Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI

(Rekayasa Sistem dan Teknologi Informasi)

Vol. 9 No. 3 (2025) 688 - 696 e-ISSN: 2580-0760

Automatic Classification of Multilanguage Scientific Papers to the Sustainable Development Goals Using Transfer Learning

Lya Hulliyyatus Suadaa[1]*, Anugerah Karta Monika[2], Berliana Sugiarti Putri[3], Yeni Rimawati[4]

[1,2]Politeknik Statistika STIS, Jakarta, Indonesia

[3,4]BPS Statistics Indonesia, Jakarta, Indonesia

[1]lya@stis.ac.id, [2]ak.monika@stis.ac.id, [3]222011595@stis.ac.id, [4]yeni.rima@bps.go.id

Abstract

The classification of scientific papers according to their relevance to Sustainable Development Goals (SDGs) is a critical task in identifying the research development status of goals. However, with the growing volume of scientific literature published worldwide in multiple languages, manual categorization of these papers has become increasingly complex and time-consuming. Furthermore, the need for a comprehensive multilingual dataset to train effective models complicates the task, as obtaining such datasets for various languages is resource intensive. This study proposes a solution to this problem by leveraging transfer learning techniques to automatically classify scientific papers into SDG labels. By fine-tuning pretrained multilingual models mBERT on SDG publication datasets in a multilabel approach, we demonstrate that transfer learning can significantly improve classification performance, even with limited labelled data, compared to SVM. Our approach enables the effective processing of scientific papers in different languages and facilitates the seamless mapping of research to the relevance of SDGs, the four pillars of SDGs, and the 17 goals of SDGs. The proposed method addresses the scalability issue in SDG classification and lays the groundwork for more efficient systems that can handle the multilingual nature of modern scientific publications.

Keywords: multilingual model; multilabel text classification; scientific papers; SDGs research

How to Cite: L. H. Suadaa, A. K. Monika, B. S. Putri, and Y. Rimawati, "Automatic Classification of Multilanguage Scientific Papers to the Sustainable Development Goals Using Transfer Learning", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 3, pp. 688 - 696, Jun. 2025.
*Permalink/*DOI: https://doi.org/10.29207/resti.v9i3.6560

Received: April 19, 2025
Accepted: June 13, 2025
Available Online: June 23, 2025

1. Introduction

The United Nations' Sustainable Development Goals (SDGs) serve as a comprehensive and universal framework designed to address the world's most pressing social, economic, and environmental challenges by the year 2030. Encompassing 17 interconnected goals—from ending poverty and hunger to ensuring quality education, gender equality, climate action, and strong institutions—the SDGs offer a shared vision and actionable roadmap for countries, organizations, and individuals to work toward a more inclusive, equitable, and sustainable future [1].

In this context, scientific research plays a vital role in supporting and advancing the SDGs by generating evidence, informing policy decisions, and fostering innovation. However, it has not effectively actualized the research findings into practical actions [2]. Therefore, systematically monitoring how scientific publications align with specific SDGs is crucial [3]. It allows policymakers, funding bodies, and researchers to assess current progress, uncover underexplored areas, and allocate resources more effectively. By identifying which goals are receiving substantial research attention and which are being overlooked, stakeholders can take informed steps to address imbalances and accelerate global progress toward sustainable development.

Scientific papers are a fundamental source of reliable evidence that underpin SDG-related research, offering insights into technological advancements, policy impacts, social dynamics, and environmental trends. These publications not only reflect the current state of knowledge across various disciplines but also contribute to shaping the direction of future initiatives aligned with the Sustainable Development Goals (SDGs). As such, analyzing and categorizing scientific literature based on its relevance to specific SDGs is essential for understanding the global research landscape and ensuring that scientific efforts are aligned with sustainable development priorities. Several studies conducted bibliometric analysis to inform and direct future initiatives aimed at advancing sustainable development [4]-[6].

However, the rapidly growing volume of scientific literature—spanning diverse fields, institutions, and regions—presents a major challenge for effective monitoring and analysis. This challenge is further compounded by the linguistic diversity of publications, as research is increasingly being published in multiple languages beyond English. As a result, the manual classification of papers into SDG categories becomes not only labour-intensive and time-consuming but also prone to inconsistencies and scalability issues. These limitations highlight the urgent need for automated and multilingual approaches that can accurately and efficiently process and classify scientific documents in support of the SDG agenda. Natural language processing (NLP) techniques have been applied to automatically process texts from article collections, including topic modelling and classification models.

Leveraging topic modelling, F. Invernici et al. [7] uncover how perspectives on the SDGs have evolved in scientific abstracts between 2006 and 2023. Specific to certain SDGs, F. Illia, et al. [8] support the analysis of research topic trends related to SDGs Goal 6 and highlight a dominant focus on Target 6.3, which pertains to water quality.

Text classification models were also developed to automatically classify articles to SDG labels in supervised mechanisms. A. Hajikhani and A. Suominen [9] developed a multiclass classification model to classify patent documents to the 17 goals of SDGs so that each document maps to a particular goal. Since the articles could be relevant to several goals, R. C. Morales-Hernández et al. [10] categorize the articles to the 17 goals of SDGs using a multilabel classification approach. Therefore, this study proposes leveraging multilabel classification models.

