

AI-driven Analysis of Higher Education Institutions' Contributions to the Sustainable Development Goals

Jordi Borrell¹, Davinia Hernández-Leo¹, Roger G. March¹, Miguel Habana¹, Patricia Santos¹, Carla Lancelotti², Josep M. Antó³

¹Department of Engineering, Universitat Pompeu Fabra, Spain, ²Department of Humanities, Universitat Pompeu Fabra, Spain. ³Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Spain.

How to cite: Borrell, J.; Hernández-Leo, D.; March, R. G.; Habana, M.; Santos, P.; Lancelotti, C.; Antó, J. M. (2026). AI-driven Analysis of Higher Education Institutions' Contributions to the Sustainable Development Goals. In: 12th International Conference on Higher Education Advances (HEAd'26). Valencia, 15-18 June 2026. <https://doi.org/10.4995/HEAd26.2026.21539>

Abstract

The integration and assessment of the Sustainable Development Goals (SDGs) in higher education are important yet resource-intensive tasks. This paper proposes a methodology for the automatic classification of academic output—including papers, theses, and syllabi—based on their thematic relevance to the 17 SDGs. Our research systematically compares the performance of traditional machine learning (Logistic Regression, Naive Bayes), deep learning (sentence embeddings with Neural Networks), and state-of-the-art large language models (LLMs) (Gemma, Llama 2, Mistral). Evaluation using the Cohen's Kappa coefficient reveals that deep learning approaches deliver superior performance, providing a classification accuracy closely aligning with expert human coders while maintaining computational efficiency. The application of the methodology to a large dataset from Universitat Pompeu Fabra (UPF) showcases the potential of the approach towards providing automated tools that aid academic administrators.

Keywords: Sustainable Development Goals; Machine Learning; Deep Learning; Large Language Models; Text Classification; Higher Education.

1. Introduction

Higher Education Institutions (HEIs) are central to the realization of the SDGs through the advancement of knowledge and the cultivation of competencies that equip citizens to actively contribute to the SDGs. The seventeen SDGs established by the United Nations General Assembly in 2015 offer a globally recognized framework for addressing sustainability, and enhancing the understanding of SDG implementation worldwide is a key objective of the UN 2030 Agenda. In response, HEIs across different geographical contexts have developed

strategies and initiatives aimed at supporting the SDGs (Sustainable Development Solutions Network, 2020; Antó, J. M. et al, 2021; Vanderfeesten et al., 2022). However, despite the growing number of institutional efforts, a persistent challenge remains: systematic assessment of the reach of SDGs HEIs initiatives. In particular, there is a lack of outcome-oriented assessment frameworks capable of measuring how HEI-led educational and research actions are directly related to SDG achievement, representing a significant gap in both research and practice (Lozano et al., 2015; Prior et al., 2025).

Recent studies have explored the use of automatic text classifiers and machine learning methods to analyze the alignment between textual content and the SDGs across different domains (Pukelis et al., 2020; Guisiano et al., 2022; Meier et al., 2025). These approaches show significant potential for supporting large-scale, systematic assessments of SDG-related activities. However, a major challenge persists in the availability and accessibility of suitable data for SDG analysis. Most existing SDG monitoring efforts rely on official governance statistics, predominantly produced by the United Nations Statistical Commission and national or regional statistical agencies. While these data sources are essential, they often fail to capture the full range of societal contributions—particularly those originating from HEI. The *Data Revolution* reports (IEAG, 2014; UN DESA, 2024) emphasize the urgent need for more comprehensive, innovative, and diversified data strategies to effectively monitor progress toward the SDGs, highlighting an opportunity for alternative data sources and analytical methods.

In this context, this study aims *to develop and evaluate an automatic text classification methodology to label and assess HEI activities according to their relevance to the SDGs*. A key reference for this work is the Open-Source Sustainable Development Goals (OSDG) initiative, which provides an open framework for classifying text data aligned with the SDGs. Pukelis et al. (2020) propose an approach that integrates heterogeneous textual sources into a unified SDG classification system using supervised machine learning techniques to identify salient features associated with each goal. Building on this foundation, our study adopts the OSDG conceptual framework to design and conduct a series of classification experiments. Our research systematically compares the performance of traditional machine learning, deep learning, and state-of-the-art large language models (LLMs). Although we do not directly employ the original OSDG models, our approach is strongly informed by their methodology. Moreover, model training is performed using the OSDG Community Dataset, which comprises over 40,000 SDG-labeled texts and represents one of the most comprehensive publicly available resources for SDG-oriented text classification. To evaluate the methodology, we use a large real dataset from *Universitat Pompeu Fabra (UPF)*, which in 2018 launched the *Planetary Wellbeing Initiative*, a long-term institutional strategy, animated by the United Nations Sustainable SDGs (Antó et al, 2021).

