Data Science and Engineering (DSE) Record, Volume 7, issue 1.

Development of research classification tools based on SDGs data collection using machine learning

Suphitcha Thawornlimpaphong [1] and Nasi Tantitharanukul [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
suphitcha_tha@cmu.ac.th

Abstract. This research develops and evaluates research classification tools for mapping research publications to the Sustainable Development Goals (SDGs) using SDG data collection and machine learning-based text classification. The proposed approach integrates Scopus textual metadata, SciVal SDG reference labels, an SDG key phrase repository, rule-based classification, and sdgBERT-based classification. Scopus was used as the source of textual metadata, including titles, abstracts, author keywords, and index keywords, while SciVal was used as the reference source for SDG labels. The Scopus dataset contained 55,702 records. After duplicate removal, 51,636 unique records remained, of which 31,660 records were successfully matched with SciVal, representing a match rate of 61.31%. The rule-based method classified 8,770 matched records, corresponding to 27.70% coverage. After excluding empty-empty cases, rule-based classification achieved 47.08% adjusted agreement and 38.22% exact match with SciVal. In contrast, sdgBERT at a threshold of 0.94 achieved 99.56% coverage, but its Micro F1, Macro F1, and adjusted agreement were 0.2061, 0.2411, and 11.71%, respectively. The results show a clear trade-off between the two approaches. Rule-based classification provides interpretable and more conservative outputs when explicit keyword evidence exists, but its coverage is limited. sdgBERT provides broader semantic coverage but shows low agreement with SciVal reference labels. Therefore, the developed tools should be used as preliminary decision-support mechanisms for SDG research classification rather than as fully automated final classification systems. Future work should develop and empirically test a hybrid approach that combines rule-based evidence, sdgBERT candidate labels, threshold-based configuration, and expert validation.

Keywords: SDG Classification, Research Classification, Machine Learning.

1 Introduction

The United Nations adopted the Sustainable Development Goals (SDGs) as a global framework to address social, economic and environmental challenges. The framework comprises 17 goals and 169 targets, ranging from poverty and health to education, climate action, responsible consumption and partnerships for sustainable development. Universities and research institutions have an important role to play in the promotion of these goals, as they generate research, knowledge and innovation that can help achieve sustainable development.

Fig. 1 Sustainable Development Goals (SDGs)[1]

In recent years, universities have increasingly used the SDGs as a framework for assessing and commu[1]nicating the societal contribution of research. Research publications are often analysed to determine what SDGs they support, particularly for institutional reporting, research strategy development and international ranking systems. However, the classification of research publications according to SDGs is not straightforward. One publication can relate to more than one SDG. For example, research on renewable energy might be relevant for SDG 7, SDG 9 and SDG 13 at the same time. Therefore, the classification of SDG research should be considered a multi-label classification problem.

The traditional SDG classification is often based on keyword matching. This approach is transparent and easy to interpret since each assigned SDG label can be traced back to specific words or phrases. However, the keyword-based methods have limitations when the SDG relevance is expressed implicitly or in the discipline-specific terms that are not included in a predefined list of keywords. This may result in relevant publications not being classified.

Machine learning and natural language processing offer alternative ways to classify SDGs. Transformer-based models (e.g. BERT) and domain-specific models (e.g. sdgBERT) can capture contextual and semantic information beyond direct keyword matching. These models can be potentially useful for the identification of SDG relevance in large-scale research publication datasets. However, the use of such models for institutional research management should be carefully evaluated. High coverage of high classification does not mean high agreement with reference labels. Important issues are threshold choice, label imbalance, interpretability and consistency with reference data of the institution.

[1] https://www.un.org/sustainabledevelopment/news/communications-material/

In this study, research classification tools are developed and evaluated with machine learning. The study combines textual metadata from Scopus, SciVal SDG reference labels, SDG key phrases, rule-based classification and sdgBERT-based classification. The aim is not to replace expert judgment but to explore how rule-based and sdgBERT-based approaches can assist in the initial classification of SDG research. The objectives of this study are the following:

1. To develop a research classification tool for mapping research publications according to the Sustainable Development Goals (SDGs) by integrating Scopus textual metadata, SciVal SDG reference labels, SDG key phrases, and sdgBERT-based classification.