Research is published in a wide range of languages across the globe, reflecting the diverse linguistic and cultural contexts in which scientific knowledge is produced. While this multilingual nature enriches the global research landscape, it also poses significant challenges for automated systems that aim to classify and analyze scientific content. Most traditional classification models are designed and trained primarily on English-language datasets, limiting their effectiveness when applied to publications in other languages. X. Luo [11] and N. Disayiram and R. A. H. M. Rupasingha [12] build machine learning models, such as Random Forest, Naïve Bayes, and Support Vector Machines (SVM), to categorize English texts. Leveraging the deep learning model through the transfer learning procedures, S. Aum and S. Choe [13] fine-tune Bidirectional Encoder Representation from Transformers (BERT) to an automatic English article classification model for systematic review. BERT is built to pre-train deep text representations by considering context from both directions, and it can be quickly adapted for classification tasks by adding a single classification layer, eliminating the need to develop a model from scratch [14].

To build robust and inclusive classification systems, it is necessary to develop models capable of understanding and processing multilingual content. The strong performance of the BERT architecture in capturing deeper contextual meaning from input texts has motivated the creation of BERT models trained on various corpora, including multilingual corpora. Devlin et al. introduced multilingual BERT (mBERT) [14], a pre-trained model developed using Wikipedia texts in 104 languages, which demonstrated strong performance in zero-shot cross-lingual transfer tasks. Various studies have utilized mBERT for processing Indonesian tasks, including the classification of toxic comments [15], hoax news [16], student feedback [17], and aspect-based sentiment analysis of tourism reviews [18], [19]. The multilingual classification models for SDGs articles have been proposed by L. Pukelis et al. [20], [21], supporting 15 languages, but not including Indonesian.

Most Indonesian research papers are published in Indonesian and English. Previous research by B. S. Putri et al. [22] highlighted a limitation related to the use of mixed-language content—specifically English and Indonesian—which affected the performance of traditional methods like TF-IDF by treating semantically similar words in different languages as unrelated. It was therefore suggested that future work should explore models using single-language input to improve term consistency. Addressing this issue, our study employs mBERT's multilingual capabilities to handle both languages simultaneously, enabling a more robust and semantically accurate classification process.

The novelty of this research lies in three key contributions. First, we leverage a multilingual transformer model, mBERT, which enables contextual understanding across languages and is particularly effective for handling Indonesian and English texts—a common characteristic of scientific papers in Indonesia. Second, unlike previous studies that often focus on monolingual datasets or limited thematic scopes, our approach applies mBERT to a diverse corpus of multilanguage scientific articles, enhancing the generalizability and inclusivity of SDG-related classification. Third, we conduct a comprehensive comparison between mBERT and a traditional machine learning baseline, SVM, to empirically validate the advantages of transformer-based models in this context. The performance of each model is evaluated to determine the most effective classifier, which is subsequently used to analyze the relevance of scientific papers published in 2024 to the SDGs. These contributions collectively demonstrate a significant advancement in automated, language-agnostic SDG classification of scientific literature.

2. Methods

This study follows the six stages of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology adopted for text mining, as illustrated in Figure 1. These stages include business understanding, data understanding, data preparation, modeling, evaluation, and deployment [23].

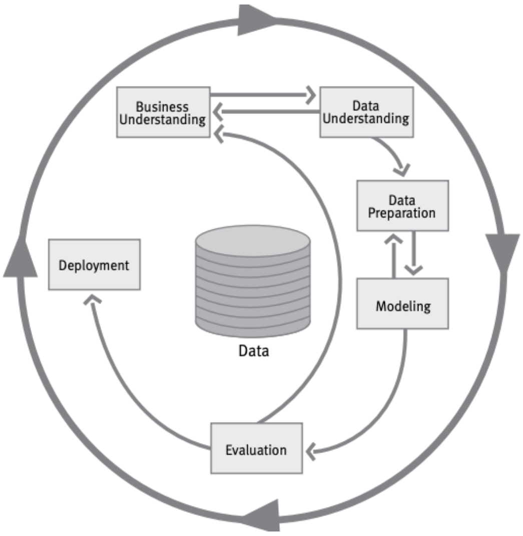


Figure 1. Life Cycle Phases of CRISP-DM [20]

2.1 Business Understanding

The primary focus of the first phase is to understand the business objectives of the research. To accomplish this, a comprehensive literature review was conducted to explore the current landscape and identify the existing gaps in tracking and analyzing research contributions related to the Sustainable Development Goals (SDGs). This review highlighted that traditional methods for assessing the alignment of scientific work with the SDGs are often manual, time-consuming, and limited in scalability—particularly when dealing with multilingual publications. Recognizing these challenges helps define a clear research direction that addresses a real-world need in both academic and policy-making contexts.

Understanding this objective is essential to ensure that the resulting model aligns with the intended purpose. The literature findings underscore a pressing demand for automated and scalable approaches that can process and classify scientific documents in multiple languages. Such methods are crucial for supporting the global SDG agenda by enabling more comprehensive and timely insights into how academic research contributes to sustainable development.

2.2 Data Understanding

Scientific publications related to SDGs are often produced in multiple languages. In countries like Indonesia, where the primary language is not English, research papers are commonly written in both Indonesian and English, highlighting the need for effective processing of multilingual texts. This phase begins with understanding the characteristics of the dataset previously compiled by B.S. Putri et al. [22]. Data was obtained using the Publish or Perish application from the Semantic Scholar database, based on keywords derived from the SDGs indicator which refers to the Publication of Bappenas in 2020 with the title SDGs Indicator Metadata. The Publish or Perish application and the semantic scholar database were chosen because they include title and abstract information and include research languages (English and Indonesian). The SDGs indicator is taken from a word or phrase that represents the indicator as a keyword used in data collection. Examples of the three keywords used to collect data are shown in Table 1.

