Scaling Sustainable Development Goal Predictions across Languages: From English to Finnish

Melany Macias, Lev Kharlashkin, Leo Huovinen, Mika Hämäläinen
Metropolia University of Applied Sciences
Helsinki, Finland
first.last@metropolia.fi

Abstract

In this paper, we leverage an exclusive English dataset to train diverse multilingual classifiers, investigating their efficacy in adapting to Finnish data. We employ an exclusively English classification dataset of UN Sustainable Development Goals (SDG) in an education context, to train various multilingual classifiers and examine how well these models can adapt to recognizing the same classes within Finnish university course descriptions. It's worth noting that Finnish, with a mere 5 million native speakers, presents a significantly less-resourced linguistic context compared to English. The best performing model in our experiments was mBART with an F1-score of 0.843.

1 Introduction

The list of 17 sustainable development goals (SDGs) established by the United Nations (UN) has gained significance in assessing the societal, humanitarian, and environmental impact of companies in EU. This is particularly relevant for large companies compelled to include robust sustainability reporting in their annual reporting to authorities[1]. As a response to this growing importance, numerous universities and educational institutions have incorporated the UN SDGs into their academic curricula. This development prompts a crucial inquiry into how educational institutions, at a higher administrative level, can ascertain which specific SDGs are being integrated into different degree programs (see Kopnina 2020; Chankseliani and McCowan 2021).

To address this concern, adopting smaller local models —specifically distil-mBERT, mBERT, mBART, and XLM-RoBERTa— tailored to the content of course descriptions proves beneficial for universities. Such an approach facilitates compliance with the General Data Protection Regulation (GDPR)[2], ensuring the preservation of data privacy more than reliance on commercial large language models. The selection of the models stems from a need to accommodate linguistic diversity and the intricacies of academic content, ensuring accurate SDG classification while respecting data confidentiality. Each model brings distinct advantages: distil-mBERT and mBERT for their efficiency and language coverage, and mBART and XLM-RoBERTa for their superior cross-lingual and contextual understanding capabilities, making them well-suited for analyzing Finnish and English course descriptions in the context of SDGs.

In our work, we experiment with the viability of producing SDG classification data automatically using a large language model (LLM) in English. We use this English only data to train several multilingual classifiers and study the scalability of these models to Finnish data. Finnish, with only 5 million native speakers, is considerably less resourced than English. This is not to say that Finnish would be particularly under-studied in the context of NLP, given the vast amount of different NLP applications available for the language (see Hämäläinen and Alnajjar 2021).

2 Related work

The examination of sustainable development within the realm of NLP has been approached from various perspectives, including investigations into fairness in NLP (Hessenthaler et al., 2022), exploration of poverty and societal sustainability through interviews (van Boven et al., 2022), analysis of argumentation mining (Fergadis et al., 2021), and community profiling (Conforti et al., 2020), among other aspects. Our approach distinguishes itself by striving to encompass all UN sustainable devel-

[1] https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 132–137

opment goals and implementing them within an educational framework.

The effect of multilingual models has been studied before in several contexts such as sentiment analysis (Hämäläinen et al., 2022) and persuasion detection (Pöyhönen et al., 2022) using parallel data in several languages. The findings suggest that translation strategies have a huge impact on the performance of the models.

Regarding the educational aspect of our study, there exists ample prior research on integrating Sustainable Development Goals (SDGs) into teaching methods (Collazo Expósito and Granados Sánchez, 2020; Rajabifard et al., 2021; Kwee, 2021). However, this previous research is non-computational, and as far as we are aware, there is no existing computational research on this subject from the standpoint of Natural Language Processing (NLP).

3 SDG Data

Our dataset, containing 5988 entries from Metropolia University, was accessed via their API[3], a resource available to staff and researchers upon formal request and approval by the university's IT department. In acquiring and handling this data, we observed strict ethical standards, including anonymization of identifiable information and adherence to the university's data use policies, ensuring the preservation of data confidentiality. This dataset, spanning from 2010 to 2023, encompasses a diverse range of courses across various departments and majors, offered in both Finnish and English. The time frame was chosen to provide a comprehensive collection of course materials, aiding in the robustness of our model training.

After the initial data collection, we utilized the Vertex AI API[4] to conduct batch processing, a pivotal step in annotating each course description with the corresponding Sustainable Development Goals (SDGs). The API facilitated the automation of this task by allowing us to process large volumes of text data and generate labels that indicate the relevance of specific SDGs to the course content.