2. Materials and Methods

2.1. Datasets

The study employs three datasets: training datasets for model development, test datasets for performance evaluation, and UPF evaluation datasets for classification using the best-performing model.

The *training data* consists of subsets of the OSDG Community Dataset, which includes over 40,000 SDG-labeled texts annotated by 3-to-9 coders. Two training sets were created: one with high inter-coder agreement but strong class imbalance (agreement thresholds from 0.6 to 1.0), and another with lower agreement thresholds but greater class balance (thresholds from 0.25 to 1.0). This design enables evaluation of the trade-off between annotation quality and dataset balance in model performance.

Two *test datasets* were constructed to reflect different levels of SDG expertise. The first comprises 300 texts annotated by three coders with standard knowledge of SDG classification, while the second includes 100 texts annotated by an SDG expert and reviewed by a second expert to ensure high reliability. This design enables model evaluation under both standard and expert knowledge conditions.

The *evaluation datasets* capture UPF's diverse HEI outputs, including research publications, PhD theses, Master and undergraduate theses, and course syllabi (Table 1). To assess how automatic SDG classification supports analysis of UPF's contribution to sustainable development, abstracts or titles were used for computational efficiency. Standard preprocessing steps included removal of null values, translation into English, and elimination of punctuation, line breaks, and extra spaces.

Table 1. UPF's evaluation datasets

Name	Period	Number of Texts
UPF's Course Syllabi	2018-2022	2,969
UPF's Master and Degree Theses	2006-2023	3,330
UPF's PhD Theses	1991-2022	2,219
UPF's Research Publications	2014-2024	15,853

2.2. SDG Classification Models

Building on previous work in educational automatic text classification (Kestin et al., 2017; Amarasinghe et al., 2023), three types of models were explored: machine learning (ML) models based on statistical approaches, deep learning (DL) models based on artificial neural network approaches, and large language models (LLMs) based on massive artificial neural networks.

For the *ML models*, we used one approach based on a multiclass and multilabel Multinomial Naive Bayes (MNB), and a second one involved 16 binary Logistic Regression (LR) models, one for each SDG. For *DL models*, we used a well-known predefined model called FastText (Zolotov et al., 2017), and we also developed a customized artificial neural network (ANN) with different embedding models (GTE Large (Li et al., 2023), E5 Large (Wang et al., 2024), and Roberta Large (Reimers et al., 2019), trained with high-quality and balanced datasets.

The selected *LLM models* were Gemma 2B and 7B (Gemma Team et al., 2024), Llama 2 7B (Touvron et al., 2023), and Mistral 7B (Jiang et al., 2023). These models required no additional training; therefore, we used basic prompt engineering and text parsing to extract the SDG labels for analysis. Using all these models, we computed the Cohen's Kappa coefficient (CKC) with the test datasets to identify the model that most closely matched human criteria.

3. Results

3.1. Model results

CKC assessment indicates that MNB (EXP: 0.335, STD: 0.423) and FastText (EXP: 0.346, STD: 0.441) have very similar behavior, with both aligning closely with standard classification criteria (Table 2). However, FastText performs slightly better than MNB, suggesting that the deep learning model is superior. Regarding the LR (EXP: 0.467, STD: 0.412) and GTE Large embedding models (EXP: 0.465, STD: 0.499), we notice that the GTE Large Embeddings model achieves parity with LR on expert criteria and significantly outperforms LR on standard criteria. Similarly, while LR is much better than FastText for expert criteria, it performs worse than FastText for standard criteria. This suggests that deep learning models generally perform better than machine learning models. We also evaluated RoBERTa Large (EXP: 0.415, STD: 0.477) and E5 Large (EXP: 0.417, STD: 0.454) with inferior accuracy and efficiency than GTE Large, indicating that proper selection of embedding models is critical to ensure robust classification.

