FinSim-4 benchmark. In terms of accuracy, our best runs respectively ranked 1^{st} and 4^{th} for the sub-tasks 1 and 2

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 218 - 223

with scores 0.848 and 0.927, respectively.

The remainder of this paper is organized as follows. In Section 2, we present the shared task FinSim-4 and the datasets for both sub-tasks. Our proposed models are detailed in Section 3. The setup and official results are described in Section 4. Finally, Section 5 concludes this paper.

2 Shared Task FinSim-4

The FinSim 2022 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. The 2022 edition proposes two sub-tasks.

2.1 Sub-task 1: ESG taxonomy extension

The first sub-task aims to extend the 'Fortia ESG taxonomy' provided by the organizers. This taxonomy was built based on different financial data providers' taxonomies as well as several sustainability and annual reports. It has twenty five different ESG concepts that belong to the ESG, split as: environment, social or governance. The organizers provide a training set which consists of terms belonging to each concept. This training set is unbalanced as one can observe in Table 1 where one can find the number of terms for each concept in the train set.

Participants were asked to complete this taxonomy to cover the rest of the terms of the original 'Fortia ESG taxonomy'. For example, given a set of terms related to the concept 'Waste management' (e.g. Hazardous Waste, Waste Reduction Initiatives), participating systems had to automatically assign to it all other adequate terms.

2.2 Sub-task 2: Sustainability classification

The second sub-task aims to automatically classify sentences into sustainable or unsustainable sentences. A sentence is considered as sustainable if it semantically mentions the Environmental or Social or Governance related factors as defined in the Fortia ESG taxonomy. Table 2 summarizes the training data provided by the organizers.

3 Proposed strategies

3.1 Sub-task 1: ESG taxonomy extension

Semantic text similarity is an important task in natural language processing applications such as information retrieval, classification, extraction, question answering and plagiarism detection. This task consists in measuring the degree of similarity between two texts and to determine whether how semantically close they are (from completely independent to fully equivalent). In our case, the terms of a same concept are considered semantically equivalent. Siamese models have been shown to be effective on the semantic analysis of sentences (Linhares Pontes et al., 2018; Reimers and Gurevych, 2019).

Our model is based on Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), a modifi-

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

Table 1: Dataset description for the ESG taxonomy extension sub-task.

Classes	#sentences
Sustainable	1223
Unsustainable	1042

Table 2: Dataset description for the sustainability subtask.

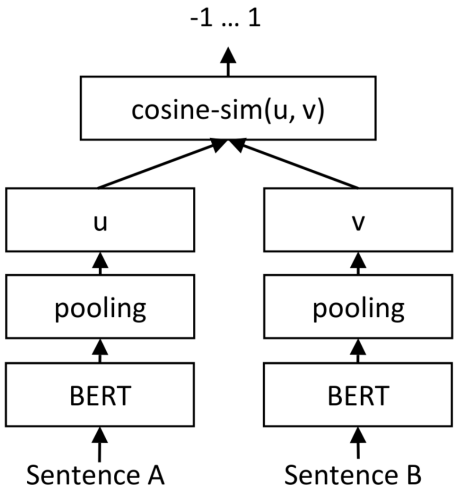

Figure 1: Sentence transformer architecture at inference to compute semantic similarity scores between two sentences.

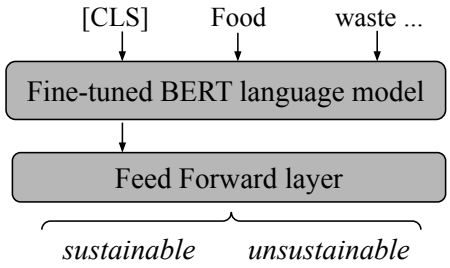

Figure 2: Architecture model for the sustainability classification task.

Sub-task	#Training	#Dev
ESG taxonomy extension	452	195
Sustainability classification	1585	680

Table 3: Details of the split of the 'Fortia ESG taxonomy' dataset to set our meta-parameters.

cation of the pre-trained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity (Figure 1). This model is trained on a parallel dataset where two paraphrases or similar semantic sentences have high cosine similarity.