Our study focuses on the following SDGs: 3 (Good Health and Well-being), 7 (Affordable and Clean Energy), 8 (Decent Work and Economic Growth), 9 (Industry, Innovation, and Infrastructure), and 10 (Reduced Inequalities), selected for their significant relevance to the university curriculum, as other goals were less represented, leading to potential data imbalance. The data is formatted as JSON objects for multilabel classification with multilingual models; one typical entry would look as follows: *"input": "Heat Distribution Systems in Buildings, the student learns: The building's heating power demand and its calculation. Different heat distribution methods and devices. Ways of adjusting the heating system. Water-circulating radiator and floor heating system. Dimensioning of pipework and selection of radiator, circulation pump, expansion vessel and safety devices. The student can calculate the heating power demand of the building, can dimension the pipework, radiators, circulation pump and expansion vessel and safety devices.", "labels": [0, 1, 0, 0, 0]*, where the input text is related to goal 7 Affordable and clean energy.

Therefore, the entries are structured as "input" - *a detailed course description* and "labels" - *binary encoding indicating the course's relevance to the selected SDGs.*

Figure 2 displays the distribution of the Sustainable Development Goal (SDG) mentioned in the training dataset.

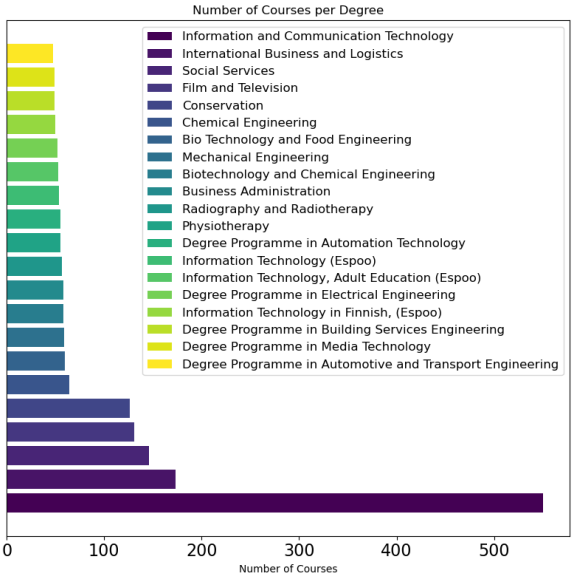

Figure 1: Distribution of courses per degree after the initial cleaning step.

4 Cross-lingual Models for SDG Prediction

In this study, we utilized four advanced multilingual models: Distil-mBERT (Sanh et al., 2019),

[3]https://wiki.metropolia.fi/display/opendata/REST-rajapinnat

[4]https://cloud.google.com/vertex-ai/docs/reference/rest

Model	Strengths	Weaknesses
Distil-mBERT	Efficient with significant language understanding	Nuance understanding may be limited
mBERT	Comprehensive language comprehension	Finnish-specific tuning may be less
mBART	Strong in deep contextual understanding	High computational requirements
XLM-RoBERTa	Excellent in cross-lingual tasks	Possible compromise in Finnish depth

Table 1: Comparative Analysis of Multilingual Models

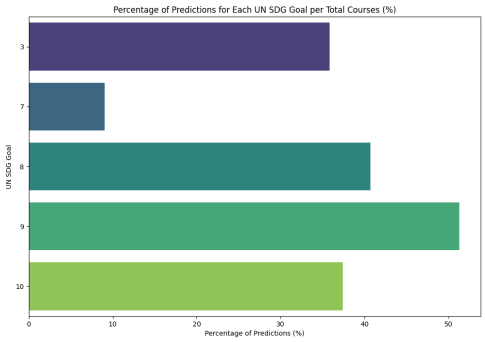

Figure 2: Distribution of SDG mentions within the training dataset.

mBERT (Devlin et al., 2019), mBART (Tang et al., 2020), and XLM-RoBERTa (Conneau et al., 2020), each uniquely suited for processing Finnish, a language with limited NLP resources. These models were selected for their balance between computational efficiency and linguistic depth, crucial for handling Finnish.

Our dataset, split into 70% training, 15% validation, and 15% testing, primarily consisted of English for training and validation, with Finnish reserved for testing. This approach was intended to test the models' transfer learning capabilities from English to Finnish.

Each model underwent fine-tuning for multilabel SDG classification using a PyTorch-based (Paszke et al., 2019) framework. Key steps in the training process included tokenization, encoding, and employing a BCEWithLogitsLoss function, as shown in Equation (1).

$$\text{BCELoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\sigma(x_i)) + (1 - y_i) \cdot \log(1 - \sigma(x_i))] \quad (1)$$

where $\sigma(x_i)$ is the sigmoid function applied to the model's output for the i^{th} sample, y_i is the true label, and N is the number of samples.