Table 2. Cohen's Kappa Coefficients for Expert (EXP) and Standard (STD) test datasets

Model	STD	EXP
MNB	0.423	0.335
LR	0.412	0.467
FastText	0.441	0.346
ANN - Roberta Large	<u>0.477</u>	0.415
ANN - E5 Large	0.454	0.417
ANN - GTE Large	0.499	<u>0.465</u>
Gemma 2B	-	0.175
Gemma 7B	-	0.322
Llama 2 7B	-	0.341
Mistral 7B	-	0.435

Regarding LLMs, the models with more parameters (7B) showed better results as expected, confirming that higher parameters lead to better classification performance (Table 2). The best LLM is Mistral 7B, with a Cohen's Kappa Coefficient of 0.435, outperforming MNB, FastText, and the Roberta and E5 Large embedding models. However, this result is still inferior to our best models, LR and GTE Large embeddings.

3.2. Exemplifying classification outcomes for the UPF dataset case

We classified all UPF activity datasets by SDG using the best-performing model (GTE Large Embedding) to estimate the *institution's contribution to sustainable development* based on the analyzed data (Figure 1). The most frequently automatically identified SDG labels (exceeding 10%) were SDG 3, SDG 9, and SDG 16. There are other secondary but not negligible SDGs with a percentage of around 5%, such as SDG 4, 5, 8, 10, and 11, indicating the importance of cross-cutting fields but also the diversity. It is interesting to notice that for all datasets, the number of documents without any SDG label is around 10% to 15%, indicating similar SDG engagement for all datasets. Results suggest that SDG 3 is predominantly associated with research articles and PhD theses; SDG 16 is mainly driven by undergraduate and Master theses; and SDG 9 is largely linked to course syllabi and undergraduate theses.

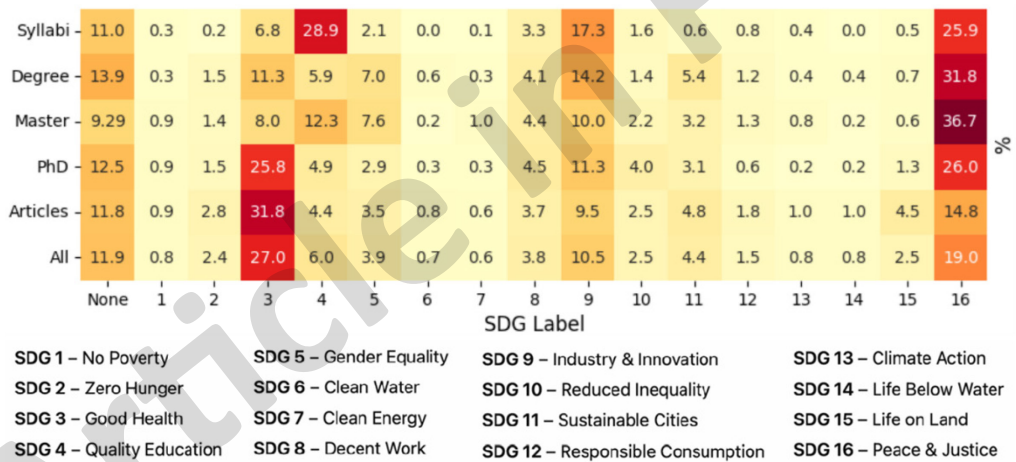


Figure 1. Percentage distribution of SDG labels for each UPF's dataset.

An additional informative indicator derived from the classification process is the *ratio of documents addressing SDGs* compared to the overall volume of documents within the UPF datasets (Figure 2). This metric enabled us to conduct a temporal assessment of both collective and dataset-specific engagement with SDG-related themes over time. For UPF, the aggregated dataset demonstrates a positive trend in SDG engagement. However, when examining course syllabi and master theses individually, their emphasis about SDGs doesn't seem to be increasing

across the years (2018-2022), whereas research-oriented outputs show a noticeable increase during this period (e.g. the percentage of SDG-related PhD theses increased around 5%).

