

# Predicting Sustainable Development Goals Using Course Descriptions - from LLMs to Conventional Foundation Models

Lev Kharlashkin, Melany Macias, Leo Huovinen and Mika Hämäläinen<sup>1</sup>

<sup>1</sup>Metropolia University of Applied Sciences, Finland

Corresponding author: Lev Kharlashkin , lev.kharlashkin@metropolia.fi

## Abstract

We present our work on predicting United Nations sustainable development goals (SDG) for university courses. We use an LLM named PaLM 2 to generate training data given a noisy human-authored course description input as input. We use this data to train several different smaller language models to predict SDGs for university courses. This work contributes to better university level adaptation of SDGs. The best performing model in our experiments was BART with an F1-score of 0.786.

## Keywords

SDG, multi label classification, LLM

## I INTRODUCTION

The United Nations (UN) has established a list of 17 sustainable development goals (SDGs)<sup>1</sup>. These goals are becoming more and more important in understanding the societal, humanitarian and environmental impact of companies in the EU given that certain large companies need to take rigorous sustainability reporting as part of their annual reporting to the authorities<sup>2</sup>.

Because of the growing importance, many universities and educational institutions have started to adopt the UN SDGs as part of their academic curricula. This raises the important question of how an educational institution can, on a higher level, know which SDGs are being taught and where in the different degree programs.

Adopting local models that are adjusted based on their course descriptions helps universities to comply with GDPR<sup>3</sup> and preserve data privacy. Additionally, by tailoring these models to the unique linguistic and curriculum quirks of the school, prediction accuracy can be increased while maintaining the security and confidentiality of sensitive data.

In our paper, we collect and clean a noisy course description dataset and use an LLM to generate SDGs for each course. We manually check and fix the LLM generated data that is used for testing. Furthermore, we fine-tune several smaller foundation models to predict SDGs based on course descriptions. Fine-tuning a smaller model makes the SDG prediction task faster and more cost-efficient.

---

<sup>1</sup><https://sdgs.un.org/goals>

<sup>2</sup>[https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting\\_en](https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting_en)

<sup>3</sup><https://eur-lex.europa.eu/eli/reg/2016/679/oj>

## II RELATED WORK

Sustainable development has been studied in the field of NLP from many different points of view such as studying fairness in NLP [Hessenthaler et al., 2022], studying poverty and societal sustainability in interviews [van Boven et al., 2022], argumentation mining [Fergadis et al., 2021] and community profiling [Conforti et al., 2020] among others. Our take differs from these in the sense that we aim to cover all UN sustainable development goals and apply them in a pedagogical context.

Perhaps the most similar prior work to ours is that of Amel-Zadeh et al. [2021]. They used more traditional methods such as word2vec [Mikolov et al., 2013] and doc2vec [Le and Mikolov, 2014] to assess how well companies align with UN SDGs. They use a dictionary of SDG goal related terms to assess the overlap of each SDG with a given company. They then train a logistic classifier, an SVM and a fully connected neural network on the embeddings. Their finding was that a combination of doc2vec and SVM gave the best results.

In terms of the pedagogical context of our research, there is plenty of prior research on incorporating SDGs as part of teaching [Collazo Expósito and Granados Sánchez, 2020, Rajabifard et al., 2021, Kwee, 2021]. This prior research is non-computational and to best of our knowledge, there is no prior research work on the topic from the NLP stand point.

## III DATA

For our work, we gathered course information from Metropolia University of Applied Sciences over their API<sup>4</sup>. This data retrieved was composed of 51,386 Finnish and English courses from the years 2004 to 2023.

### 3.1 Data Preprocessing

The dataset presented significant challenges in terms of variability and noise, attributed to the subjective nature of course descriptions provided by individual instructors. The length of these descriptions varied widely, and in some cases, the course objectives were either missing or contained redundant or extraneous information.

The study focused on courses offered between 2021 and 2023 to capture recent curricular trends. We imposed a character limit of 500 to 2000 for the combined length of course descriptions and objectives to maintain an optimal balance of detail and brevity. Courses outside this range were excluded. Moreover, the study was confined to courses conducted in English to maintain consistency in language processing. At this point, the data was composed of 8708 courses, with 103 unique disciplines.