Table 1. Examples of The Three Keywords Used to Collect Data

Indicator Code	SDGs Indicators	Keywords	Research Title
1.1.1	Extreme Poverty Rates	extreme poverty	Decision Support System Determines BLT Recipients with Extreme Poverty Using the Weight Product Method in Klambir V Kebun Village
7.1.1	Electrification Ratio	electrification	Web-Based Electrification Ratio Monitoring System at PT PLN (Persero) in the South Sulawesi Region
13.2.2	Total greenhouse gas emissions per year	greenhouse gas emission	Inventory of Greenhouse Gas Emissions Based on Land Use in Bogor City

The non-SDGs paper were collected using keywords that have the possibility of being related to non-SDGs, such as formulas, pure mathematics, and others. After collection, the dataset was manually labelled and filtered, resulting in a composition of 66,5% SDGs-related articles and 33,5% non-SDGs articles. A summary of the dataset characteristics is presented in Table 2. Three annotators labelled the dataset using SDGs indicator labels because it directly represented the relevance of the SDGs, 4 pillars and 17 goals of SDGs. The reliability of labels is measured using Krippendorff's alpha, a coefficient for measuring the consistency of scoring among two or more annotators against the same unit of analysis. The alpha values of the 4 pillars of SDGs and the 17 goals of SDGs are 0.869 and 0.815, respectively. It proves that the data quality is relatively good.

Table 2. Summary of Dataset Characteristics (n=8.090)

Category	Subcategory	Count	Percentage (%)
SDGs Classification	SDGs-related	5.379	66,5
	Non-SDGs	2.711	33,5
Title Language	Indonesian	4.576	56,6
	English	3.514	43,4
Abstract Language	Indonesian	3.274	40,5
	English	4.816	59,5

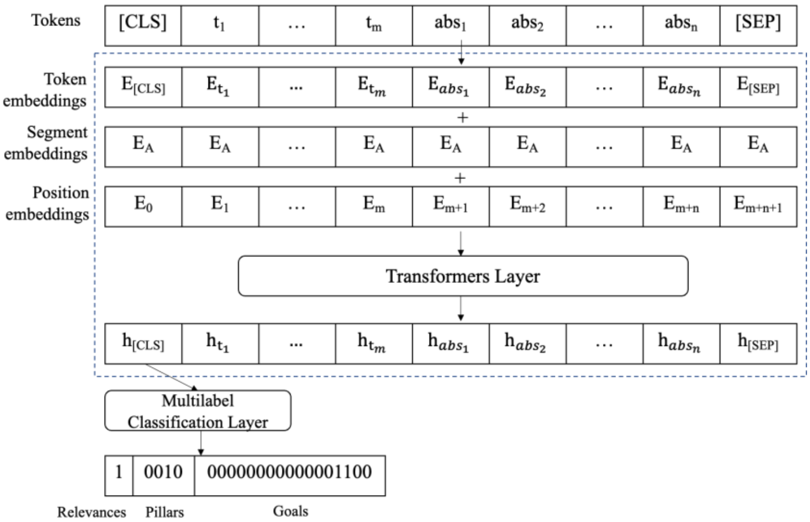


Figure 2. mBERT Architectures for multilabel SDGs

In this study, the target labels of classification are the relevance to SDGs (Yes/No), the 4 pillars of SDGs (social, economic, environmental, and law and governance), and the 17 goals of SDGs. The four pillars of the SDGs, defined by the National Development Planning Agency (Bappenas), summarize the key dimensions of the SDG framework and correspond to specific goals. The social pillar aligns with Goals 1 through 5; the economic pillar covers Goals 7, 8, 9, 10, and 17; the environmental pillar is linked to Goals 6 and 11 through 15; while the law and governance pillar corresponds to Goal 16 [24].

2.3. Data Preparation

In the data preparation phase, this study used a dataset compiled by B.S. Putri et al. [22] and subsequently split it into 80% of training, 10% of validation, and 10% of testing set. This division supported the fine-tuning of a pre-trained multilingual language model and ensured reliable performance evaluation.

The text preprocessing was applied, including case-folding, stopwords elimination, stemming, and tokenization. In case-folding process, all letters convert to the lowercase using the Pandas library. The stopword elimination process uses the Spacy library, while the stemming process to remove word suffixes and tokenization to separate text data into words uses the NLTK library.

Given the multilingual nature of the texts, all preprocessing steps were designed to maintain consistency across both Indonesian and English content. Cleaning procedures were applied by removing punctuation marks and other unnecessary characters from the texts. A pre-trained tokenizer was used to convert each input text into sub-word tokens, effectively addressing out-of-vocabulary issues. Following the standard input format of transformer-based models, a [CLS] token was added at the beginning of each sequence, followed by the concatenation of article's title and abstract, then [SEP] token to mark the end of the input [9]. These segments were then combined and converted into token IDs based on the model's vocabulary. For a machine learning model using the word vector constructed from the B.S. Putri et al. [22] dataset, all characters were converted to lowercase through case folding to reduce the vector dimension. The target labels in this study were defined for three classification objectives. The first was a classification task to identify whether a publication is relevant to the SDGs (Yes/No). The second was a multilabel classification task that assigned one or more of the 4 pillar SDGs to each SDGs-related article. The third was a multilabel classification task that assigned one or more of the 17 SDG goals to the article.