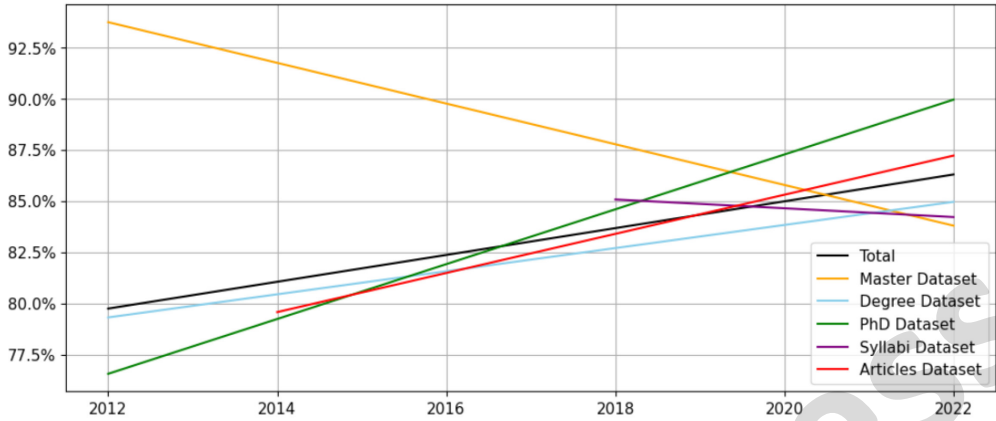


Figure 2. Regression lines of SDG-related over total documents in UPF's dataset (2012-2022).

4. Discussion and Conclusions

This study addresses a gap identified in the literature: the lack of scalable, outcome-oriented frameworks for assessing how both research and teaching efforts in HEIs contribute to the SDGs. This paper proposes a text-classification methodology that shows the potential of machine learning-based methods to complement SDG monitoring systems, by supporting exploratory, institution-level analyses that are otherwise difficult to capture. The methodology proposed shows that it is possible to develop models tailored to the specific institutional contexts. We evaluate and showcase the approach with a real large dataset of a HEI, where the achieved CKC value of 0.5 for a multiclass, multilabel task suggest that the approach has an acceptable validity for this type of institutional assessment.

The results of our study are useful to identify methodological challenges that need further attention. The finding that model performance and SDG distributions are influenced by training data quality, balance, and domain relevance, as well as by modeling choices and computational constraints suggest that these are issues that studies in this field should consider carefully. Analysis of the training corpus revealed annotation variability and class imbalance across SDGs, likely affecting sensitivity for underrepresented goals. In addition, the predominance of news-based texts introduced a domain mismatch with HEI-specific outputs such as theses, syllabi, and research publications, increasing classification uncertainty. These limitations reflect broader challenges identified in the literature regarding SDG data availability and highlight the need for more diverse and context-sensitive data sources. Consequently, both the potential and the limitations should be carefully considered when transferring it to other HEI contexts.

Despite these constraints, SDGs 3, 16, and 4 consistently emerged as the most represented across UPF activities. While these results may suggest potential model bias, they are also institutionally plausible: SDG 16 reflects the prominence of social sciences and humanities at UPF, SDG 3 aligns with its strong biomedical research profile, and SDG 4 is likely inflated in course syllabi due to the prevalence of educational terminology. The results also show interesting differences between HEI research and teaching activities across SDGs emphasis and time. Rather than definitive measurements, these findings should be interpreted as indicative signals of institutional engagement, reinforcing the usefulness of automated methods for identifying broad trends rather than precise SDG attribution.

Analyzing diverse HEI datasets against the SDGs can be significantly enhanced through temporal analysis. By evaluating the evolution of these metrics over time, institutions can identify emerging trends and refine policies to better align with sustainability targets. Moreover, disaggregating these datasets by department, degree program, or research topic enables a more nuanced diagnostic of institutional performance. This multidimensional monitoring framework—integrating activity type, time, and topic—facilitates a comprehensive map of institutional interactions, revealing critical opportunities for targeted strategic improvement.

Future work should focus on improving training data quality and domain specificity, leveraging advances in embedding models, increasing sensitivity in the classifiers, exploring approaches to enhance transparency and trust (through explainability), and extending the approach to other HEI cases. From a policy perspective, further research should examine how such automatic analysis tools can be integrated into institutional reporting and strategic planning processes, support longitudinal monitoring of SDG engagement, and enable coherent alignment between university activities and institutional strategies contributing to sustainable development.