Figure 1 depicts the distribution of English courses per the top 15 degrees after the initial cleaning step, which illustrates the diverse curricular offerings within the analyzed period. Notably, the 'Information and Communication Technology' discipline demonstrates a significantly higher volume of courses, underscoring the sector's expansion and its pivotal role in contemporary education landscapes. This visual representation also serves to highlight the curricular focus areas that are apparent within the institution, guiding the subsequent analysis stages to probe into the qualitative aspects of course content more deeply.

Our standardization process involved several steps, specifically the removal of entries with missing course descriptions or objectives, language detection using the Spacy NLP library [Honnibal

---

<sup>4</sup><https://wiki.metropolia.fi/display/.opendata/REST-rajapinnat>

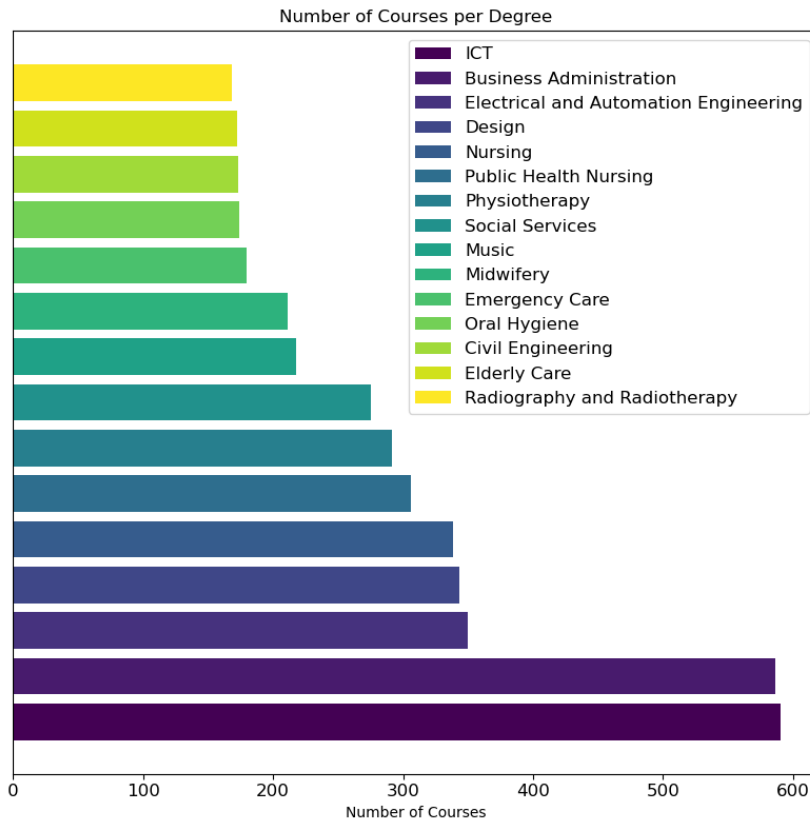


Figure 1: Distribution of courses per degree after the initial cleaning step.

and Montani, 2017] to remove Finnish courses and retain only English courses, and elimination of duplicates present from courses offered in multiple years.

The dataset, in its finalized state, consisted of 2125 courses in English, each defined by three key elements: name, description, and objective.

### 3.2 Generating SDGs

We use PaLM 2 [Anil et al., 2023] over Vertex AI API<sup>5</sup> to generate the training data for the SDG prediction models. In particular, we use *text-bison-32k* from Model Garden<sup>6</sup>. PaLM 2 is an LLM that takes in a prompt and produces an output based on the prompt in a similar fashion to ChatGPT [OpenAI, 2023].

In search of the most effective prompt, we employed the prompting IDE tool Prompterator [Sučik et al., 2023]. Thus, to ensure the quality of the model’s outputs, we took a small sample of data for batch processing and manually reviewed the model’s responses using Prompterator.

This evaluation helped us confirm the appropriateness of the SDG predictions for subsequent training. Moreover, batch processing was instrumental in handling the dataset efficiently, allowing for the dynamic integration of each course’s metadata into the prompt template. The responses collected from the model included the SDG goals deemed most relevant by the LLM, as shown in Table 1.