2.4 Modelling

In this phase, a transfer learning approach was applied using Multilingual Bidirectional Encoder Representations from Transformers (mBERT). This transformer-based language model was pre-trained on large-scale multilingual corpora [14]. mBERT supports over 100 languages, including Indonesian and English, which are prevalent in the dataset used in this study. mBERT has demonstrated strong cross-lingual capabilities, making it suitable for tasks involving multiple languages [25]-[27].

In this study, mBERT was employed to classify scientific publications to the SDGs in a multilabel approach. The architecture of the mBERT model is shown in Figure 2. As shown in the figure, the input tokens started with a special token CLS and ended with a special token SEP. Token t_1 to t_m refers to title tokens with a length of m, and abs_1 to abs_n refers to abstract tokens with a length of n. The input tokens are fed to the Transformer layer and yield the hidden states. The final hidden representation of the [CLS] token is fed to the multilabel classification layer and results in the sequences of 0 and 1 representing the labels. The first sequence is the relevance label of the article to SDGs. The second to fifth sequences are the 4 pillars of SDGs. The remaining sequences are the 17 goals of SDGs. As a baseline model, SVM using a one-vs-rest mechanism was developed for multilabel classification with the same sequence output labels.

2.5 Evaluation

In the evaluation phase, this study assessed the model's classification performance using standard classification metrics including accuracy, average precision which is calculated from recall and precision, and F1 score.

The evaluation metrics for accuracy, recall (R), precision (P), average precision, and F1 score are listed in Equation1 1 - 5.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$Average\ Precision = \sum_n \quad (R_n - R_{n-1})P_n \quad (4)$$

$$F1\ Score = \frac{2\ x\ P\ x\ R}{P + R} \quad (5)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. P_n and R_n are the precision and recall at the n-th threshold. In this study, we calculate the weighted-averaged F1 score, which is calculated by taking the average of all per-class F1 scores, with the contribution of each class weighted according to its support.

2.6 Deployment

At the deployment stage, a prototype inference process was conducted to simulate classification on unseen paper. A full deployment (e.g., integration into a live system or end-user tool) was not implemented. The unseen paper was collected from Garba Rujukan Digital (GARUDA) with the keyword "Sustainable Development Goals" released in 2024. GARUDA is a scientific reference discovery platform in Indonesia that provides access to scholarly works created by Indonesian researchers and academics. GARUDA's coverage consists of domestic e-journals, conferences, and research reports [28]. Using a classification model, the data of the scientific papers are categorized into the categories of relevance of SDGs (SDGs-NonSDGS), relevance of the 4 pillars of the SDGs, and relevance of 17 SDGs goals.

The results of the classification were visualized through a series of informative charts to provide insights into the relevance and distribution of scientific papers related to the Sustainable Development Goals (SDGs). First, a pie chart illustrates the overall proportion of papers classified as SDG-relevant versus non-SDG-relevant, providing a high-level view of the alignment of research with sustainability themes. Second, a bar diagram displays the distribution of SDG-related scientific papers by language, highlighting the multilingual nature of the dataset and the role of both Indonesian and English publications in contributing to the SDG agenda. Lastly, another bar chart presents the distribution of papers according to their relevance to the four pillars of the SDGs as well as their association with the specific 17 SDG goals. These visualizations collectively offer a comprehensive understanding of the scope and focus areas of the analyzed research corpus.

3. Results and Discussions

This section presents the outcomes of the study, including the performance evaluation of the proposed classification models, analysis of the prediction results, and interpretation of the findings in relation to existing research. The performance of the baseline SVM and fine-tuned multilingual BERT (mBERT) model was compared using key metrics such as accuracy, average precision, and F1. The classification outputs were further analyzed to explore the distribution of scientific papers across SDG-related and non-SDG categories, language usage, and alignment with the four SDG pillars and the 17 goals. This discussion highlights the advantages of the multilingual multilabel approach, examines the implications of the results, and reflects on the contributions and limitations of the study within the broader context of SDG-related research monitoring.

3.1 Experimental Results

The mBERT model was fine-tuned to the SDGs classification task, resulting in the relevance to SDGs, the 4 pillars of SDGs, and the 17 goals of SDGs using a maximum of 10 epochs, batch size of 4, Adam optimizer, and learning rate of 1×10^{-5}. The accuracy of the fine-tuned model in the fine-tuning process is shown in Figure 3.

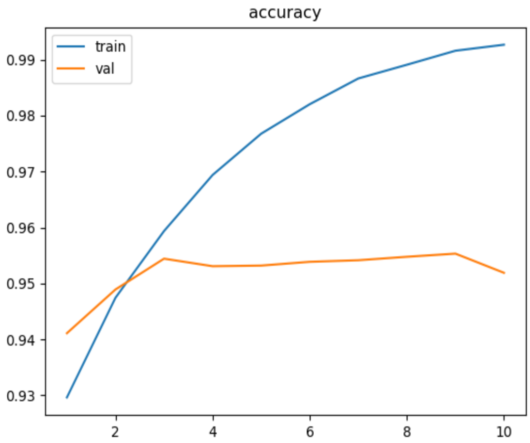


Figure 3. Accuracy of Fine-tuned mBERT-base-cased for Each Epoch in Train and Validation Sets

Based on Figure 3, the best accuracy was achieved in the ninth epoch. Therefore, the final model used for the next phase is the model at the ninth epoch. The experimental result of our proposed model compared to

SVM (linear kernel and regularization parameter C=1) as a baseline is shown in Tables 3, 4, and 5.