2. To evaluate and compare the roles, coverage, agreement, exact match, classification performance, and limitations of rule-based classification and sdgBERT-based classification using SciVal SDG labels as reference data.

2 Literature Review

2.1 SDG Research Mapping

The Sustainable Development Goals (SDGs) are a global set of sustainable development objectives that comprise 17 goals targeting social, economic and environmental dimensions (United Nations, 2020). In higher education and research institutions, SDGs have been increasingly used to analyze the contribution of research outputs towards sustainability-related goals. However, mapping research publications to SDGs is not straightforward as a publication can be associated with more than one goal. This makes the SDG research classification consistent with the multi-label classification, where an instance can be related to more than one class simultaneously (Tsoumakas & Katakis, 2007; Zhang & Zhou, 2014).

Several works have used bibliometric metadata and text mining for supporting SDG research mapping. Wang et al. (2023) proposed Auckland Approach, which maps research publications to SDGs using titles, abstracts and keywords. This approach uses n-gram analysis and Term Frequency-Inverse Document Frequency (TF-IDF) to identify SDG-related terms, then expert review to improve contextual relevance. This demonstrates the importance of both computational text analysis and domain validation for SDG mapping.

2.2 Rule-Based and Text Mining Approaches

Text mining techniques such as frequency analysis and TF-IDF are often used to construct or refine keyword repositories. Jung et al. (2022) applied text mining to examine the relationship between ISO standards and SDGs, showing that keyword overlap can reveal relationships between technical standards and sustainability themes. However, rule-based classification depends heavily on the completeness of the keyword repository. If a publication is relevant to an SDG but does not contain expected terms, the method may fail to classify it. Therefore, rule-based classification is useful as an interpretable evidence layer, but it may have limited coverage when applied to large and diverse research datasets.

2.3 Transformer-Based and Multi-Label SDG Classification

Machine learning and deep learning have been increasingly applied to SDG-related text classification. Lee et al. (2023) used a machine learning approach to classify development aid activities according to SDGs, while Angin et al. (2022) applied RoBERTa to classify sustainability reports. Pukelis et al. (2022) also introduced OSDG 2.0 as a multilingual tool for classifying text data according to the United Nations SDGs. These studies indicate that AI-based methods can support large-scale SDG text analysis.

Transformer models are relevant to SDG classification because they can capture contextual relationships in text. The Transformer architecture introduced by Vaswani et al. (2017) uses self-attention mechanisms to model relationships among tokens. BERT further improved contextual representation through bidirectional pretraining (Devlin et al., 2019), while RoBERTa improved BERT's training strategy and achieved stronger performance across NLP tasks (Liu et al., 2019). For domain-specific tasks, domain-adaptive pretraining can improve model performance when the target corpus differs from general pretraining data (Gururangan et al., 2020). In the sustainability domain, sdgBERT was developed to capture sustainability-related discourse and SDG-related semantic patterns (Sadick et al., 2026).

Since SDG classification is a multi-label problem, evaluation should not rely on a single metric. Micro and Macro Precision, Recall, and F1-score provide different perspectives on classification performance, while Hamming Loss measures incorrect label assignments across label dimensions (Tsoumakas & Katakis, 2007; Zhang & Zhou, 2014). In addition, coverage is important because it indicates the proportion of records that receive at least one predicted SDG label. For models such as sdgBERT, threshold selection is also important because a lower threshold may increase coverage and recall, while a higher threshold may improve precision and reduce label-level errors.

2.4 Research Gap

Prior works show the evolution of SDG classification from keyword-based approaches to Transformer-based semantic classification. Keyword-based methods also offer transparency and traceability, while Transformer-based models provide broader semantic capability (Wang et al., 2023; Angin et al., 2022; Pukelis et al., 2022). However, three research gaps remain. First, threshold sensitivity in the multi-label classification of SDG is still underexplored. Second, few studies evaluate SDG classification using institutional research datasets that combine Scopus textual metadata with SciVal reference labels. Third, hybrid approaches that combine rule-based evidence, candidate labels based on Transformer, and validation by experts have not been sufficiently tested.