Our final prompt to the model was the following one appended with a course description of

<sup>5</sup><https://cloud.google.com/vertex-ai/docs/reference/rest>

<sup>6</sup><https://cloud.google.com/vertex-ai/docs/generative-ai/learn/models>

Parameter	Value
Prompt	Your goal is to identify UN SDG Goals relevant to students. Given a course name, the student learns: course content and course objective. Answer the question: What are the top few most relevant sustainable development goals to this course? Your task is to return only the numbers of the top few goals separated by commas. Also, never use the goal number 4.
Temp.	0.2
Token Limit	500

Table 1: Final prompt specification used for SDG goal generation

each course: *Your goal is to identify UN SDG Goals relevant to students. Given a course name, the student learns: course content and course objective. Answer the question: What are the top few most relevant sustainable development goals to this course? Your task is to return only the numbers of the top few goals separated by commas. Also, never use the goal number 4.*

The purpose of selecting a lower temperature setting (0.2) was to limit the output variability of the model and promote accuracy. We discovered empirically that a token limit of 500 was adequate to enable the model to produce thorough answers without being overly verbose.

### 3.3 Data Preparation for Training

After SDG predictions were generated by the LLM, the dataset underwent a meticulous cleaning process. Initially, labels generated by the language model were extracted, stripping away the prompts included in the input. Subsequently, we refined the labels to solely represent SDG numbers, with a particular exclusion of Goal 4: Quality education. This goal was excluded because it was over-represented in the data - virtually every single university course contributes to quality education.

For compatibility with multi-label classification models, we encoded the list of SDGs relevant to each course into a binary format. Consequently, the dataset for model training comprised two components: the input—encompassing the course name, description, and objectives—and the output—a binary vector denoting pertinent SDGs.

An evaluation of the SDG distribution within the training data was conducted to ascertain dataset quality and representation balance, the results of which are depicted in Figure 2.

The percentage distribution of the model-generated SDG forecasts is shown in Figure 2. This figure omits Goal 4 (Quality Education) by design and reveals that certain goals, such as 2 (Zero Hunger), 14 (Life Below Water), and 15 (Life on Land), are less frequently associated with the course descriptions at Metropolia. This skewness in the data reflects the varied emphasis of SDGs in the actual course content.

After the quality of the dataset was confirmed, it was split 70:15:15 into subsets for training, validation, and testing. This allocation prevents overfitting during training and enables a thorough assessment of the model’s performance across unknown data.

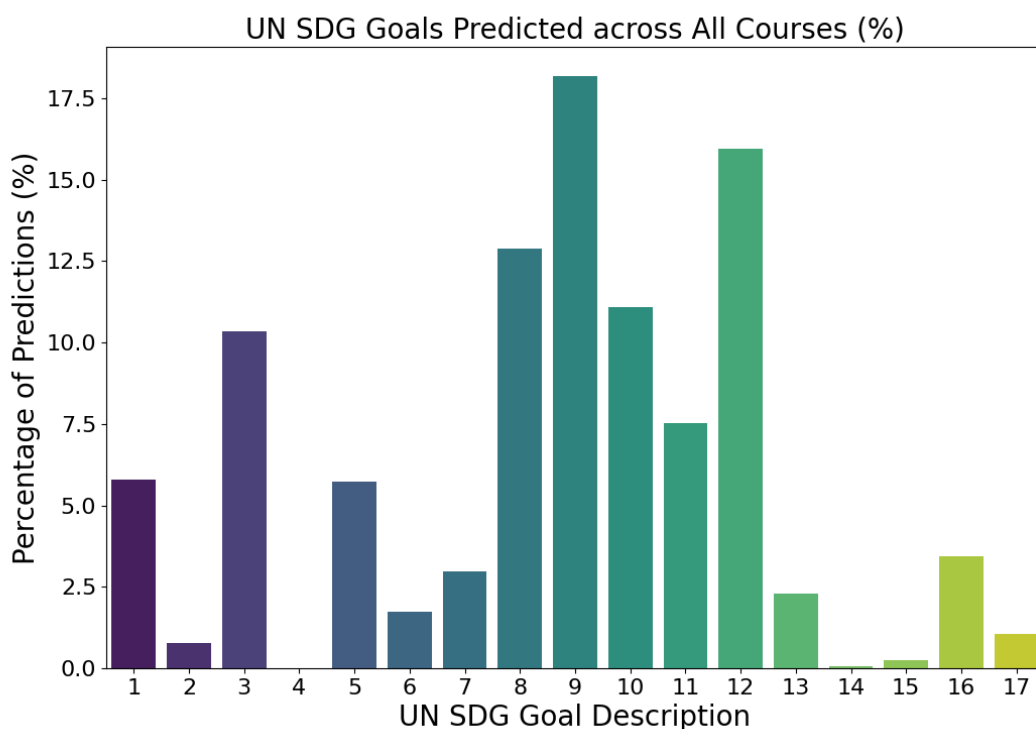


Figure 2: Distribution of SDG mentions within the training dataset.

