Shared Task for Cross-lingual Classification of Corporate Social Responsibility (CSR) Themes and Topics

Yola Nayekoo[1], Sophia Katrenko[1], Véronique Hoste[2], Aaron Maladry[2], Els Lefever[2]
[1]EcoVadis, Avenue de la Grande Armée 43, Paris, France
[2]Language and Translation Technology Team, Ghent University, Groot-Brittanniëlaan 45, Ghent, Belgium
{ynayekoo, skatrenko}@ecovadis.com
{veronique.hoste, aaron.maladry, els.lefever}@UGent.be

Abstract

This paper provides an overview of the Shared Task for Cross-lingual Classification of CSR Themes and Topics. We framed the task as two separate sub-tasks: one cross-lingual multi-class CSR theme recognition task for English, French and simplified Chinese and one multi-label fine-grained classification task of CSR topics for Environment (ENV) and Labor and Human Rights (LAB) themes in English. The participants were provided with URLs and annotations for both tasks. Several teams downloaded the data, of which two teams submitted a system for both sub-tasks. In this overview paper, we discuss the set-up of the task and our main findings.

Keywords: multilingual CSR, multi-label classification, CSR theme detection

1. Introduction

Today, business organizations are expected to report on matters that affect environment, the economy and people (impact materiality) as well as matters that influence enterprise value (financial materiality). Corporations are held accountable for impacts across the entire value chain and recognize the need for sustainable procurement, to reduce the risk of supply chain disruption, protect their brands and reputation, and facilitate access to capital. As a consequence, there is a growing need and interest in processing Corporate Social Responsibility content originating from both business organizations and media. Laws and regulations such as FCPA in the US, Sapin II and the UK Bribery Act have made companies even more liable for knowing about sustainability infractions, yet the information is difficult to uncover, anticipate, and manage.

For over 16 years, EcoVadis has been measuring the quality of a company's sustainability management system through its policies, actions and results. It has been screening a large variety of specialized sources and newspapers to identify CSR-related content and assess it with respect to CSR themes and criteria (topics). A key distinguishing element of EcoVadis' sustainability monitoring platform is the integration of this external input to augment company-provided documentation and data sources. Sustainability analysts assess news items in a variety of languages (e.g., English, Spanish, French) on how they impact the quality and effectiveness of the sustainability management system or reflect positive innovation. The analyzed results are then integrated as part of the EcoVadis sustainability rating, and are displayed on the EcoVadis scorecard, which allows businesses to monitor the sustainability performance of their trading partners as well as their continuous improvement actions.

Despite the progress in automatic information extraction in the last decades, no datasets or methodologies are available yet aiming at automatic CSR theme detection. This shared task, which is co-organized by EcoVadis[1], a business sustainability ratings provider, and by the Language and Translation Technology Team[2] (LT3) from Ghent University provides the NLP community with data sets in multiple languages (English, French, and simplified Chinese) for CSR news analysis and will shed light on the feasibility of cross-lingual CSR theme detection. In addition, we also provide data sets to gain insights into fine-grained topic classification for two large CSR themes, viz. Environment (ENV) and Labor and Human Rights (LAB) in English.

The remainder of the paper is organized as follows. In Section 2, we discuss related research linked to the analysis of financial and social responsibility information. The shared task setup is described in depth in Section 3 and includes details on the two sub-tasks, dataset annotation and the experimental data selected for both tasks. In Section 4, we list the results of the participating teams for both task A and B. We end the paper with the main conclusions of the shared task and some prospects for follow-up research.

2. Related Research

The task of detecting corporate social responsibility themes and topics is operationalized as a classification task (cf. supra). Text classification is by nature the most fundamental task in NLP (Li et al., 2020). With the advent of deep neural networks,

[1] `www.ecovadis.com`

[2] `www.lt3.ugent.be`

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 283–291
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

and transformer-based language models in particular, approaches to text classification have drastically changed. More traditional machine learning models were feature-based, and incorporated manually crafted features relying on linguistic insights and knowledge from experts. Neural networks, however, utilize the text data itself to derive the embeddings used as input for the model. Deep learning models "integrate feature engineering into the model fitting process by learning a set of nonlinear transformations that serve to map features directly to outputs" (Li et al., 2020). These neural models have shown to perform very well for a wide range of NLP tasks, and they automatically provide meaningful semantic representations without the need of human-designed features or rules. They do, however, also come with important drawbacks: they require huge amounts of data and computational resources, and they are *black boxes*, viz. it is hard to investigate what information is really captured by the model, or to trace back why certain predictions (or errors) are made.

