Translation of article 10.22146/jnteti.v14i1.16265

A Multilevel and Hierarchical Approach for Multilabel Classification Model in SDGs Research

Berliana Sugiarti Putri[1], Lya Hulliyyatus Suadaa[1,2], Efri Diah Utami[2]

[1] Program Studi Komputasi Statistik, Politeknik Statistika STIS, Jakarta Timur, DKI Jakarta 13330, Indonesia
[2] Unit Kajian Sains Data, Politeknik Statistika STIS, Jakarta Timur, DKI Jakarta 13330, Indonesia

[Received: 23 September 2024, Revised: 11 December 2024, Accepted: 24 February 2025]
Corresponding Author: Lya Hulliyyatus Suadaa (email: lya@stis.ac.id)

ABSTRACT — The growing number of research publications complicates the identification of the implementation of research publications, especially related to sustainable development goals (SDGs). The research publication categorization into SDG levels has not been conducted. The Center for Research and Community Service (Pusat Penelitian dan Pengabdian Masyarakat, PPPM) Politeknik Statistika (Polstat) STIS needs this to monitor lecturers in implementing SDGs. This study aimed to implement and evaluate problem transformation methods and machine learning classification algorithms with a multilevel and hierarchical approach to categorize research publications into SDG levels. Problem transformation methods used were binary relevance, label powerset (LP), and classifier chains. Machine learning classification algorithms used were logistic regression (LR) and support vector machine (SVM). The inputs included titles, abstracts, and titles and abstracts. The best filter model that classified data into SDGs-non-SDGs was the model with titles and SVM, with an accuracy of 0.8634. The best level model for classifying data to SDG level was the model using titles, LP, and SVM with multilevel approaches. The level model classified data into four pillars, goals, targets, and indicators of SDGs, with an accuracy of 0.8067, 0.7501, 0.6792, and 0.6194, respectively. In comparison to other inputs with more comprehensive information, the results showed that title inputs yielded the best accuracy due to the simultaneous use of English and Indonesian. Future research can modify the model to utilize a single language input to optimize the term frequency-inverse document frequency (TF-IDF) process, hence, the word meanings from each language are not considered different important words.

KEYWORDS — Machine Learning, Problem Transformation Methods, Multilevel Approach, Hierarchy Approach.

I. INTRODUCTION

The world, through the commitment of United Nations (UN) member states, has committed to the Sustainable Development Goals (SDGs) as the primary goal for improving the community's quality of life. This agreement demonstrates the universal urgency of this quality of life [1]. Indonesia is encouraged to pursue sustainable development through the SDGs. Based on data released by Statistics Indonesia (Badan Pusat Statistik, BPS), the percentage of people living in poverty in Indonesia decreased by 1.86%, from 11.22% in 2015 to 9.36% in 2023 [2]. The SDGs also encouraged an increase in the gross participation rate (*angka partisipasi kasar*, APK) and pure participation rate (*angka partisipasi murni*, APM) in Indonesia for almost in all levels of education from 2015 to 2022 [3], [4]. The decrease and increase in value do not solely stop Indonesia's efforts and role in pursuing the goals and targets of the SDGs.

As a form of concrete action, universities can implement the SDGs through research, which aligns with one of their obligations stated in the *tri dharma* of higher education, namely conducting research. Research is an activity carried out by implementing scientific rules and methods to collect information, data, and insights relevant to the testing of scientific disciplines. The implementation of research is expected to improve the quality of higher education and fulfill national development [5].

The government contributes to the real implementation of the SDGs by requiring lectures seeking academic positions to conduct scientific research and publication, as stipulated in the Regulation of the Minister of State Apparatus Empowerment and Bureaucratic Reform Number 17 of 2013 [6]. The research results must be disseminated through seminars, publications, and/or patents by universities, as stipulated in the Law of the Republic of Indonesia Number 12 of 2012 [5]. The government's commitment through its regulations has increased research in quantity. Based on data published by World Class University (WCU) analysis, the number of research publications demonstrates an upward trend annually across all university affiliates [7].

According to WCU analysis data, a positive trend related to the quantity of research publications also occurred in Politeknik Statistika (Polstat) STIS [7]. It demonstrates that Polstat STIS actively contributes to the progress of research in Indonesia. In fact, the Polstat STIS actively organizes the national seminar on official statistics annually and the International Conference on Data Science and Official Statistics (ICDSOS) every two years. This activity became a forum for speakers to publish their research results [8].

