Multi-label Classification of Scientific Articles Based on Sustainable Development Goals Using Pre-trained Large Language Model

Adhistianita Safira Husna
School of Computing
Telkom University
Bandung, Indonesia
adhisti@student.telkomuniversity.ac.id

Ade Romadhony
School of Computing
Telkom University
Bandung, Indonesia
aderomadhony@telkomuniversity.ac.id

Suryo Adhi Wibowo
School of Electrical Engineering
Telkom University
Bandung, Indonesia
suryoadhiwibowo@telkomuniversity.ac.id

***Abstract*—The Sustainable Development Goals (SDGs) are 17 global development agendas established by the United Nations (UN) through the United Nations Sustainable Development Group (UNSDG) to achieve peace and prosperity for people and the planet now and in the future. There have been many works through scientific articles written to solve problems related to these goals. However, the current classification system for scientific articles is limited to English. This classification system is crucial as it bridges the gap between locally produced Indonesian scientific knowledge and the global SDGs framework, enabling the identification, utilization, and amplification of Indonesian research contributions towards achieving sustainable development worldwide. This research develops a multi-label text classification system to classify Telkom University's thesis and scientific papers into SDGs goals in Indonesian with pre-trained transformer-based methods, focusing on single-language (IndoBERT) and multilingual (mBERT, XLM-RoBERTa, and mBART) transfer learning models. Evaluation was conducted based on the following metrics: F1-score, precision, recall, subset accuracy, and Hamming Loss; with two approaches: pre-trained word embedding with Multi-Layer Perceptron (MLP) classifier and end-to-end fine-tuning classifier. Experimental results showed XLM-RoBERTa-large outperforming others as multilingual model and achieved optimal performance with F1-score of 85.66% in end-to-end approach while the base model also recorded F1-score of 74.23% in pre-train word embedding with MLP classifier approach. These results are also due to the challenge of imbalance in the dataset. Hence, these results demonstrate the effectiveness of the transformer model for classifying multi-labelled text and the potential for the system to be used on Indonesian scientific papers.**

Keywords—Fine-Tuning, Indonesian, Multi-label Classification, Multi-layer Perceptron, Transformer

I. INTRODUCTION

The Sustainable Development Goals (SDGs) developed by the United Nations (UN) have become a global agenda adopted by various countries, including Indonesia. The achievement of these 17 sustainable goals is regularly monitored by the National Development Planning Agency (Bappenas)[1], with academic research being one of the important indicators in measuring progress. As an educational institution that actively carries out the Tri Dharma of higher education, Telkom University contributes through research and concrete activities, especially in thesis and research papers for final task. However, the large number of scientific papers produced requires an SDGs-based automatic classification system to accurately map the university's contribution. This is a challenge in itself considering that research related to multi-label classification of SDGs in Indonesia is still limited.

Applied system of the SDGs classification on the SINTA[2] platform currently only includes limited Scopus indexed journals using queries and not all local publications are integrated yet. Some previous research, such as the study of Putri et al. [1], used the SVM algorithm with a label powerset (LP) approach to classify PPPM STIS policy documents into SDGs pillars, goals, targets, and indicators, although the resulting accuracy still needs to be improved (61%-80%).

Subroto et al. [2] developed BERT and DistilBERT-based models on the Indonesian-language GARUDA dataset, but only focused on the classification of three classes (SDG 3, SDG 4, and non-SDGs), not multi-label. Meanwhile, the IndoGovBERT model by Riyadi et al. [3] shows superior performance in government document processing, but the dataset used has different characteristics from academic scientific papers.

IndoBERT, a language model tailored for the Indonesian language [4], has been effectively applied in various text classification tasks, demonstrating its versatility and efficacy. Its applications in classifications from sentiment analysis [5], [6] and hate speech detection [7] to more complex tasks like extreme multilabel classification [8]. IndoBERT's ability to understand and process Indonesian text makes it a valuable tool in these domains.

However, the main problem in local research is the lack of Indonesian multi-label datasets with 16 labels, as well as the non-optimal exploration of transformer-based methods for this case. Multilingual models such as mBERT (multilingual BERT) [9] have proven effective in international studies, such as Aurora's research that achieved a f1-score of 91.2% on the English SDGs dataset [10]. Research by Macias et al. [11] also showed the superiority of multilingual models such as XLM-RoBERTa [12] and mBART [13] on the Finnish SDGs dataset with an f1-score of 84.3%. These findings open up opportunities to adapt similar models in the Indonesian context.

This research aims to develop a transfer learning method using LLM for Indonesian text multi-label classification task to map Telkom University's scientific papers into SDGs goals. In-depth evaluation of model performance, label distribution analysis, and architecture optimization are conducted to overcome the limitations of previous research. The research results are expected to be an academic reference as well as a real contribution in accelerating the monitoring of SDGs in Indonesia through the automation of scientific document analysis.

[1] https://sdgs.bappenas.go.id/
[2] https://sinta.kemdikbud.go.id/scopus