Acknowledgments

This work has been partially supported by PID2023-146692OB-C33 and 2023-26 CEX2021-001195-M granted by MICIU/AEI/10.13039/501100011033 and the Center for Planetary Wellbeing of UPF. DHL (Serra Hünter) also acknowledges the support of ICREA Academia.

References

- Amarasinghe, I., Marques, F., Ortiz-Beltrán, A., et al. (2023). Generative Pre-trained Transformers for Coding Text Data? An Analysis with Classroom Orchestration Data. European Conference on Technology-Enhanced Learning, Springer LNCS 14200, 32-43.
- Antó, J. M., Martí, J. L., Casals, J., et al. (2021). The Planetary Wellbeing Initiative: Pursuing the Sustainable Development Goals in Higher Education. *Sustainability*, 13(6), 3372.
- Gemma Team, et al. (2024). GemMa: Open models based on Gemini research and technology. arXiv. <https://doi.org/10.48550/arxiv.2403.08295>

- Guisiano, J. E., Chiky, R., & De Mello, J. (2022). SDG-Meter: A deep learning based tool for automatic text classification of the Sustainable Development Goals. In N. T. Nguyen, T. K. Tran, et al. (Eds.), *Intelligent Information and Database Systems* (pp. 259–271). Springer.
- Jiang, A. Q., Sablayrolles, A., Arthur, M., Bamford, C., et al. (2023). Mistral 7B. arXiv <https://doi.org/10.48550/arxiv.2310.06825>
- Kestin, T., van den Belt, M., Denby, L., et al. (2017). Getting started with the SDGs in universities: A guide for universities, higher education institutions, and the academic sector. <https://apo.org.au/sites/default/files/resource-files/2017-08/apo-nid105606.pdf>
- Li, Z., Zhang, X., Zhang, Y., Long, D., et al. (2023). Towards General Text Embeddings with Multi-stage Contrastive Learning. arXiv <https://doi.org/10.48550/arxiv.2308.0328>
- Lozano, R., Ceulemans, K., et al. (2015). A review of commitment and implementation of sustainable development in higher education: results from a worldwide survey. *Journal of Cleaner Production*, 108, 1–18. <https://doi.org/10.1016/j.jclepro.2014.09.048>
- Meier, D. S., Mata, R., & Wulff, D. U. (2025). text2sdg: An R package to monitor sustainable development goals from text. *The R Journal*, 16(1), 83–95.
- Prior, D. D., Seshadrinath, S. M., et al. (2025). Measuring sustainable development goals (SDGs) in higher education through semantic matching. *Studies in Higher Education*, 50(7), 1556–1569. <https://doi.org/10.1080/03075079.2024.2386625>
- Pukelis, L., Bautista-Puig, N., Statulevičiūtė, G., Stančiauskas, et al. (2022). OSDG 2.0: a multilingual tool for classifying text data by UN Sustainable Development Goals (SDGs). arXiv. <https://doi.org/10.48550/arxiv.2211.11252>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In 9th Int. Joint Conference on Natural Language Processing, 3982–3992.
- Sustainable Development Solutions Network. (2020). Accelerating education for the SDGs in universities: A guide for universities, colleges, and tertiary and higher education institutions. UN report <https://ap-unsdsn.org/regional-initiatives/universities-sdgs/education-for-sdgs/>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv <https://doi.org/10.48550/arxiv.2307.09288>
- United Nations. Department of Economic and Social Affairs. (2024). The sustainable development goals report 2024. UN report, <https://unstats.un.org/sdgs/report/2024/The-Sustainable-Development-Goals-Report-2024>
- United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development. (2014). A world that counts: Mobilising a data revolution for sustainable development. United Nations. <https://www.undatarevolution.org/report/>
- Vanderfeesten, M., Jaworek, R., & Keßler, L. (2022). AI for mapping multi-lingual academic papers to the United Nations' Sustainable Development Goals (SDGs) (Version 1.0) [Report]. Zenodo. <https://doi.org/10.5281/zenodo.6487606>
- Wang, Q., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 Text Embeddings: a Technical report. arXiv <https://doi.org/10.48550/arxiv.2402.05672>
- Zolotov, V., & Kung, D. S. (2017). Analysis and optimization of FastText Linear Text Classifier. arXiv <https://doi.org/10.48550/arxiv.1702.05531>