The increase in the quantity of research publications makes the identifying process of the implementation of research publications increasingly difficult, including the identification of implementation at the level of SDGs. This is needed by the Center for Research and Community Service (Pusat Penelitian dan Pengabdian Masyarakat, PPPM) Polstat STIS to monitor the real action of SDGs by lecturers as one of the concrete manifestations of the tri dharma of higher education.

The categorization of lecturer research publications into the SDGs level can be done automatically by implementing a multilabel classification model with a multilevel and hierarchical approach. Multilabel classification is a type of classification in which an instance can have multiple classes at once (multilabel) [9]. The multilevel approach organizes classes into levels, without paying attention to the interrelationships between levels. The hierarchical approach arranges classes into levels by observing the hierarchical interrelationships [10], [11].

At the basic level, publications will be filtered into SDGs or non-SDGs. SDG publications are classified into four levels. At the first level, publications are classified into one or several categories on the four pillars of SDGs in a multilabel manner, namely the social, economic, environmental, and law and governance pillars. The four pillars of the SDGs are the concise dimensions of the SDGs set by the National Development Planning Agency (Badan Perencanaan Pembangunan Nasional, Bappenas) and are related to the goals of the SDGs. The social pillars relate to goals 1, 2, 3, 4, and 5. The economic pillars relate to goals 7, 8, 9, 10, and 17. The environmental pillars relate to goals 6, 11, 12, 13, 14, and 15. The pillars of law and governance relate to goal 16 [12]. At the second level, publications are classified into one or more categories of the 17 SDG goals on a multilabel basis. At the third level, publications are classified into one or more categories in a multilabel manner from Indonesia's 143 SDG targets. At the fourth level, publications are classified into one or more categories in a multilabel manner from Indonesia's 289 SDG indicators.

The problem transformation method was utilized to categorize research publication into SDG levels, thereby overcoming multilabel data. The method turns a multilabel problem into one or more single-label problems so that it can be solved using a single-label algorithm [13]. The classification algorithm used was machine learning classification algorithms, such as logistic regression (LR) and support vector machine (SVM), which are proven to provide the best performance in the classification model [13]. In addition, both algorithms are capable of handling high-dimensional data [14]. Machine learning is a category of artificial intelligence that allows computers to learn on their own from data [15]. The application of machine learning depends on the type of problem. Classification is categorized into a supervised learning paradigm known as learning through historical data.

The best model was selected to categorize the research publications of Polstat STIS lecturers into the SDG levels. The results of the publication category into the SDG levels were visualized to gain useful insights for the Polstat STIS.

Based on this background, this study aimed to implement and evaluate problem transformation methods and machine learning classification algorithms, with a multilevel and hierarchical approach to categorize research publications into SDG levels. This study also contributed to building datasets which were used in implementing and evaluating the model to meet these objectives. In addition, the best model from the results of the implementation and evaluation of problem transformation methods and machine learning classification algorithms was also applied to categorize the research publications of Polstat STIS lecturers into the SDG levels as one of the real realizations of the built model.

II. METHODOLOGY

This study focused on the implementation and evaluation of a multilabel classification model with a multilevel and hierarchical approach to categorize the research publications of Polstat STIS lecturers into the SDGs level. In addition, not all SDG targets and indicators were used because some of them are still being developed in Indonesia. The amount of data for each label was limited to a minimum of five, assuming the value already represents the labeling. Labeling was done based on the title and abstract of the publication, assuming that the title and abstract have represented the content of the research. In addition, not all content on a publication can be found.

The documents collected were research publications, excluding books. The dataset used started in 2016 because the idea of SDGs was inaugurated in 2015. The classification model processed multilingual inputs, namely Indonesian and English at the same time.

The stages of this research refer to the stages in the Cross-Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is an analytical project framework consisting of business understanding, data understanding, data preparation, modelling, evaluation, and deployment [16].

A. BUSINESS UNDERSTANDING

Business understanding aims to seek understanding related to goals and requirements, namely the definition of problems and steps to achieve goals [16]. The process in this study includes analysis of client needs and literature study. Both processes aim to understand the goals and requirements in categorizing research publications into the SDG levels.