Detecting fine-grained CSR topics for Environment and Labor and Human Rights in English (Task B) is a multi-label classification task: the topic labels are nonexclusive and there are no restrictions about the number of classes that need to be assigned to the instances. Most recent multi-label text classification research has addressed this uncertainty of the number of labels, mainly by recasting the multi-label classification task into a multi-task problem (Lin et al., 2022). Another challenge, however, is to construct a better semantic representation space when feeding multi-label instances. As mentioned by Lin et al. (2023), the semantic space becomes "susceptible to distractions" when confronted with multi-label samples, and the boundaries between the classes become "blurred". They propose to deploy contrastive learning techniques to improve multi-label classification tasks.

While a large body of literature has already been devoted to the topic of CSR and CSR communication (Crane and Glozer, 2016), the application of natural language processing techniques in the domain of corporate social responsibility is fairly new and recently also gained some further visibility through the organization of the First Computing Social Responsibility Workshop which was held in collocation with LREC-2022 (Wan and Huang, 2022). Current NLP work on corporate social responsibility often deals with the collection and (automatic) analysis of CSR reports, i.e. regular reports published by a company or an organization about the economic, environmental and social impacts caused by its activities. The work on CSR reports among others describes the collection of corpora of CSR reports (Händschke et al., 2018; Purver et al., 2022), the analysis of financial and corporate social responsibility reports with respect to the Task Force on Climate-related Financial Disclosures (TCFD) questions that guide sustainability reporting (Luccioni et al., 2020), Global Reporting Initiative (GRI) topics detection from CSR reports (Polignano et al., 2022), the development of a Word Embedding-based Inclusion Model (WEIM) in CSR reports (Lu et al., 2022), etc.

CSR-related topics have furthermore also been investigated in social media, and among others, deal with sentiment analysis of Environmental, Social and Governance (ESG)-related social media posts (Park et al., 2022), such as for example the detection of human rights on social media (Pilankar et al., 2022).

However, to the best of our knowledge, there are no publicly available datasets that would enable CSR theme detection or a more fine-grained CSR topic detection per theme. Furthermore, the fact that the majority of studies have also been conducted on English with limited experimentation on other languages, motivated us to set up a shared task for cross-lingual classification of corporate social responsibility (CSR) themes and topics.

3. Shared Task Setup

3.1. Pilot study

To assess the feasibility of the task, we conducted a pilot study, resulting in over 1,034 annotated news items in English, 54 items in Spanish, over 250 items in French, and 24 articles in simplified Chinese. CSR theme detection includes the classification of news into one of four CSR themes:

1. Environment (ENV), which deals with factors that affect the natural environment such as carbon emissions, natural resources, energy efficiency, waste management, and raw material sourcing.

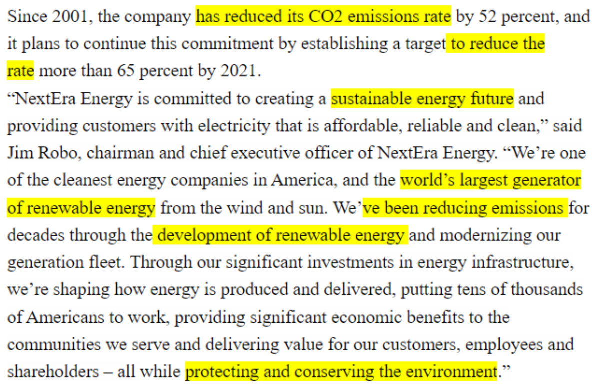
2. Labor and Human Rights (LAB), discussing topics such as human rights, labor standards, diversity and inclusion or career management and training.

3. Fair Business Practices (FBP), reflecting on anti-competitive practices, corruption, and responsible information management.

4. Sustainable Procurement (SUP), which includes supplier environmental and social practices

The pilot study showed that the ENV and LAB themes were predominant followed by FBP and SUP for all four languages. The most frequent topics within the ENV theme were *Materials, Chemicals, & Waste*, and *Environmental Services & Ad-*

vocacy, while *Employee Health & Safety* and *Labor Practices and Human Rights* were the most reported topics within the LAB theme. An overview of the different topics in both ENV and LAB themes is given in Table 1. In the case of CSR topic detection, we observed that articles may be assigned two labels, while the assignment of three or more labels was less common.



Since 2001, the company has reduced its CO2 emissions rate by 52 percent, and it plans to continue this commitment by establishing a target to reduce the rate more than 65 percent by 2021.