#### IV SDG PREDICTION MODELS

We try out fine-tuning several models for multi-class classification task using Transformers Python library [Wolf et al., 2020]. We selected BERT [Devlin et al., 2019], mBERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019], XLM-RoBERTa [Conneau et al., 2020], and BART [Lewis et al., 2019] due to their top-tier results in multi-label classification tasks. XLM-RoBERTa and mBERT, in particular, were chosen to explore their capabilities of a multilingual model for potential future purposes.

The models are trained to receive course information as input and are trained to predict the corresponding 3 most relevant SDG goals as output. This aligns with the format of our training data. An illustrative example of the input data and the model's expected output is presented in Table 2. This binary output represents the relevance of specific SDG goals to the given course information, with the example indicating relevance to goals 3, 5 and 8.

Field	Value
Input	"Clinical Practice, the student learns: Clinical Practice in nursing environment. Students can apply the theoretical and clinical competence required by the clinical practice environment to the nursing care of clients/patients- can maintain and promote the health of clients/patients and their significant others in a client-oriented way in nursing care- follow the ethical guidelines and principles of nursing- work responsibly as members of work groups and work community- can assess their professional competence and develop it further."
Output	[0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Table 2: Example of course information input and the expected binary output for SDG prediction.

Training and evaluation of the models were carried out on Puhti, a Finnish research supercomputer provided by CSC - IT Center for Science, which facilitated the necessary computational

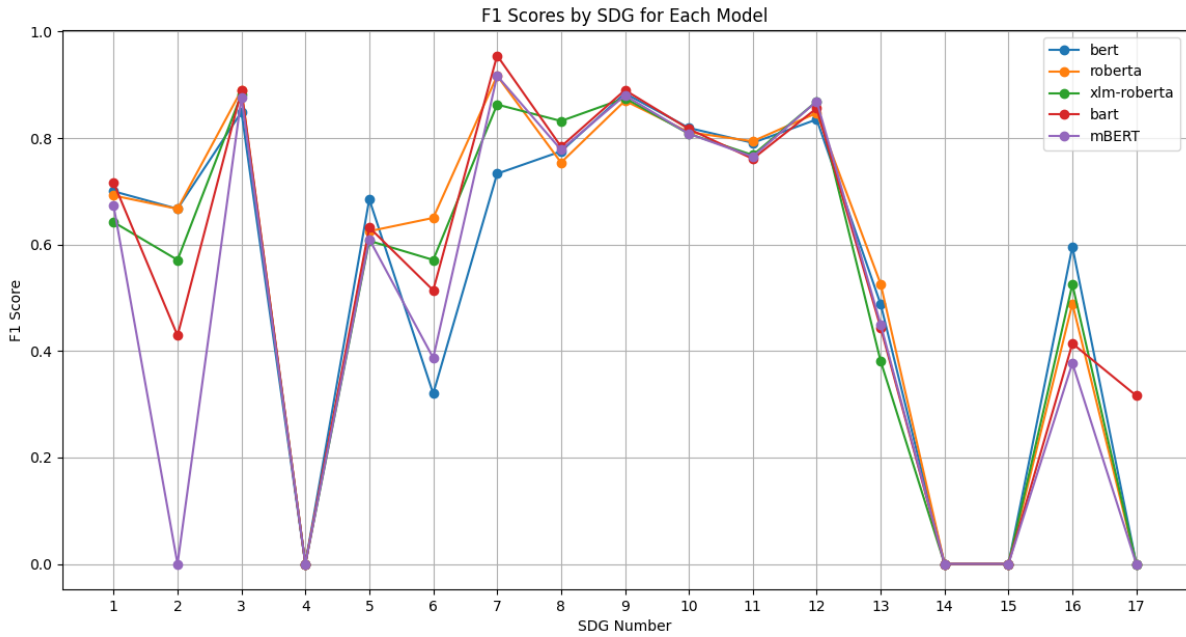


Figure 3: F1 Scores by SDG for Each Model

resources [CSC - IT Center for Science, 2023]. Using the V100 GPU’s large memory and parallel processing power, the models were trained on a single node to effectively handle our dataset.