This study addresses these gaps by creating and assessing research classification tools using machine learning based on SDG data collection. It combines Scopus textual metadata, SciVal SDG reference labels, SDG key phrases, rule-based classification and sdgBERT-based classification. To clarify the practical roles and limitations of each approach, we compare the methods in terms of coverage, agreement, exact match, multi-label metrics, and threshold sensitivity analysis.

3 Data and Methodology

3.1 Research Design and Workflow

This study adopted a quantitative research design to develop and evaluate research classification tools for mapping research publications to the Sustainable Development Goals (SDGs). The methodological framework was designed as an SDG-based research classification workflow, as shown in Fig. 2. Rather than treating the workflow as a linear data-processing procedure, the framework explains how different data sources and classification approaches were integrated to support SDG label assignment and evaluation.

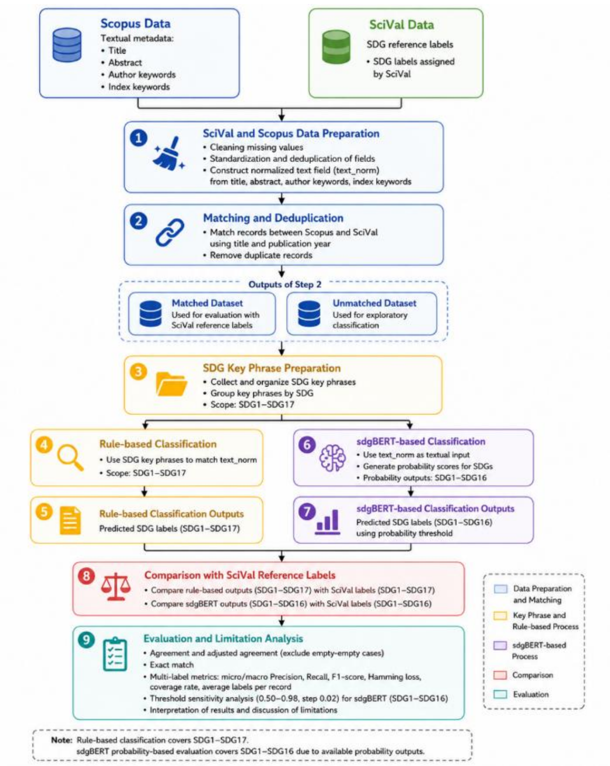

Fig. 2 Methodological Workflow for SDG-Based Research Classification

The workflow consists of five main components. First, research publication metadata were obtained from Scopus, while SDG reference labels were obtained from SciVal. Second, both datasets were cleaned, standardized, normalized, and matched using publication title and publication year. Third, the matched and unmatched datasets were separated according to whether SciVal reference labels were available. Fourth, two classification approaches were independently applied: rule-based classification using SDG key phrases and sdgBERT-based classification using machine learning-based semantic prediction. Finally, the classification outputs were compared with SciVal reference labels and evaluated using agreement-based and multi-label classification metrics.

The matched dataset was used for quantitative evaluation because it contained both Scopus textual metadata and SciVal SDG reference labels. The unmatched dataset was retained for exploratory classification because it contained Scopus textual metadata but did not have matched SciVal reference labels. Therefore, unmatched records were not interpreted as non-SDG publications.

3.2 Data Sources and Study Scope

The data source of this study consisted of research publication records from Thai higher education institutions classified by the Ministry of Higher Education, Science, Research and Innovation (MHESI) as Group 2 institutions, which are institutions oriented toward technology development and innovation promotion. The study used bibliographic and SDG-related data from two main sources: Scopus and SciVal.

Scopus was used as the source of textual metadata for research publications. The relevant fields included publication title, abstract, author keywords, index keywords, and publication year. These fields were selected because they provide the textual evidence required for SDG-related text classification.

SciVal was used as the reference source for SDG labels. The SDG labels assigned by SciVal were treated as reference labels for evaluation, not as absolute ground truth. This distinction is important because SDG relevance may differ depending on the classification criteria, interpretation of research context, and availability of textual evidence.