"NextEra Energy is committed to creating a sustainable energy future and providing customers with electricity that is affordable, reliable and clean," said Jim Robo, chairman and chief executive officer of NextEra Energy. "We're one of the cleanest energy companies in America, and the world's largest generator of renewable energy from the wind and sun. We've been reducing emissions for decades through the development of renewable energy and modernizing our generation fleet. Through our significant investments in energy infrastructure, we're shaping how energy is produced and delivered, putting tens of thousands of Americans to work, providing significant economic benefits to the communities we serve and delivering value for our customers, employees and shareholders – all while protecting and conserving the environment."

Figure 1: An example of CSR news (**CSR theme**: ENV, **CSR topic**: ENERGY CONSUMPTION & GHG).

An example of CSR news for the Environment CSR theme is given in Fig. 1. The parts in yellow are indicated by the annotators as triggers for the chosen label, but are not part of the shared task.

3.2. Task Description

The shared task includes two sub-tasks:

- **Task A:** Cross-lingual CSR theme recognition (English, French, simplified Chinese): cross-lingual, multi-class classification task with the following labels: Environment (ENV), Labor and Human Rights (LAB), Fair Business Practices (FBP), Sustainable Procurement (SUP).

- **Task B:** Fine-grained multi-label classification of CSR topics (English) for Environment (ENV) and Labor and Human Rights (LAB) themes.

Task A is framed as a multi-class classification task, for which participants output for each news article in the different languages a CSR label. **Task B** is a multi-label classification problem whereby an article may be assigned multiple topics from Table 1 within the specified theme (e.g., an article with two topics, *Air Pollution* and *Customer Health and Safety*, within the ENV theme). While we encouraged participants to contribute to both sub-tasks, they could also decide to participate in **Task A** or **Task B** only.

CSR theme	topic
ENV	Air Pollution
	Biodiversity
	Customer Health & Safety
	Energy Consumption & GHGs
	Environmental Services & Advocacy
	Materials, Chemicals & Waste
	Product End-of-Life
	Product Use
	Water
LAB	Career Management & Training
	Child Labor, Forced Labor & Human Trafficking
	Diversity, Equity, and Inclusion
	Employee Health & Safety
	External Stakeholder Human Rights
	Labor Practices and Human Rights
	Social Dialogue
	Social Discrimination
	Working Conditions

Table 1: List of CSR topics for the ENV and LAB themes.

3.3. Dataset Construction for the Shared Task

We aimed at collecting and annotating at least 1,500 publicly available English news articles with CSR themes for the training set and at least 500 news items per LAB and ENV CSR theme. Articles covering the two largest themes (ENV, LAB) were annotated with underlying CSR topics to produce the dataset for **Task B**. The datasets for both sub-tasks were constructed from publicly available content. As no personal data was used, we did not anticipate risks with respect to ethics, privacy or security.

Dataset quality and annotators In line with the pilot study, dataset quality was ensured by engaging highly qualified CSR experts as annotators, monitoring inter-annotator agreement and resolving disagreements. Every document was independently annotated by two trained CSR analysts and disagreements were resolved through discussion in pairs to arrive at the final list of annotations (Oortwijn et al., 2021).

Annotation scheme For cross-lingual CSR theme recognition, the annotation was done at the news item level whereby each URL was classified into one of four CSR themes: ENV, LAB, FBP, or SUP. The subset of news items labeled with the ENV and LAB themes was subsequently further annotated into one or more CSR topics. The data set shared with participants included news item URLs and the corresponding labels.

3.4. Experimental data

The annotated data was split using stratified random sampling to build training and test sets for English. For the remaining languages in **Task A**, only test sets were made available. The label distribution for the training (English) and test data (English, French and Chinese) from Task A is presented in Table 2 and the corresponding figures for Task B are given in Table 3. Recall that Task B was set up as a multi-label classification task. When we consider the distribution of the labels across both themes (Figure 2), we can observe that one or two labels were assigned to the large majority of instances, whereas up to 10% of the instances received three or even more labels.

	TRAIN	TEST		
	English	English	French	Chinese
ENV	708	164	70	70
FBP	197	48	21	25
LAB	662	149	70	40
SUP	41	2	1	0
Total	1608	363	162	135

Table 2: Label distribution for Task A

ENV	TRAIN	TEST
Air pollution	36	6
Biodiversity	62	11
Customers Health and Safety	62	19
Energy Consumption, GHGs	366	80
Env. Services & Advocacy	242	79
Materials,Chemicals, Waste	112	32
Product End of Life	73	20
Product Use	44	7
Water	71	16
LAB	TRAIN	TEST
Career Mgmt & Training	77	18
Child Labor, Forced Labor, Human Trafficking	7	1
Diversity, Equity, Inclusion	149	35
Employee Health, Safety	138	37
Ext. Stakeh. Human Rights	14	3
Labor Pract. & Human Rights	47	24
Social Dialogue	52	14
Social Discrimination	18	5
Working Conditions	201	60

Table 3: Label distribution for Task B

4. Methodology of Participating Teams

For this shared task, two teams submitted results for both sub-tasks: Team Kosar & Van Nooten (Van Nooten et al., 2024) and Team TredenceAICoE (Sharma et al., 2024).