## V RESULTS

Understanding the efficacy of a given algorithm in the context of multi-class SDG classification depends on the evaluation of model performance. The models’ performance was evaluated using precision, recall, and F1-score, which offer a thorough understanding of the models’ capabilities—especially in light of the inherent class imbalance in our dataset. The performance metrics for each model are shown in the following table, emphasizing their advantages and disadvantages for the given task.

Model	Precision	Recall	F1-Score
<b>BERT</b>	0.765	0.798	0.781
<b>mBERT</b>	0.762	0.795	0.778
<b>RoBERTa</b>	0.768	0.802	0.785
<b>XLM-RoBERTa</b>	0.767	0.801	0.784
<b>BART</b>	<b>0.769</b>	<b>0.803</b>	<b>0.786</b>

Table 3: Models performance based on the micro scores

The table 3 shows BERT’s precision of 0.765 and F1-score of 0.781 reflect its proficiency in categorizing instances correctly, demonstrating a reliable balance between precision and recall.

In contrast, BART outperforms other models with the highest F1-score, suggesting superior model efficacy due to its advanced pretraining methodology. The nuanced performance variations across the models underscore the significance of model selection tailored to specific NLP tasks’ requirements.

The displayed F1 scores reveal that model performance fluctuates across the SDGs, with BART

and mBERT often outperforming others, particularly in SDGs 7, 8, and 9. The lower F1 scores for SDGs 14,15 and 17 suggest that data imbalances pose challenges, affecting the models' ability to generalize effectively in these areas. Such patterns indicate the need for enhanced data strategies to address the imbalances and optimize model performance.

## VI CONCLUSIONS

In this paper, we introduced a novel approach to predicting UN SDGs for university courses, employing PaLM 2 large language model to generate training data from course descriptions. Through the utilization of various smaller language models, we successfully trained models to predict SDGs for university courses. Notably, the best-performing model in our experiments was BART, achieving an F1-score of 0.786.

This research contributes to advancing the integration of SDGs at the university level, providing a valuable methodology for enhancing the adaptation of sustainable development principles in higher education. The findings open avenues for further research and implementation of similar approaches to foster sustainable practices in academic institutions worldwide.

## References

- Amir Amel-Zadeh, Mike Chen, George Mussalli, and Michael Weinberg. Nlp for sdgs: Measuring corporate alignment with the sustainable development goals. *Columbia Business School Research Paper*, 2021.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- Leslie Mahe Collazo Expósito and Jesús Granados Sánchez. 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, 2020.
- Costanza Conforti, Stephanie Hirmer, Dai Morgan, Marco Basaldella, and Yau Ben Or. Natural language processing for achieving sustainable development: the case of neural labelling to enhance community profiling. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8427–8444, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.677. URL <https://aclanthology.org/2020.emnlp-main.677>.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.
- CSC - IT Center for Science. Puhti supercomputer. <https://www.csc.fi/en/-/puhti>, 2023. Accessed: 2023-12-16.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. Argumentation mining in scientific literature for sustainable development. In Khalid Al-Khatib, Yufang Hou, and Manfred Stede, editors, *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.argmining-1.10. URL <https://aclanthology.org/2021.argmining-1.10>.
- Marius Hessenthaler, Emma Strubell, Dirk Hovy, and Anne Lauscher. Bridging fairness and environmental sustainability in natural language processing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7817–7836, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.533. URL <https://aclanthology.org/2022.emnlp-main.533>.
- Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. <https://spacy.io/>, 2017. Accessed: 2023-12-16.

- Ching Ting Tany Kwee. 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, 2021.
- Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 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. *arXiv preprint arXiv:1907.11692*, 2019.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Abbas Rajabifard, Masoud Kahalimoghadam, Elisa Lumantarna, Nilupa Herath, Felix Kin Peng Hui, and Zahra Assarkhaniki. Applying sdgs as a systematic approach for incorporating sustainability in higher education. *International Journal of Sustainability in Higher Education*, 22(6):1266–1284, 2021.
- Samuel Sučík, Daniel Skala, Andrej Švec, Peter Hraška, and Marek Šuppa. Prompterator: Iterate efficiently towards more effective prompts. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 471–478, Singapore, December 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.emnlp-demo.43>.
- Goya van Boven, Stephanie Hirmer, and Costanza Conforti. 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 Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2007–2021, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.216>.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL <https://aclanthology.org/2020.emnlp-demos.6>.