In addition to Scopus and SciVal, an SDG key phrase repository was used as a domain-specific knowledge source for rule-based classification. The repository contained SDG-related key phrases for SDG 1 to SDG 17 and was used to identify explicit textual evidence associated with each SDG.

3.3 Data Preparation and Record Matching

Data preparation was conducted to create a consistent textual representation of each publication. Missing values were cleaned, relevant fields were standardized, and duplicate records were removed. The textual input used for classification was constructed by combining the title, abstract, author keywords, and index keywords into a normalized text field.

The textual representation of each publication can be expressed as:

$$\text{Text} = \text{Title} + \text{Abstract} + \text{Author Keywords} + \text{Index Keywords} \qquad (1)$$

The original Scopus dataset contained 55,702 records. After duplicate removal, 51,636 unique records remained. The SciVal dataset contained 34,349 records after cleaning, of which 10,268 records contained extracted SDG labels.

Record matching between Scopus and SciVal was conducted using publication title and publication year. This matching process produced 31,660 matched records, corresponding to 61.31% of the unique Scopus records. The remaining 19,976 records were unmatched. The matched dataset was used for evaluation, while the unmatched dataset was retained for exploratory classification.

3.4 Classification Framework

The classification framework consisted of two independent approaches: rule-based classification and sdgBERT-based classification.

The rule-based classification approach used phrase-level matching between normalized publication text and the SDG key phrase repository. Each matched key phrase contributed evidence to its corresponding SDG label. This approach covered SDG 1 to SDG 17. Its main advantage was interpretability because each predicted label could be traced back to explicit matched phrases. However, the method was limited by the completeness of the key phrase repository and could not detect SDG relevance expressed implicitly or through non-matching terminology.

The sdgBERT-based classification approach used a Transformer-based language model to generate SDG probability scores from the normalized publication text. The model input was tokenized with a maximum sequence length of 512 tokens. The model produced probability scores for SDG labels, and predicted labels were selected using a probability threshold. In this study, the main threshold used for classification was 0.94. Threshold sensitivity analysis was also conducted from 0.50 to 0.98 with increments of 0.02.

In the available sdgBERT probability output, probability columns were available for SDG 1 to SDG 16. Therefore, sdgBERT probability-based evaluation covered SDG 1 to SDG 16, while rule-based classification covered SDG 1 to SDG 17.

3.5 Comparison and Evaluation Strategy

The outputs from rule-based classification and sdgBERT-based classification were compared with SciVal reference labels. The purpose of the evaluation was not to identify a single universally superior method, but to examine the roles, strengths, limitations, and trade-offs of the two approaches.

Agreement was calculated using Jaccard similarity between the predicted SDG label set and the SciVal reference label set:

$$\text{Agreement} = |\text{Predicted Labels} \cap \text{Reference Labels}| \,/\, |\text{Predicted Labels} \cup \text{Reference Labels}| \quad (3)$$

Exact match was calculated when the predicted label set was identical to the SciVal reference label set. An adjusted evaluation was also performed by excluding empty-empty cases, where both the predicted output and SciVal reference labels contained no SDG labels. This adjustment was necessary because counting empty-empty cases as full agreement could inflate agreement scores without reflecting actual SDG label alignment.

Multi-label classification metrics were also used, including Micro Precision, Micro Recall, Micro F1-score, Macro Precision, Macro Recall, Macro F1-score, Hamming Loss, coverage rate, and average labels per record. Coverage was calculated as the proportion of records that received at least one predicted SDG label. These metrics were used together to interpret classification performance because SDG research classification is a multi-label problem, and no single metric can fully explain the effectiveness of the classification tools.

4 Result

4.1 Data Preparation Results

SciVal was used as the reference source for SDG labels. After cleaning, the SciVal dataset contained 34,349 records, of which 10,268 records contained extracted SDG labels.

Scopus was used as the source of textual metadata. The original Scopus dataset contained 55,702 records. After duplicate removal, 51,636 unique records remained.

Table 1. Results of Data Preparation

Model	**Result**
Original Scopus records	55,702
Unique Scopus records after deduplication	51,636
SciVal records after cleaning	34,349
SciVal records with extracted SDG labels	10,268

4.2 Record Matching Results

Record matching between Scopus and SciVal was performed using publication title and publication year. The results showed that 31,660 records were successfully matched, representing 61.31% of the unique Scopus records. The remaining 19,976 records were unmatched, representing 38.69%.