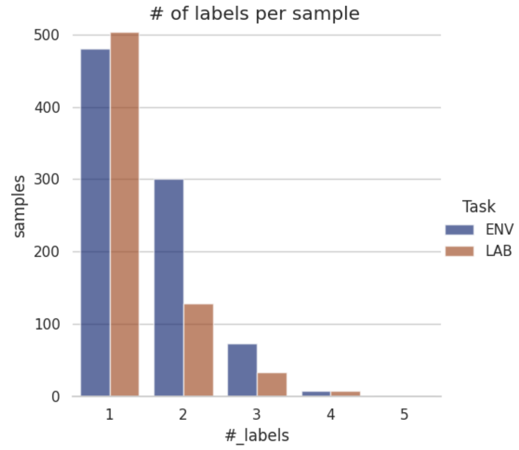


Figure 2: Number of samples with 1,2,3,4 or 5 labels for Task B (test and training data combined)

4.1. Data Collection, Cleaning and Augmentation

The two participating teams employed different libraries to scrape the content of the provided URLs. Kosar & Van Nooten used the Trafilatura library (Barbaresi, 2021), which was not able to scrape the content of all of the provided URLs. TredenceAICoE, on the other hand, used Newspaper3K[3] and was able to scrape not only the content of the pages but also the titles of the pages, which they leveraged as additional input information. After collecting the data, only Team Kosar & Van Nooten cleaned the data, using GPT 3.5[4].

Since the organizers only provided training data for English for both tasks, but Task A also involves testing on Chinese and French, both teams made use of data augmentation to obtain training and development data. Again, the two teams used different systems to translate the training data. Whilst Team Kosar & Van Nooten used the Google Translate API, Team TredenceAICoE used two language-specific transformer models to translate to French Helsinki-NLP/opus-mt-tc-big-en-fr (Tiedemann and Thottingal, 2020) and Chinese (Helsinki-NLP/opus-mt-en-zh).

In addition to generating data in the other languages, both teams used data augmentation to create additional samples to address class and label imbalance. Team TredenceAICoE used GPT 4[5] to generate samples for the minority classes and Team Kosar & Van Nooten used Mixtral[6] to

[3] https://github.com/codelucas/newspaper

[4] https://platform.openai.com/docs/models/overview

[5] https://platform.openai.com/docs/models/overview

[6] https://docs.together.ai/docs/

paraphrase each sample in the train set.

4.2. Methodologies

Both teams tested a wide variety of systems, with the most notable being zero-shot prompting with advanced prompts for GPT 3.5 and GPT 4 (Kosar & Van Nooten). Regardless, the best approach for both teams was still fine-tuning pre-trained transformer models.

For their final systems, Kosar & Van Nooten use the XLM-RoBERTa-large model for Task A and monolingual RoBERTa-large model for Task B (Conneau et al., 2020). Similarly,Team TredenceAICoE used MDeBERTa (He et al., 2023) for Task A (as it is multilingual) and Longformer (Beltagy et al., 2020) for Task B. As most scraped articles extend beyond the standard token length of 512, Longformer is a sensible model choice. However, this was not the only measure the team took to address the exceeding text length. They also divided the dataset into multiple sequences or chunks, and experimented with a Variable Selection Network (VSN) (Lim et al., 2021) to provide selected additional information to the classification layer for improved predictions.

In addition to creating augmented data to combat class imbalance, both teams also modified the training procedure. While Team TredenceAICoE made use of Dynamic Weighted Loss to increase the probabilities of underrepresented classes, Team Kosar & Van Nooten employed an advanced variation of contrastive learning that is specifically aimed at dealing with multi-label classification.