Table 3. Evaluation Results of Classification Model to SDGs Relevance on Testing Data.

Model	Acc	Avg Prec	F1
SVM	0.8504	0.9202	0.8568
Fine-tuned mBERT-base-cased	0.8628	0.8865	0.8639

Table 3 shows the results of the classification model to SDGs relevance. The SVM model, while achieving a respectable accuracy of 0.8504 and F1-score of 0.8568, was slightly outperformed by the fine-tuned mBERT model with accuracy of 0.8628 and F1-score of 0.8639. Although the average precision of mBERT (0.8865) was slightly lower than that of SVM, the improvement in F1-score suggests that mBERT provided more consistent predictions.

Table 4. Evaluation Results of Classification Model to 4 Pillars of

Model	Acc	Avg Prec	F1
SVM	0.6022	0.6300	0.7641
Fine-tuned mBERT-base-cased	0.6989	0.6925	0.8030

Table 4 shows the results of the classification model for the 4 pillars of SDGs. The fine-tuned mBERT model achieved a notable improvement across all major evaluation metrics: an accuracy of 0.6989, an average precision of 0.6925, and an F1-score of 0.8030. In contrast, the SVM model recorded significantly lower scores, with an accuracy of 0.6022, an average precision of 0.6300, and an F1-score of 0.7641.

These results emphasize the effectiveness of leveraging pre-trained multilingual models mBERT for multilabel classification in complex and multilingual domains such as SDG research mapping. The substantial gain in accuracy (+9.7%) and F1-score (+3.9%) highlights mBERT's superior ability to capture contextual relationships and semantic nuances in scientific texts across different languages. While the average precision also improved, the relatively modest gain suggests that while mBERT is better at achieving a balance between precision and recall, there may still be room to enhance precision further through hyperparameter tuning or by incorporating more domain-specific training data.

Table 5. Evaluation Results of Classification Model to 17 Goals of SDGs on Testing Data.

Model	Acc	Avg Prec	F1
SVM	0.4461	0.4864	0.6552
Fine-tuned mBERT-base-cased	0.5948	0.5524	0.7230

Table 5 shows the results of the multilabel classification model for the 17 goals of SDGs. The SVM model achieved an accuracy of 0.4461, an average precision of 0.4864, and an F1-score of 0.6552. In contrast, the fine-tuned mBERT-base-cased model yielded substantially better results, with an accuracy of 0.5948, an average precision of 0.5524, and an F1-score of 0.7230. These results demonstrate a considerable performance gain across all metrics, particularly in accuracy (+14.9%) and F1-score (+6.8%), suggesting that mBERT is more effective in classifying more label sequences in the multilabel approach. This performance improvement is largely attributed to mBERT's ability to generate deep contextual representations and capture semantic relationships across different languages, which is essential for handling multilingual scientific texts. Unlike traditional models like SVM that rely on sparse, surface-level features such as TF-IDF, mBERT leverages pre-trained language understanding to disambiguate meaning, manage language variability, and maintain consistent word representations—even when dealing with mixed-language content.

3.2 Prediction Result Analysis

Data on scientific papers in 2024 has been collected as many as 392. Data is collected through the GARUDA website. The SDG labels of the article are predicted using the fine-tuned mBERT as our best model.

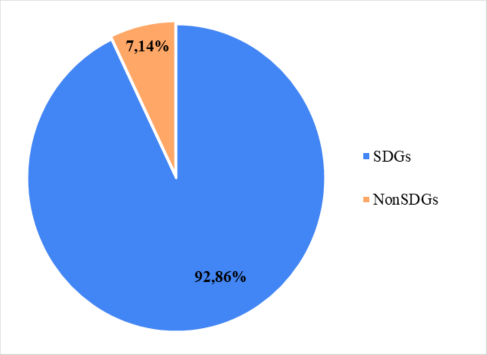
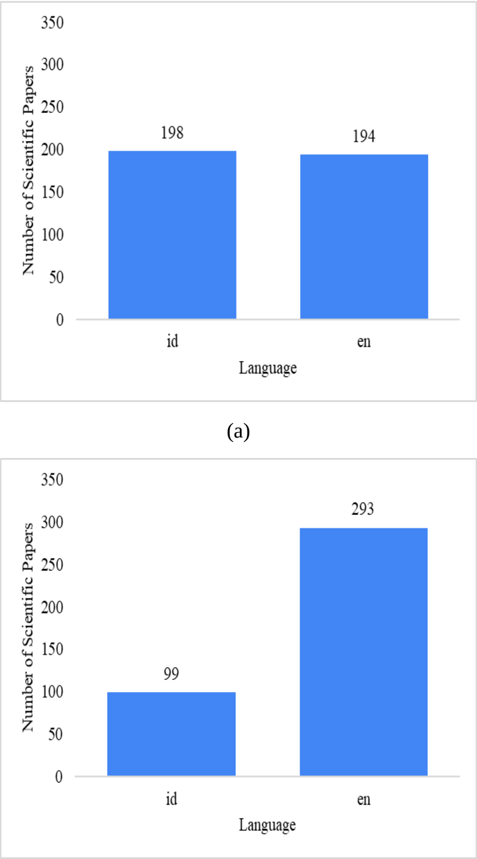


Figure 4. Distribution of SDGs and non-SDG on scientific papers

Figure 4 explains the distribution of scientific articles on SDGs and non-SDGs. The percentages are 92.86% and 7.14% for SDG and non-SDG, respectively. These results show that the majority of research in Indonesia by 2024 has already contributed to the progress of the SDGs.