Table 2. Results of Record Matching and Deduplication

Item	Result	Percentage
Unique Scopus records	51,636	100.00%
Matched records	31,660	61.31%
Unmatched records	19,976	38.69%

The matched dataset was used for evaluation against SciVal reference labels. The unmatched dataset was retained as an exploratory classification pool.

4.3 Rule-Based Classification Results

The rule-based method classified 8,770 records out of 31,660 matched records, giving a coverage rate of 27.70%. For the unmatched dataset, it classified 4,713 records out of 19,976 records, giving a coverage rate of 23.59%.

Table 3. Results of Rule-Based Classification

Dataset	Total Records	Classified Records	Coverage
Matched records	31,660	8,770	27.70%
Unmatched records	19,976	4,713	23.59%

These results show that the rule-based method provides transparent and traceable outputs. However, its coverage is limited because it depends on explicit matches with predefined SDG key phrases.

4.4 sdgBERT-Based Classification Results

At a threshold of 0.94, sdgBERT achieved 99.56% coverage on the matched dataset, with an average of 1.6157 labels per record. This indicates that the model assigned at least one SDG label to almost all records. However, classification performance against SciVal reference labels remained low.

Table 4. sdgBERT-Based Classification Results at Threshold 0.94

Metric	Value
Label scope	SDG 1–16
Coverage rate	99.56%
Average labels per record	1.6157
Micro Precision	0.1293
Micro Recall	0.5082
Micro F1-score	0.2061
Macro Precision	0.2331
Macro Recall	0.4228
Macro F1-score	0.2411
Hamming Loss	0.1006

These results indicate that sdgBERT provides broad coverage, but many predicted labels do not align with SciVal reference labels. Therefore, high coverage should not be interpreted as high agreement-based accuracy.

4.5 Agreement and Exact Match Analysis

Agreement and exact match were calculated for both rule-based and sdgBERT-based classification. The evaluation was performed before and after excluding empty-empty cases.

Table 5. Agreement and Exact Match Before and After Excluding Empty-Empty Cases

Comparison	Label Scope	Evaluation Scope	Agreement	Exact Match	Empty-Empty Cases	Empty-Empty Rate
SciVal vs Rule-Based	SDG 1– 17	All records	80.72%	77.50%	20,129	63.58%
SciVal vs Rule-Based	SDG 1– 17	Excluding empty-empty cases	47.08%	38.22%	20,129	63.58%
SciVal vs sdgBERT	SDG 1– 16	All records	12.08%	5.39%	133	0.42%
SciVal vs sdgBERT	SDG 1– 16	Excluding empty-empty cases	11.71%	4.99%	133	0.42%

The initial rule-based agreement was 80.72%, with exact match of 77.50%. However, 20,129 records, or 63.58% of the matched dataset, were empty-empty cases. After excluding these cases, the adjusted rule-based agreement decreased to 47.08%, and exact match decreased to 38.22%.

This finding shows that the initial rule-based agreement should not be interpreted as strong classification performance. It was inflated by cases where neither SciVal nor the rule-based method assigned SDG labels.

In contrast, sdgBERT had only 133 empty-empty cases, representing 0.42% of the dataset. After excluding empty-empty cases, its agreement changed only slightly from 12.08% to 11.71%. This indicates that low sdgBERT agreement was mainly caused by differences between sdgBERT predictions and SciVal reference labels, rather than by empty-empty cases.

4.6 Multi-Label Metric Evaluation

The classification outputs were evaluated using standard multi-label metrics. The results show a clear trade-off between the two methods.

Table 6. Multi-Label Metrics of Rule-Based and sdgBERT-Based Classification

Method	Label Scope	Micro Precision	Micro Recall	Micro F1	Macro Precision	Macro Recall	Macro F1	Hamming Loss	Coverage
Rule-Based	SDG 1– 17	0.6721	0.6279	0.6493	0.6234	0.6940	0.6171	0.0164	27.70%
sdgBERT	SDG 1– 16	0.1293	0.5082	0.2061	0.2331	0.4228	0.2411	0.1006	99.56%

The rule-based method achieved higher Micro Precision, Micro F1, Macro Precision, Macro Recall, and Macro F1 than sdgBERT. It also achieved lower Hamming Loss. This indicates that when rule-based classification assigned labels, its predictions were more consistent with SciVal reference labels.