5. System Evaluation & Results

5.1. Evaluation

To evaluate system performance for Tasks A and B, the prediction of coarse-grained CSR themes and fine-grained CSR topics for environment and Labor and Human rights, we used standard evaluation measures, including accuracy, precision, recall and F1-score. The results are ranked according to weighted F1-score to account for the difference in sample sizes for English (363), Chinese (135) and French (162) and the different class distributions across the CSR themes and topics. However, in addition to the weighted F1-score, we also provide macro-averaged F1-scores to describe the performance on the labels with low sample counts. One of the participating teams (Kosar & Van Nooten) reported encountering issues while scraping the text for some of the test samples. As a result, they could not provide any predictions for 43 samples for Task A and 19 samples for Task B. Naturally, the results for Team Kosar & Van Nooten are better when only considering the samples they had access to. However, as the performance difference is small and does not impact the ranking for either of the tasks, we present the results on the complete test set. More concretely, if we assume the missing predictions are wrong for Task A, this results in a drop of 3% across all labels. For Task B, we assume that none of the labels are present in the prediction, resulting in a drop of 0.4% across all tasks and subsets for Team Kosar & Van Nooten.

	team	acc.	prec.	rec.	f-m	f-w
EN	A-J	0.90	0.96	0.90	0.61	0.93
	TRED	**0.95**	**0.97**	**0.95**	**0.77**	**0.96**
ZH	A-J	0.58	0.67	0.58	0.35	0.60
	TRED	**0.78**	**0.88**	**0.78**	**0.61**	**0.81**
FR	A-J	0.82	0.93	0.82	0.65	0.87
	TRED	**0.94**	**0.95**	**0.94**	**0.87**	**0.94**
avg.	A-J	0.76	0.86	0.76	0.54	0.80
	TRED	**0.89**	**0.93**	**0.89**	**0.75**	**0.90**

Table 4: Summary of the results for Task A with detailed information on the performance per language. Precision (prec.), recall (rec.) and F1 scores (f-w) are weighted averages across all classes. To describe the performance on minority classes, we also show macro-averaged F1 (f-m).

	team	acc.	prec.	rec.	f1
ENV	A-J	0.76	0.91	0.75	0.82
	TRED	**0.89**	**0.97**	**0.86**	**0.91**
FBP	A-J	0.76	0.88	0.66	0.75
	TRED	**0.89**	**0.81**	**0.98**	**0.88**
LAB	A-J	0.76	0.77	0.84	0.79
	TRED	**0.89**	**0.95**	**0.89**	**0.91**
SUP	**A-J**	0.86	**0.35**	**0.75**	**0.48**
	TRED	**0.95**	0.32	**0.75**	0.44
avg.	A-J	0.76	0.85	0.76	0.80
	TRED	**0.89**	**0.93**	**0.89**	**0.90**

Table 5: Summary of the results for Task A with detailed information on the performance per label. The concluding row with averaged values (*avg.*) reports the weighted averaged F1-score.

5.2. Results for Task A

As shown in Table 5, Team TredenceAICoE attained the highest overall scores for Task A with a macro-averaged F1-score of 75% and a weighted F1-score of 90% across all labels. Except for the SUP category - for which the test data, depending on the language, merely contains between 0 and 2 instances -, their system consistently outperformed the system of their competitors on the sustainability labels (as illustrated in Figure 4), but mostly attained these increased scores by performing better on French and Chinese (illustrated in Figure 3).

inference-models

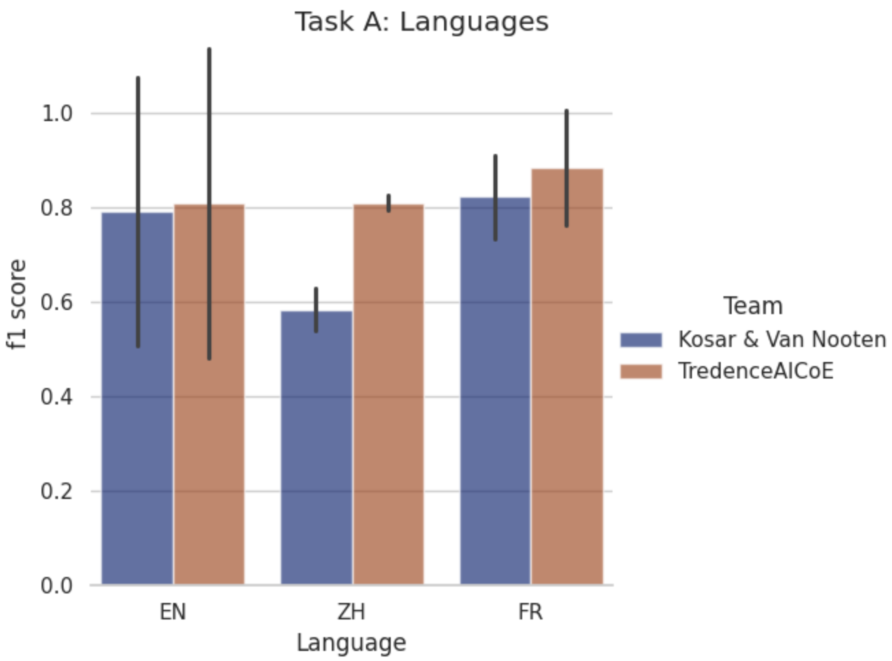


Figure 3: Weighted F1-scores on the test set for Task A (theme multi-class classification) per language. The vertical lines describe the standard deviation for the different labels.