Based on the title and abstract, the language of the paper consists of Indonesian and English. The distribution of data by language is presented in Figure 5. Indonesian dominated data based on titles, as many as 198 data, while English dominated data based on abstracts, as many as 293 data. It shows that scientific papers in GARUDA still maintain Indonesian as an identity in the title section. However, it adopts English in the abstract to expand accessibility at the international level.

Based on the prediction results, the distribution of scientific papers data for the four pillars of the SDGs is visualized in Figure 6. The economics category has the highest number of papers, with 213 studies. The law and governance category has the least number of papers, as many as 7 studies. This number shows that research in Indonesia in 2024 tends to discuss economics. On the other hand, the topic of law and governance can be a topic of proposal for Indonesian researchers in conducting research.



(a)

(b)

Figure 5. Distribution of research publications by language and metadata, (a) title and (b) abstract

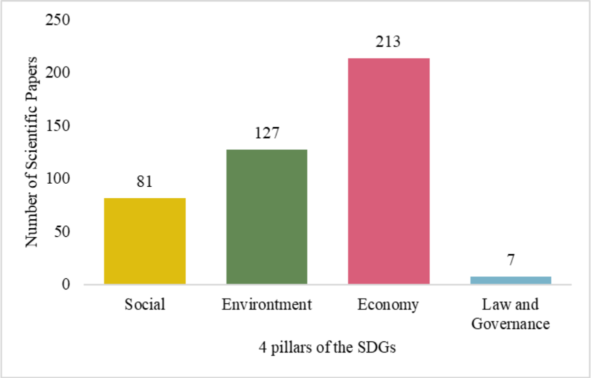


Figure 6. Distribution of SDG research publications based on 4 pillars of the SDGs

The distribution of scientific papers to the 17 goals of SDGs is visualized in Figure 7. Goal 8 had the most number of papers, with 110 studies, followed by goal 12, with 75 studies. Goal 8 discusses decent work and economic growth, while goal 12 discusses responsible consumption and production. The categories with the least number of publications are goals 16 and 17. Goal 16 discusses peace, justice, and resilient institutions, while goal 17 discusses partnerships to achieve goals. From these results, the topics in goals 16 and 17 can be some of the proposed topics for Indonesian researchers in conducting research.

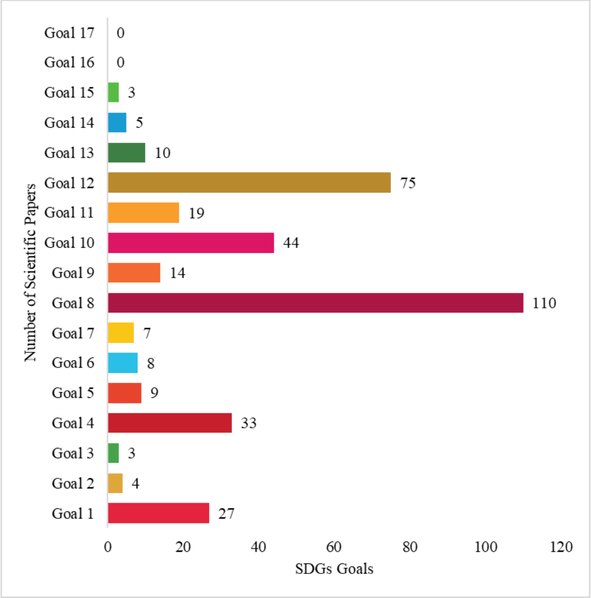
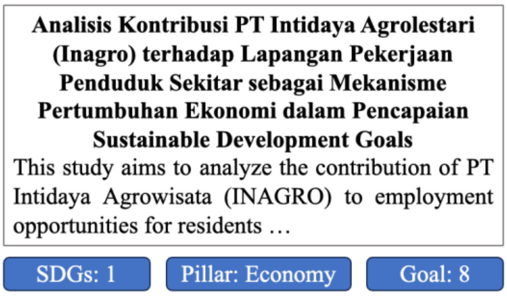


Figure 7. Distribution of SDGs scientific papers based on SDGs objectives



Analisis Kontribusi PT Intidaya Agrolestari (Inagro) terhadap Lapangan Pekerjaan Penduduk Sekitar sebagai Mekanisme Pertumbuhan Ekonomi dalam Pencapaian Sustainable Development Goals

This study aims to analyze the contribution of PT Intidaya Agrowisata (INAGRO) to employment opportunities for residents …

SDGs: 1
Pillar: Economy
Goal: 8

Figure 8. Examples of scientific papers categorized into Goal 8 SDGs

Figure 8 shows two examples of scientific papers categorized into the Goal 8 of SDGs. This is in accordance with the words contained in the title and abstract of the scientific papers exemplified. In paper 1, the words "employment", "economic growth", "employment opportunities", and "economic growth" are clearly stated in the title and abstract which explains why the paper is categorized into the goals of the 8 SDGs. In addition, the words "decent work", "economic growth", "unemployment", "decent work", and "economic growth" are also the reasons for the second paper to be categorized into the goals of the 8 SDGs. The existence of two languages listed in both papers also shows that the model accommodates both languages well, namely Indonesian and English.