1) CLIENT NEEDS ANALYSIS

Client needs analysis was carried out to obtain information related to background and specification of needs. The analysis was carried out through a discussion between PPPM and researchers. PPPM participated in the discussion were the head, secretary, and staff who were in charge of conducting research data management. The results indicate a necessity for conducting a study focused on the implementation and evaluation of the multilabel classification model with a multilevel and hierarchical approach to categorize the research publications of Polstat STIS lecturers into the SDG levels.

2) LITERATURE STUDY

Literature studies were carried out by collecting information related to the results of discussions between researchers and clients in the analysis of client needs. The literature study activity in this study consisted of searching for background information, objectives, and methods to implement and evaluate the model.

B. DATA UNDERSTANDING

The data understanding phase commenced with collecting raw data that were identified for quality to gain initial insights. These insights were used to form hypotheses for the next stage [16]. This process consisted of building a dataset with three processes, namely data collection, data labeling, and reliability tests between assessors.

1) DATA COLLECTION

Data were collected through the Publish or Perish application from the Semantic Scholar database with keywords specified by the researcher. The reason for retrieving publication information from the Publish or Perish application and the Semantic Scholar database was because both include title and abstract information used to represent the documents in this study.

The keyword was the SDG indicators by taking a phrase showing the intention of the SDG indicators. The use of such keywords aimed to ensure that all indicators were represented in the data. In addition, these keywords were assumed to have already covered the SDG levels. The keywords refer to the publication of SDG Indicator Metadata published by Bappenas

TABLE I
EXAMPLE OF RESEARCH PUBLICATION DATA WITH THE KEYWORD "POVERTY IN VARIOUS DIMENSIONS" AND ITS LABELING

Publication				
Analysis of Factors Affecting Poverty in Manado City Poverty is an important issue in the dimension of human development in Indonesia. This study aims to determine ...				
Label				
SDGs	***4 Pillar***	***Goal***	***Target***	***Indicator***
1	Social	1;4	1.2;4.1;4.2	1.2.2;4.1.2;4.2.2

in 2020 in Indonesian and English. An example of publication data based on keywords and their labeling is shown in Table I. The data are included in the SDG research, categorized into social pillars, goal 1 with no poverty and goal 4 with quality education. The data are also categorized into target 1.2, which is a derivative of goal, 1 as well as target 4.1 and target 4.2, which is a derivative of goal 4. The data are also categorized into indicators 1.2.2, 4.1.2, and 4.2.2. Indicator 1.2.2 represents the percentage of men, women and children of all ages, living in poverty in various dimensions. According to the national definition, indicator 4.1.2 represents the completion rate of education at the elementary/equivalent, junior high school/equivalent, and high school/equivalent levels, while indicator 4.2.2 represents the participation rate in organized learning by gender.

The publications that had been collected were then filtered to categorize publications labeled SDGs and non-SDGs using the model API from the open-source SDG (OSDG). In this study, the percentage of publications labeled SDGs and non-SDGs was not specifically determined.

In addition, 23 keywords with the possibility of being related to non-SDGs were determined, such as formulas and pure mathematics. Non-SDG publications were collected from the Publish or Perish application and the Semantic Scholar database of 500 for each keyword. Finally, the publication data were validated with a model from the OSDG.

In addition, data on research publications of Polstat STIS lecturers were also collected to be categorized into SDG levels. Data were collected from the Science and Technology Index (Sinta) website, which is synchronized with the Scopus, Web of Science (WoS), Garuda, and Google Scholar databases. The Sinta website is a web-based research information system that measures the performance of science and technology [17].

The collected data were selected to obtain research that included each indicator. The selection consisted of filtering the year of publication; deleting publications other than English and Indonesia; deleting publications without author, title, or abstract; eliminating data duplication; and ensuring that each keyword had a minimum of five, assuming that the value already represented the label.

2) DATA LABELLING

Three annotators labeled the data with labels in the form of SDG indicators represented by indicator codes. The final label was determined by the principle of majority voting. The annotators at this stage were Polstat STIS students who had studied Advanced Official Statistics courses containing SDGs material, assuming that annotators already had basic knowledge of the SDGs.