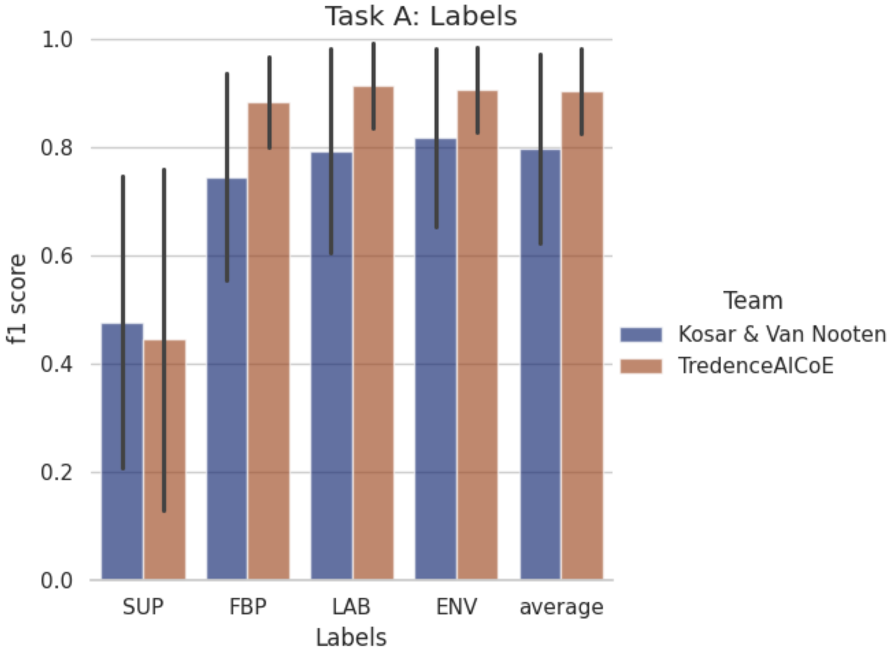


Figure 4: Weighted F1-scores on the test set for Task A (theme multi-class classification) for each individual label. The vertical lines describe the standard deviation for the different languages.

5.3. Results for Task B

For Task B, the fine-grained multi-label classification for ENV and LAB, Team Kosar & Van Nooten reached the highest overall weighted F1-score (*f-w*) of 88.1%. From their macro-averaged F1-score (*f-m*) of 73.8% (shown in Table 6), we could derive that across the two CSR themes, Team Kosar & Van Nooten is better at identifying the less prominent labels.

On the ENV sub-task, there is no significant difference in performance between the two teams, with weighted F1-scores of 87.3% vs 87.7% and an equal accuracy score of 87.5%. As shown in

	team	acc.	prec.	rec.	f-m	f-w
ENV	A-J	**0.875**	0.877	**0.875**	**0.745**	0.873
	TRED	**0.875**	**0.884**	**0.875**	0.725	**0.877**
LAB	**A-J**	**0.892**	0.888	**0.892**	**0.731**	**0.889**
	TRED	0.875	**0.901**	0.875	0.682	0.879
avg.	**A-J**	**0.884**	0.882	**0.884**	**0.738**	**0.881**
	TRED	0.875	**0.893**	0.875	0.704	0.878

Table 6: Summary of the results for Task B. Precision (prec.), Recall (rec.) and F1 scores (f-w) weighted averages across all labels. To describe the performance on minority classes, we also show macro-averaged F1 (f-m).

Figure 5, the teams had varying scores depending on the label, with Team Kosar & Van Nooten scoring 8% higher for “Biodiversity” and 10% lower for “Energy Consumption & GHGs”. However, along with the other labels, the score difference averaged out to 0.3% (weighted F1-score).

On the LAB sub-task, Team Kosar & Van Nooten did achieve a weighted F1-score of 88.9%, which is higher compared to the 87.9% score of the other team. On the LAB subset of Task B, the winning team scored 7.5% higher on the label for “Labor Practices and Human Rights”, while scoring 5% lower on the label for “Working Conditions” (illustrated in Figure 6). Along with some minor differences (both ups and downs) on the other labels, the notable performance difference on the label for “Labor Practices and Human Rights” seems to be the game-changer for Task B.

6. Conclusion

Both participating teams developed highly advanced systems that were directly modified to deal with the two specific sub-tasks. The modifications of Team Kosar & Van Nooten address class/label imbalance, cross-lingual transfer and multi-label cooccurence with data augmentation, machine translation and a special variant of contrastive learning for multi-label classification. Their competitors, Team TredenceAICoE, also employed data augmentation to create additional samples for the unseen languages and underrepresented classes. However, their efforts specifically address the exceeding text length of the articles using a Variable Selection Network for chunking.