We consider all terms about a concept as paraphrases because they share the same semantic information. For instance, the terms 'carbon footprint' and 'carbon data' should have similar sentence representation because they share the same concept 'carbon factor'; meanwhile, the terms 'Water Risk Assessment' and 'Transition to a circular economy' do not share the same concept and, consequently, their representations should have different sentence representation.

With the SBERT model, we project all terms on the same dimensional space and then, we train our logistic regression model[1] to analyze and classify them to their corresponding concept classes.

3.2 Sub-task 2: Sustainability classification

For this sub-task, we combine a BERT-based language model (Liu et al., 2019) with a feed-forward multi-layer perceptron to extract the context of sentences and classify them into 'sustainable' or 'unsustainable'. The architecture of our model is described in Figure 2.

We took the representation of the [CLS] token at the last layer of these models and we added a feed-forward layer to classify a input sentence as '*sustainable*' or '*unsustainable*'.

4 Experimental setup and evaluation

4.1 Evaluation metrics

All runs were ranked based on mean rank and accuracy for the first sub-task and only accuracy for the second sub-task. The mean rank is the average of the ranks for all observations within each sample.

Accuracy determines how close the candidates' predictions are to their true labels:

$$accuracy = \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} 1(\hat{y}_i = y_i), \quad (1)$$

where $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the corresponding true value.

4.2 Experimental evaluation

In order to select the best pre-trained models for each sub-task, we split the training datasets into 70% training and 30% for development. Table 3 shows the number of examples in the resulting training and development split for our analysis.

For the first sub-task, we selected the sentence BERT models: '*bert-base-nli-mean-tokens*'[2], '*all-roberta-large-v1*'[3], and '*paraphrase-mpnet-base-v2*'[4]. The first and second pre-trained SBERT

[1] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[2] https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

[3] https://huggingface.co/sentence-transformers/all-roberta-large-v1

[4] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

SBERT model	Mean rank	Accuracy
bert-base-nli-mean-tokens	1.502	0.764
all-roberta-large-v1	1.461	0.779
paraphrase-mpnet-base-v2	**1.349**	**0.810**

Table 4: Results of our approach (Section 3.1) using different SBERT models for the first sub-task.

BERT model	Accuracy
distilbert-base-uncased	0.906
bert-base-uncased	0.921
roberta-base	**0.922**

Table 5: Results of our approach (Section 3.2) using different BERT-based language models for the second sub-task.

models are based on the well-know BERT-based language models (BERT and RoBERTa language models, respectively). The third pre-trained model was trained on the paraphrase dataset where two paraphrases have close representation. Table 4 shows the results for each pre-trained model. The '*paraphrase-mpnet-base-v2*' achieved the best results for both metrics. We assume that the analysis of paraphrases is similar to the analysis of terms that share the same concept, which allowed this model to outperform the other models.

For the second sub-task, we selected the BERT language models: DistilBERT (Sanh et al., 2019), BERT, and RoBERTa. RoBERTa (Robustly Optimized BERT Pre-training Approach) is an extension of BERT with changes to the pre-training procedure (Liu et al., 2019). They trained their model with bigger batches and over more data with long sentences. They also removed the next sentence prediction objective and dynamically changed the masking pattern applied to the training data. In this case, the RoBERTa language model outperformed the other models (Table 5).

4.3 Official results

We submitted two runs for ESG taxonomy extension. The first run used the approach described in Section 3.1 to train our model on the training data (Fortia ESG taxonomy). For the second run, we extended the Fortia ESG taxonomy with our in-house ESG taxonomy[5] and we used the same procedure to train the model. Our ESG taxonomy consists of a total of 65 terms spread across 22 concepts. For both runs, we used the pre-trained SBERT model '*paraphrase-mpnet-base-v2*'.

[5]No terms from our ESG taxonomy appear in the test data set published by the organizers.

Team name	Mean rank	Accuracy
kaka_1	1.441	0.745
kaka_2	1.670	0.662
kaka_3	1.545	0.752
JETSONS_1	1.972	0.607
LIPI_subtask1_1	1.517	0.710
LIPI_subtask1_2	1.669	0.703
TCSWITM_1	1.462	0.772
TCSWITM_2	1.448	0.779
vishleshak_task1	1.614	0.683
Baseline1	2.276	0.462
Baseline2	1.524	0.745
ours_wo_extended_data	1.262	0.834
ours_with_extended_data	**1.255**	**0.848**

Table 6: Official results for the first sub-task. Our approaches are listed at the bottom of the table. The best results are in bold. Our model *ours_wo_extended_data* was trained on the original training data provided by the organizers and the version *ours_with_extended_data* was trained on the original data set combined with our taxonomy.