However, the rule-based method covered only 27.70% of the matched dataset. Therefore, its stronger metric values should be interpreted together with its limited coverage. The method is conservative and interpretable, but it leaves many records unclassified.

sdgBERT achieved much higher coverage at 99.56%, meaning that it assigned at least one SDG label to almost all records. However, its lower precision and F1-scores indicate that many predicted labels did not match SciVal reference labels. Therefore, sdgBERT is more suitable for preliminary candidate label generation than for final automatic classification.

4.7 Threshold Sensitivity Analysis

Threshold sensitivity analysis was conducted using sdgBERT probability outputs for SDG 1 to SDG 16. Threshold values from 0.50 to 0.98 were tested. The best threshold based on Micro F1 was 0.98.

Threshold 0.98 achieved a slightly higher Micro F1-score and lower Hamming Loss than threshold 0.94. However, it also reduced coverage and recall.

Table 7. Comparison Between Threshold 0.94 and 0.98

Threshold	Coverage	Avg. Labels	Micro Precision	Micro Recall	Micro F1	Macro F1	Hamming Loss
0.94	99.56%	1.6157	0.1293	0.5082	0.2061	0.2411	0.1006
0.98	94.36%	1.1174	0.1477	0.4017	0.2160	0.2273	0.0749

These results show that threshold selection is not merely a technical parameter, but an operational decision. A lower threshold such as 0.94 is more suitable for broad preliminary screening because it maximizes coverage and reduces the risk of missing potentially relevant SDG records. A higher threshold such as 0.98 is more suitable for conservative classification because it reduces the number of assigned labels and lowers Hamming Loss.

5 Discussion

The results demonstrate that rule-based and sdgBERT-based classification have different strengths and limitations.

The rule-based approach is highly interpretable. Its predictions can be traced to matched SDG key phrases, making it useful for expert review and institutional reporting. It also achieved stronger agreement-based performance than sdgBERT when it was able to assign labels. However, its coverage was limited. This limitation is expected because the method depends on explicit lexical matches. If a publication is contextually related to an SDG but does not contain relevant key phrases, the rule-based method cannot detect it.

sdgBERT showed the opposite behavior. It achieved very high coverage and assigned labels to almost all records. This indicates that Transformer-based semantic classification can support broad screening across large datasets. However, its low agreement with SciVal labels indicates that model predictions were not sufficiently aligned with the reference labels. This may be caused by several factors, including differences between the model's training distribution and the institutional dataset, label imbalance, different interpretations of SDG relevance, and threshold configuration. In addition, SciVal labels should be treated as reference labels rather than absolute ground truth for all forms of SDG relevance.

The analysis of empty-empty cases is an important methodological finding. The initial rule-based agreement of 80.72% appeared high, but it was strongly influenced by cases in which both SciVal and rule-based outputs contained no SDG labels. After excluding these cases, the adjusted agreement decreased to 47.08%. This demonstrates that agreement metrics can be misleading if unlabeled-unlabeled cases are counted as full agreement. For SDG classification, adjusted agreement provides a more meaningful interpretation of actual label alignment.

The results also clarify the meaning of performance in SDG classification. If performance is defined as agreement with SciVal and label-level reliability, rule-based classification performs better. If performance is defined as coverage and semantic candidate generation, sdgBERT performs better. Therefore, neither method is universally superior. Their value depends on the intended use case.

For institutional SDG reporting, a practical classification tool should not rely entirely on either method alone. Rule-based classification can serve as an interpretable evidence layer, while sdgBERT can serve as a semantic candidate generation layer. Records where both methods agree may be treated as higher-confidence cases. Records where they disagree should be prioritized for expert review.

6 Conclusion

This study developed and evaluated research classification tools based on SDG data collection using machine learning. The tools integrated Scopus textual metadata, SciVal SDG reference labels, an SDG key phrase repository, rule-based classification, and sdgBERT-based classification.