For Task A, TredenceAICoE attained the highest score across all three classes. Their fine-tuned MDeBERTa most notably outperformed the other system on the Chinese subset of the data. Likely, their choice of using a transformer model for translation aided them in exceeding the results of the other team, who used the Google Translate API for translation.

For Task B, the advanced multi-label approach of Kosar & Van Nooten with their particular variant of contrastive learning allowed them to beat the performance of their opponents on the LAB subset of Task B.

7. Acknowledgments

The authors thank Clay Liu, Yi-Shiuan Iza Sun, Manisha Babubudjnauth, Ravi Ramruttun, Hans Lachanah, Viduchy Goorvadoo, Swastee Pomny for making the data sets available. This work was supported by Ghent University under grant BOF.24Y.2021.0019.01

8. Bibliographical References

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Andrew Crane and Sarah Glozer. 2016. Researching corporate social responsibility communication: Themes, opportunities and challenges. *Journal of Management Studies*, 53:7.

Sebastian G.M. Händschke, Sven Buechel, Jan Goldenstein, Philipp Poschmann, Tinghui Duan, Peter Walgenbach, and Udo Hahn. 2018. A corpus of corporate annual and social responsibility

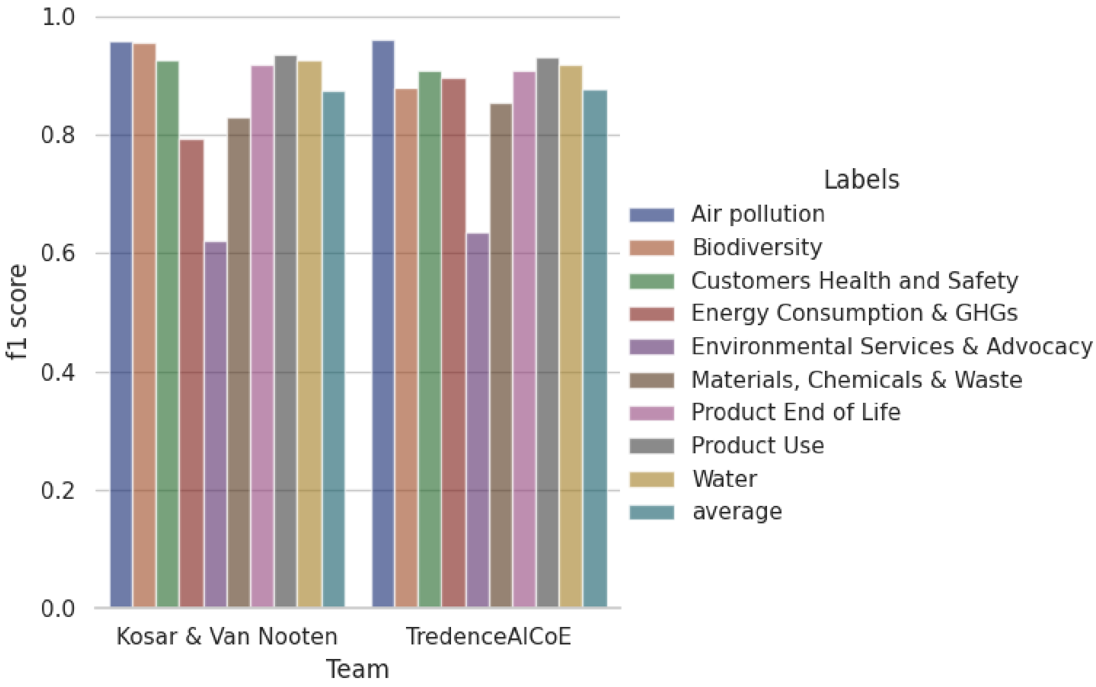


Figure 5: Weighted F1-scores on the test set for Task B (multi-label classification) for each individual label in the ENV subset.