3.3 Discussions

The results of this study underscore the effectiveness of leveraging transfer learning, particularly through fine-tuning the multilingual BERT (mBERT) model, for the task of classifying scientific publications according to their relevance to the Sustainable Development Goals (SDGs). Compared to the traditional Support Vector Machine (SVM) baseline, the mBERT-based model demonstrated superior performance, especially in scenarios with limited labeled data—highlighting one of the key strengths of pre-trained language models in low-resource settings.

A major advantage of using mBERT is its ability to process and understand texts across multiple languages without requiring separate models or extensive retraining for each language. This feature directly addresses the multilingual challenge posed by the global and diverse nature of scientific research. The model’s strong performance suggests that multilingual transformer-based architectures hold significant potential in building scalable, inclusive classification systems that can track research contributions toward the SDG agenda more effectively.

Moreover, the results validate the viability of automated SDG classification systems in supporting policy-making, research funding decisions, and global monitoring efforts. As the demand for evidence-based decision-making grows, such systems can provide timely and consistent insights into how scientific research aligns with global development priorities.

However, some limitations remain. While mBERT showed improved performance, the quality and balance of the training data still play a crucial role in model outcomes. The absence of comprehensive, high-quality multilingual SDG-labeled datasets continues to be a bottleneck. Future research could focus on creating more robust multilingual corpora and exploring other models, including the implementation of LLMs as studied by T. Fankhauser and S. Clematide [29] and an LLM-augmented knowledge graph as proposed by W. Benjira, et al. [30].

While this study utilizes the same dataset as the prior work by B. S. Putri et al. [22], a direct comparison of performance metrics is not presented due to fundamental differences in model design and classification objectives. The previous study implemented a multi-level classification approach, where separate models were trained to predict SDG labels at different levels or stages, effectively treating the task as a hierarchical classification problem. In contrast, our approach employs a unified multilabel classification framework, allowing simultaneous prediction of multiple SDGs in a single inference process. This difference significantly impacts how the models are trained, how label dependencies are handled, and how performance is measured. Therefore, a direct side-by-side comparison could be misleading. Instead, this study focuses on demonstrating the advantages of the multilabel approach in terms of scalability, efficiency, and the ability to capture overlapping SDG relevance in multilingual scientific articles.

In the context of the CRISP-DM framework, the deployment stage typically involves delivering the model into a production environment where it can be used by end users or integrated into operational systems. However, in this study, the deployment phase was limited to a prototype inference simulation rather than full-scale implementation. This limitation was primarily due to constraints in infrastructure, time, and access to institutional platforms where such a system could be integrated. As a result, while the model's performance was validated on unseen data in a controlled setting, automated classification in a live environment remains future work.

4. Conclusions

This study presents a multilingual, transformer-based approach for classifying scientific publications according to their relevance to the Sustainable Development Goals (SDGs) and relation to four pillars of SDGs and the 17 goals. By fine-tuning the pretrained mBERT-base-cased model on SDG-labeled data, we demonstrate its superiority over a traditional Support Vector Machine (SVM) model, particularly in handling multilingual texts and achieving more reliable multilabel classification outcomes. These findings highlight mBERT’s ability to understand and generalize across diverse linguistic contexts, making it a scalable and effective solution for global research monitoring aligned with the SDG framework.

References

[1] United Nations Development Programme (UNDP), “What are the Sustainable Development Goals?” Accessed: Jun. 18, 2025. [Online]. Available: https://www.undp.org/sustainable-development-goals

[2] P. Berrone, H. E. Rousseau, J. E. Ricart, E. Brito, and A. Giuliodori, “How can research contribute to the implementation of sustainable development goals? An interpretive review of SDG literature in management,” *International Journal of Management Reviews*, vol. 25, no. 2, pp. 318–339, Apr. 2023, doi: 10.1111/ijmr.12331.

[3] S. Sorooshian, “The sustainable development goals of the United Nations: A comparative midterm research review,” *J Clean Prod*, vol. 453, p. 142272, May 2024, doi: 10.1016/j.jclepro.2024.142272.

[4] M. Mishra *et al.*, “A bibliometric analysis of sustainable development goals (SDGs): a review of progress, challenges, and opportunities,” *Environ Dev Sustain*, vol. 26, no. 5, pp. 11101–11143, May 2023, doi: 10.1007/s10668-023-03225-w.

[5] F. Indana and R. W. Pahlevi, “A bibliometric approach to Sustainable Development Goals (SDGs) systematic analysis,” *Cogent Business & Management*, vol. 10, no. 2, Dec. 2023, doi: 10.1080/23311975.2023.2224174.

[6] N. V. Diniz, D. R. Cunha, M. de Santana Porte, C. B. M. Oliveira, and F. de Freitas Fernandes, “A bibliometric analysis of sustainable development goals in the maritime industry and port sector,” *Reg Stud Mar Sci*, vol. 69, p. 103319, Jan. 2024, doi: 10.1016/j.rsma.2023.103319.

[7] F. Invernici, F. Curati, J. Jakimov, A. Samavi, and A. Bernasconi, “Capturing research literature attitude towards Sustainable Development Goals: an LLM-based topic modeling approach,” Nov. 2024, doi: 10.1186/s40537-025-01189-4.