The workflow successfully processed 55,702 Scopus records, removed duplicates to obtain 51,636 unique records, and matched 31,660 records with SciVal reference data. The rule-based classification method achieved 27.70% coverage on the matched dataset and 47.08% adjusted agreement after excluding empty-empty cases. The sdgBERT-based method achieved 99.56% coverage at threshold 0.94, but its Micro F1, Macro F1, and adjusted agreement were low.

The findings show a trade-off between interpretability and coverage. Rule-based classification is more conservative, interpretable, and aligned with SciVal when explicit keyword evidence is available, but its coverage is limited. sdgBERT provides broad semantic screening and candidate label generation, but its agreement with SciVal reference labels is insufficient for fully automated final classification.

Therefore, the developed tools should be used as preliminary decision-support mechanisms rather than complete automated SDG classification systems. They can support institutional research data analysis by generating candidate SDG labels and identifying records for further expert validation.

7 Limitations and Future Work

This study has several limitations. First, SciVal reference labels were not available for all records. Therefore, direct quantitative evaluation was limited to matched records. Second, the rule-based method depended on a predefined SDG key phrase repository, which limited its ability to detect implicit or context-dependent SDG relevance. Third, sdgBERT achieved high coverage but low agreement with SciVal, indicating the need for further calibration and fine-tuning. Fourth, the present study did not empirically test a fully integrated hybrid model. Fifth, AI probability-based evaluation was limited to

SDG 1 to SDG 16 because SDG 17 probability outputs were not available in the model output file.

Future work should focus on four directions. First, the SDG key phrase repository should be expanded using institution-specific and discipline-specific research terminology. Second, sdgBERT should be fine-tuned using SciVal-labeled institutional data to improve alignment with reference labels. Third, threshold selection should be developed as an objective-based configuration mechanism, allowing users to choose thresholds based on the intended use case, such as broad discovery, formal reporting, or expert review prioritization. Fourth, a true hybrid SDG classification framework should be developed and empirically evaluated. Such a framework should combine rule-based evidence, sdgBERT probability scores, confidence-level classification, and expert validation.

References

1. Angin, M., Taşdemir, B., Yılmaz, C. A., Demiralp, G., Atay, M., Angın, P., & Dikmener, G. (2022). A RoBERTa approach for automated processing of sustainability reports. Sustainability, 14(23), 16139. https://doi.org/10.3390/su142316139

2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (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 (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

3. Jung, D., Kim, J., & Koo, S. H. (2022). A study on the activation of ISO standards to achieve SDGs: Using text mining technique. Journal of Standards & Certification Safety, 12(3), 93–104. https://doi.org/10.34139/jscs.2022.12.3.93

4. Lee, J., Park, S., & Kim, H. (2023). Machine learning-based classification of development aid activities aligned with the Sustainable Development Goals. Sustainability, 15(3), 2458. https://doi.org/10.3390/su15032458

5. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

6. Pukelis, L., Bautista-Puig, N., Statulevičiūtė, G., Stančiauskas, V., Dikmener, G., & Akylbekova, D. (2022). OSDG 2.0: A multilingual tool for classifying text data by UN Sustainable Development Goals (SDGs). arXiv preprint arXiv:2211.11252. https://doi.org/10.48550/arXiv.2211.11252

7. Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 3(3), 1–13.

https://doi.org/10.4018/jdwm.2007070101

8. United Nations. (2020). The Sustainable Development Goals report 2020. United Nations. https://unstats.un.org/sdgs/report/2020/

9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762

10. Wang, W., Kang, W., & Mu, J. (2023). Mapping research to the Sustainable Development Goals (SDGs). Research Square. https://doi.org/10.21203/rs.3.rs-2544385/v2

11. Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8), 1819–1837. https://doi.org/10.1109/TKDE.2013.39

12. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2004.10964

13. Sadick, A.-M., Hasan, A., & Ahiaga-Dagbui, D. D. (2026). Modeling sustainability discourse in the construction industry: A deep-learning approach. Journal of Construction Engineering and Management, 152(4). https://doi.org/10.1061/JCEMD4.COENG-16205