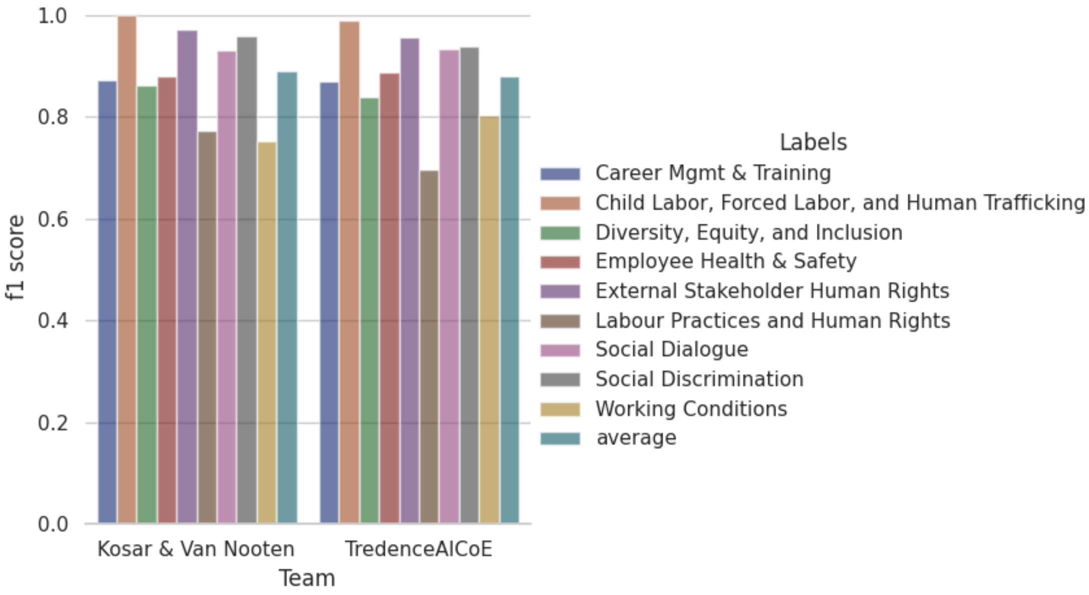


Figure 6: Weighted F1-scores on the test set for Task B (multi-label classification) for each individual label in the LAB subset.

reports: 280 million tokens of balanced organizational writing. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 20–31, Melbourne, Australia. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Qian Li, Hao Peng, Jianxin Li, Congyin Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2020. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13:1–41.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

Nankai Lin, Sihui Fu, Xiaotian Lin, and Lianxi Wang. 2022. Multi-label emotion classification based on adversarial multi-task learning. *Information Processing and Management*, 59(6).

Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. 2023. An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8730–8744, Toronto, Canada. Association for Computational Linguistics.

Lu Lu, Jinghang Gu, and Chu-Ren Huang. 2022. Inclusion in CSR reports: The lens from a data-driven machine learning model. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 46–51, Marseille, France. European Language Resources Association.

Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. 2020. Analyzing Sustainability Reports Using Natural Language Processing. *CoRR*, abs/2011.08073.

Yvette Oortwijn, Thijs Ossenkoppele, and Arianna Betti. 2021. Interrater Disagreement Resolution: A Systematic Procedure to Reach Consensus in Annotation Tasks. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 131–141, Online. Association for Computational Linguistics.

Joonbeom Park, Woojoo Choi, and Sang-Uk Jung. 2022. Exploring Trends in Environmental, Social, and Governance Themes and their Sentimental Value over Time. *Frontiers in Psychology*, 13.

Yash Pilankar, Rejwanul Haque, Mohammed Hasanuzzaman, Paul Stynes, and Pramod Pathak. 2022. Detecting violation of human rights via social media. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 40–45, Marseille, France. European Language Resources Association.

Marco Polignano, Nicola Bellantuono, Francesco Paolo Lagrasta, Sergio Caputo, Pierpaolo Pontrandolfo, and Giovanni Semeraro. 2022. An NLP Approach for the Analysis of Global Reporting Initiative Indexes from Corporate Sustainability Reports. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 1–8, Marseille, France. European Language Resources Association.

Matthew Purver, Matej Martinc, Riste Ichev, Igor Lončarski, Katarina Sitar Šuštar, Aljoša Valentinčič, and Senja Pollak. 2022. Tracking changes in ESG representation: Initial investigations in UK annual reports. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 9–14, Marseille, France. European Language Resources Association.

Shubham Sharma, Himanshu Janbandhu, and Ankush Chopra. 2024. Improving Cross-Lingual CSR Classification using Pretrained Transformers with Variable Selection Networks and Data Augmentation . In *Proceedings of the Joint Workshop of FinNLP-KDF-ECONLP@LREC-COLING 2024*, Torino, Italy.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Jens Van Nooten, Andriy Kosar, Guy De Pauw, and Walter Daelemans. 2024. Advancing CSR Theme and Topic Classification: LLMs and Training Enhancement Insights. In *Proceedings of the Joint Workshop of FinNLP-KDF-ECONLP@LREC-COLING 2024*, Torino, Italy.

Mingyu Wan and Chu-Ren Huang, editors. 2022. *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France.

9. Language Resource References