[8] F. Illia, R. Nooraeni, and L. H. Suadaa, “Implementation of Topic Modeling in the Analysis of Topic Trends in SDGs Goal 6 Research,” in *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICEEI59426.2023.10346917.

[9] A. Hajikhani and A. Suominen, “Mapping the sustainable development goals (SDGs) in science, technology and innovation: application of machine learning in SDG-oriented

artefact detection," *Scientometrics*, vol. 127, no. 11, pp. 6661–6693, Nov. 2022, doi: 10.1007/s11192-022-04358-x.

[10] R. C. Morales-Hernandez, J. G. Jaguey, and D. Becerra-Alonso, "A Comparison of Multi-Label Text Classification Models in Research Articles Labeled With Sustainable Development Goals," *IEEE Access*, vol. 10, pp. 123534–123548, 2022, doi: 10.1109/ACCESS.2022.3223094.

[11] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Engineering Journal*, vol. 60, no. 3, pp. 3401–3409, Jun. 2021, doi: 10.1016/j.aej.2021.02.009.

[12] N. Disayiram and R. A. H. M. Rupasingha, "A Comparative Study of Classifying English News Articles Using Machine Learning Algorithms," in *2022 Trends in Electrical, Electronics, Computer Engineering Conference (TEECCON)*, IEEE, May 2022, pp. 50–55. doi: 10.1109/TEECCON54414.2022.9854832.

[13] S. Aum and S. Choe, "srBERT: automatic article classification model for systematic review using BERT," *Syst Rev*, vol. 10, no. 1, p. 285, Dec. 2021, doi: 10.1186/s13643-021-01763-w.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018.

[15] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, p. 1071, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[16] L. B. Hutama and D. Suhartono, "Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic," *Informatica*, vol. 46, no. 8, Nov. 2022, doi: 10.31449/inf.v46i8.4336.

[17] F. Indriani, R. A. Nugroho, M. R. Faisal, and D. Kartini, "Comparative Evaluation of IndoBERT, IndoBERTweet, and mBERT for Multilabel Student Feedback Classification," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 6, pp. 748–757, Dec. 2024, doi: 10.29207/resti.v8i6.6100.

[18] A. N. Azhar and M. L. Khodra, "Fine-tuning Pretrained Multilingual BERT Model for Indonesian Aspect-based Sentiment Analysis," in *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, IEEE, Sep. 2020, pp. 1–6. doi: 10.1109/ICAICTA49861.2020.9428882.

[19] C. A. Bahri and L. H. Suadaa, "Aspect-Based Sentiment Analysis in Bromo Tengger Semeru National Park Indonesia Based on Google Maps User Reviews," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 17, no. 1, p. 79, Feb. 2023, doi: 10.22146/ijccs.77354.

[20] L. Pukelis, N. B. Puig, M. Skrynik, and V. Stanciauskas, "OSDG -- Open-Source Approach to Classify Text Data by UN Sustainable Development Goals (SDGs)," May 2020.

[21] L. Pukelis, N. Bautista-Puig, G. Statulevičiūtė, V. Stančiauskas, G. Dikmener, and D. Akylbekova, "OSDG 2.0: a multilingual tool for classifying text data by UN Sustainable Development Goals (SDGs)," Nov. 2022.

[22] Berliana Sugiarti Putri, Lya Hulliyyatus Suadaa, and Efri Diah Utami, "A Multilevel and Hierarchical Approach for Multilabel Classification Model in SDGs Research," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 14, no. 1, pp. 52–61, Feb. 2025, doi: 10.22146/jnteti.v14i1.16265.

[23] P. Chapman *et al.*, *CRISP-DM 1.0 Step-by-step data mining guide*. SPSS Inc, 2000.

[24] Bappenas, "Peraturan Menteri Perencanaan Pembangunan Nasional/ Kepala Badan Perencanaan Pembangunan Nasional Republik Indonesia Nomor 7 Tahun 2018 Tentang Koordinasi, Perencanaan, Pemantauan, Evaluasi, Dan Pelaporan Pelaksanaan Tujuan Pembangunan Berkelanjutan," 2018.

[25] S. Wu and M. Dredze, "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 833–844. doi: 10.18653/v1/D19-1077.

[26] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4996–5001. doi: 10.18653/v1/P19-1493.

[27] S. Wu and M. Dredze, "Are All Languages Created Equal in Multilingual BERT?," in *Proceedings of the 5th Workshop on Representation Learning for NLP*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 120–130. doi: 10.18653/v1/2020.repl4nlp-1.16.

[28] GARUDA, "Application Usage Manual of GARUDA: GARBA RUJUKAN DIGITAL," 2022. Accessed: Jun. 18, 2025. [Online]. Available: https://drive.google.com/file/d/1QEpm6q5KVSp2SW_Ai5RZ9ZYah_8YVDUw/view

[29] T. Fankhauser and S. Clematide, "SDG Classification Using Instruction-Tuned LLM," in *Proceedings of the 9th edition of the Swiss Text Analytics Conference*, Chur, Switzerland: Association for Computational Linguistics, 2024, pp. 148–156.

[30] W. Benjira, F. Atigui, B. Bucher, M. Grim-Yefsah, and N. Travers, "Automated mapping between SDG indicators and open data: An LLM-augmented knowledge graph approach," *Data Knowl Eng*, vol. 156, p. 102405, Mar. 2025, doi: 10.1016/j.datak.2024.102405.