This study not only demonstrates the effectiveness of these models in a multilingual context but also sheds light on the scaling behavior of LLMs, particularly in adapting from high-resource to low-resource languages. The findings provide valuable insights into the adaptability of multilingual models, with a special focus on Finnish, illustrating the broader applicability of these models in diverse linguistic settings.

The exploration of these multilingual models in predicting SDGs in Finnish highlights significant insights into the scaling behavior and adaptability of LLMs. Our strategic data split, along with the tailored training and architecture of each model, demonstrates our approach in tackling the challenges of language representation within NLP. This methodology is particularly pertinent in understanding how LLMs perform across languages with varying resource levels.

5 Results and Evaluation

Table 2 provides an overview of the model's performance based on micro-average scores. The F1-Score, which balances precision and recall, showcases the models' effectiveness in multilabel classification across Sustainable Development Goals (SDGs). Notably, the mBART outperforms the others with an F1-Score of 0.843, indicating its robustness in handling diverse SDGs.

Model	Precision	Recall	F1-Score
distil-mBERT	0.749	0.557	0.547
mBERT	0.798	0.678	0.716
mBART	**0.825**	**0.867**	**0.843**
XLM-RoBERTa	0.842	0.824	0.829

Table 2: Models performance based on the micro scores

In addition, figure 3 depicts the F1 scores, which represent the harmonic mean of precision and recall for each SDG. Our in-depth analysis of the classifiers' performance on a per-label (per SDG) basis provides subtle insights into their prediction capability and limitations.

These per-label findings highlight the importance of model architecture and training corpus diversity in addressing the unique linguistic issues given by each SDG. The variations in performance

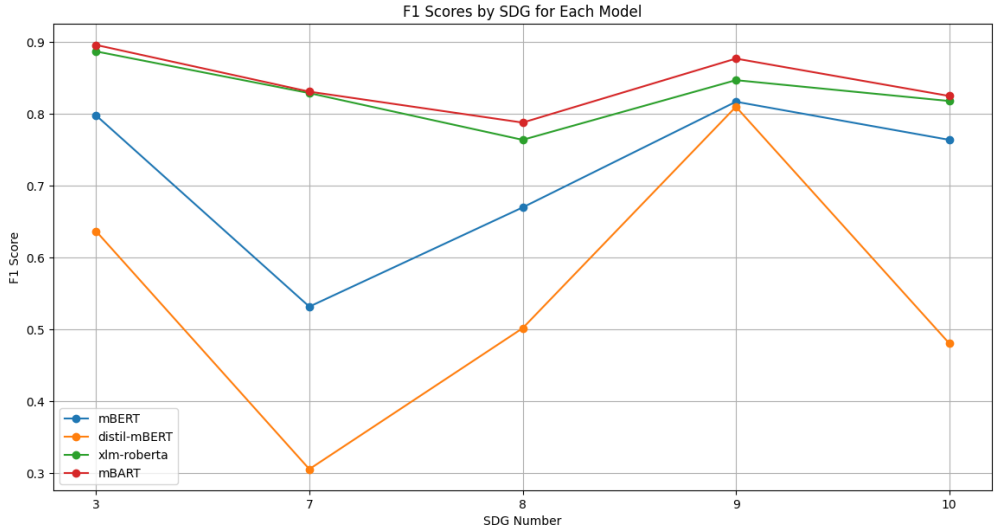

Figure 3: F1 Scores by SDG for Each Model

across objectives indicate that, while certain SDGs are well-represented and easier to forecast with current NLP models, others require additional research and targeted data enrichment to improve model performance.

6 Conclusions

This study takes a novel approach to assessing the integration of the United Nations Sustainable Development Goals (SDGs) into university curricula. We investigated the effectiveness of multilingual classifiers in adapting to Finnish data, a language with significantly fewer resources in the field of Natural Language Processing (NLP). Our study sought to ascertain the scalability of these models in recognizing SDG-related content within Finnish university course descriptions.

The findings show that multilingual models can bridge the language gap effectively, with notable success in identifying SDG-related content across multiple languages. The performance varied across SDGs, with some models excelling in some areas while struggling in others. This variation emphasizes the significance of model selection based on the target language's specific characteristics and linguistic nuances.

Our work, which aligns with workshop talks on scaling behavior across linguistic settings, advances the NLP community's understanding of multilingual model applicability in resolving language resource inequities by utilizing Finnish as a case study. The study opens the door to further investigation into the use of multilingual models for other low-resource languages.

In conclusion, the study successfully demonstrates the feasibility of using English-trained multilingual models to process and analyze data in Finnish, a language with limited resources. This method not only provides a useful tool for educational institutions' sustainability reporting, but it also contributes to a better understanding of multilingual NLP applications.