Official results for the first sub-task are listed in Table 6. Both of our runs achieved the best results for mean rank and accuracy. In fact, our siamese model provided a better semantic representation of terms and outperformed the other approaches. The extension of the training data with our taxonomy enabled our model to better analyze the context of terms and their corresponding concepts and, consequently, improved the accuracy of 0.014 points.

We also submitted two runs for the second sub-task. The first run follows the same idea described in Section 3.1 to represent the sentences by using SBERT. Then, the logistic regression classifies these sentence representations into only two classes: '*sustainable*' and '*unsustainable*'. The second run uses the deep-learning model described in Section 3.2. Our model uses the pre-trained RoBERTa language model and two feed-forward layers to classify a sentence into '*sustainable*' or '*unsustainable*'.

Official results for the second sub-task are listed in Table 7. Our runs achieved the fourth best result. The combination of fine-tuned RoBERTa language model and feed-forward layers outperformed both baselines as well as our run with SBERT and logis-

Team name	Accuracy
kaka_4	0.927
kaka_2	**0.946**
CompLx_1	0.936
FORMICA2_1	0.883
FORMICA2_2	0.888
LIPI_1	0.922
LIPI_2	0.932
TCSWITM_1	0.873
vishleshak_task2	0.912
JETSONS_1	0.927
Baseline1	0.497
Baseline2	0.819
ours_sbert_logistic_regression	0.907
ours_roberta_with_ffnn	0.927

Table 7: Official results for the second sub-task. Our approaches are listed at the bottom of the table. The best results are in bold.

tic regression. Our models performed well (over 92% accuracy) and was ranked among the top 5 systems (0.19 points below the best-performing system).

5 Conclusion

This paper described the joint effort of the L3i laboratory of the University of La Rochelle and the Trading Central Labs in the *Shared Task FinSim-4* evaluation campaign for the task of ESG in financial documents. For this task, we developed BERT-based models. Our model based on siamese sentence analysis achieved the best results for the first sub-task. For the second sub-task, our approach based on the RoBERTa model got the fourth position.

Acknowledgments

This work has been partially supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region, France.

References

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, pages 1081–1091.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. 2021. A hybrid approach to measure semantic relatedness in biomedical concepts. *arXiv preprint arXiv:2101.10196*.

Juyeon Kang, Ismail El Maarouf, Sandra Bellato, and Mei Gan. 2021. FinSim-3: The 3rd shared task on learning semantic similarities for the financial domain. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 31–35, Online. -.

Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. 2020. Txtract: Taxonomy-aware knowledge extraction for thousands of product categories. *arXiv preprint arXiv:2004.13852*.

Elvys Linhares Pontes, Stéphane Huet, Andréa Carneiro Linhares, and Juan-Manuel Torres-Moreno. 2018. Predicting the semantic textual similarity with Siamese CNN and LSTM. In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pages 311–320, Rennes, France. ATALA.

Hao Liu, Yehoshua Perl, and James Geller. 2020. Concept placement using bert trained by transforming and summarizing biomedical ontology structure. *Journal of Biomedical Informatics*, 112:103607.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

Ismail El Maarouf, Youness Mansar, Virginie Mouilleron, and Dialekti Valsamou-Stanislawski. 2020. 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. -.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2925–2933.

Youness Mansar, Juyeon Kang, and Ismail El Maarouf. 2021. The finsim-2 2021 shared task: Learning semantic similarities for the financial domain. WWW '21, page 288–292, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. 2020a. Taxoexpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *Proceedings of The Web Conference 2020*, pages 486–497.

Si Shen, Xiao Liu, Hao Sun, and Dongbbo Wang. 2020b. Biomedical knowledge discovery based on sentence-bert. *Proceedings of the Association for Information Science and Technology*, 57(1):e362.

MK Vijaymeena and K Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28.