7 Limitations

While our study provides insights into the use of multilingual models for Sustainable Development Goal (SDG) prediction in Finnish, it is crucial to consider several limitations:

- **Subset of UN SDGs**: Our research focused on a subset of the UN SDGs. Extending our approach to encompass all SDGs would provide a more comprehensive understanding of the models' capabilities across a broader range of sustainability topics.

- **Model Size and Performance Trade-offs**: We employed models like 'mBART-large-50', 'distilbert-base-multilingual-cased', 'bert-base-multilingual-cased', and 'xlm-roberta-base', each varying significantly in size and architecture. A detailed comparative analysis of these models reveals notable trade-offs

between their sizes and their precision and F1 scores. Larger models tend to offer better performance but at the cost of increased computational resources and complexity. This aspect is particularly relevant in the context of scaling behavior in LLMs.

- **Language Resource Limitations**: While Finnish is considered a low-resourced language in NLP, it is still better represented than many of the world's approximately 7,000 languages. Our findings for Finnish may not directly translate to other low-resource languages, especially those with very limited digital presence or NLP tools.

- **Domain Specificity**: Our study was confined to the academic context of Metropolia University of Applied Sciences. The models' performance may not generalize to other educational institutions, especially those offering a different range of academic disciplines.

- **Potential Model Biases**: Our classifiers, while effective in a controlled environment, may have learned an oversimplified version of the problem domain. There is also a risk of unknown biases when these models are applied in real-world settings.

In light of these limitations, our study should be viewed as a stepping stone towards understanding the scalability and adaptability of multilingual models in handling low-resourced languages, particularly in the domain of educational sustainability reporting.

8 Ethics statement

Our study's focus on Finnish and English in multilingual models raises concerns about their performance and potential biases in other languages, especially those underrepresented in NLP research. While we ensured adherence to data privacy and consent in our methodology, the limited scope, centered on one university's curriculum, may not fully represent other educational contexts or disciplines.

The findings highlight the need for broader linguistic representation in NLP models to ensure fairness and mitigate biases. Future research should extend to diverse languages and educational settings, adhering to ethical research standards and prioritizing equitable representation in NLP applications.

References

Maia Chankseliani and Tristan McCowan. 2021. Higher education and the sustainable development goals. *Higher Education*, 81(1):1–8.

Leslie Mahe Collazo Expósito and Jesús Granados Sánchez. 2020. Implementation of sdgs in university teaching: a course for professional development of teachers in education for sustainability for a transformative action. *Sustainability*, 12(19):8267.

Costanza Conforti, Stephanie Hirmer, Dai Morgan, Marco Basaldella, and Yau Ben Or. 2020. Natural language processing for achieving sustainable development: the case of neural labelling to enhance community profiling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8427–8444, Online. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mika Hämäläinen and Khalid Alnajjar. 2021. The current state of finnish nlp. In *Proceedings of the Seventh International Workshop on Computational Linguistics of Uralic Languages*, pages 65–72.

Mika Hämäläinen, Khalid Alnajjar, and Thierry Poibeau. 2022. Video games as a corpus: Sentiment analysis using fallout new vegas dialog. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*, pages 1–4.

Marius Hessenthaler, Emma Strubell, Dirk Hovy, and Anne Lauscher. 2022. Bridging fairness and environmental sustainability in natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages

7817–7836, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Helen Kopnina. 2020. Education for the future? critical evaluation of education for sustainable development goals. *The Journal of Environmental Education*, 51(4):280–291.

Ching Ting Tany Kwee. 2021. I want to teach sustainable development in my english classroom: A case study of incorporating sustainable development goals in english teaching. *Sustainability*, 13(8):4195.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Teemu Pöyhönen, Mika Hämäläinen, and Khalid Alnajjar. 2022. Multilingual persuasion detection: Video games as an invaluable data source for nlp. In *Proceedings of the 2022 DiGRA International Conference*. DiGRA.

Abbas Rajabifard, Masoud Kahalimoghadam, Elisa Lumantarna, Nilupa Herath, Felix Kin Peng Hui, and Zahra Assarkhaniki. 2021. Applying sdgs as a systematic approach for incorporating sustainability in higher education. *International Journal of Sustainability in Higher Education*, 22(6):1266–1284.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Goya van Boven, Stephanie Hirmer, and Costanza Conforti. 2022. At the intersection of NLP and sustainable development: Exploring the impact of demographic-aware text representations in modeling value on a corpus of interviews. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2007–2021, Marseille, France. European Language Resources Association.