3) RELIABILITY TEST BETWEEN ASSESSORS

The reliability test between appraisers was carried out to find out the agreement of the appraiser in categorizing the data. The measure used was the Krippendorff's alpha coefficient. An agreement is said to be sufficient if the value of Krippendorff's alpha coefficient is at least in the range of 0.60 – 0.80, with an interpretation in the form of an actual agreement [18].

Krippendorff's alpha coefficient in this study was calculated for each level. The coefficient calculation was not done for each category at each level due to the large number of categories at each level. In addition, Krippendorff's alpha coefficient also accommodates multivalued fillings [19].

C. DATA PREPARATION

The data preparation phase included activities to build the final dataset. The final data were used for model development [16]. In this study, this process included data preprocessing, multilabeling binarizer, and feature extraction. This phase was carried out for all datasets that have gone through the dataset building process.

1) DATA PREPROCESSING

The first step was case folding. At this stage, the letters were converted to lower case or lowercase letters. This study used the Pandas library to perform the case folding stage.

Following the previous stage was stopwords and characters eliminaton. In this study, the SpaCy library was employed to eliminate English or Indonesian stopwords due to its extensive collection of stopwords [20]. In addition, the elimination of Indonesian stopwords was carried out according to the researched list of stopwords [21]. This study also eliminated nonalphabetic characters using regex libraries.

After that, stemming was conducted. This stage was done to eliminate suffixes so that the basic word was obtained. Stemming for English data used the NLTK library, while Indonesian data used the Sastrawi library.

Subsequent stage was tokenization. Tokenization is the stage of separating text data into tokens or words which are smaller units for training the model [22]. At this stage, the NLTK library was used.

2) MULTILABEL BINARIZER

Multilabeling binarizer is the process of converting target data to a binary multilabel format. Columns represent unique labels and rows correspond to instances. This stage used the MultiLabelBinarizer class from the scikit-learn library.

3) FEATURE EXTRACTION

The feature extraction stage was carried out using the term frequency-inverse document frequency (TF-IDF) technique. TF-IDF is a text feature extraction that maps each word and multiplies TF and IDF as weighting in the document [23]. TF-IDF uses the TfidfVectorizer class from the scikit-learn library with the following (1) and (2).

$$TF_{(t,d)} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (1)$$

$$IDF_{(t)} = log\frac{1+n}{1+df(t)} + 1 \quad (2)$$

where $f_{t,d}$ is the frequency of the word t in the document d, $\sum_{t' \in d} f_{t',d}$ is the total word t in each document d, n is the number of training documents used, and $df(t)$ is the number of training documents that contain the term.

D. MODELING

This study built two types of models, namely the SDG-Non-SDG research classification model as a filter model and the multilabel classification model into the SDG level as a level model. The filter model is a multiclass classification model, while the level model is a multilabel classification model. Therefore, only the level model uses the problem transformation method. The level model is built using a multilevel and hierarchical approach. The dataset input consisted of a title, abstract, and title and abstract.

The development of filter models and level models with a multilevel approach involved grid search in the search for the best parameters and a 5-fold cross-validation mechanism. In the grid search process, max_features values were also determined using pipelines from the scikit-learn library. The construction of a level model with a hierarchical approach also used parameters from a level model with a multilevel approach.

1) FILTER MODEL DEVELOPMENT

The filter model is a model that categorizes research publications into the SDG or non-SDG categories. This model used machine learning classification algorithms, namely LR and SVM obtained from scikit-learn libraries.

2) LEVEL MODEL DEVELOPMENT

The level model is a multilabel classification model that categorizes publications into SDG levels consisting of four pillars, goals, targets, and SDG indicators. Furthermore, classification algorithms were used to categorize the data.

In the development of a model with a multilevel approach, the problem transformation methods used were binary relevance (BR), powerset label (LP), and classifier chains (CC). BR is an algorithm that breaks down a multilabel classification problem into an independent binary classification problem. LP is an algorithm that makes every unique occurrence of a set of labels in the training data considered a class for the newly transformed dataset. CC turns multilabel learning problems into binary classification problem chains. Each classifier combines the classes predicted by the previous classifier as a new feature [24], [25]. An illustration of the method is shown in Figure 1.

The classification algorithms used, namely LR and SVM, have been proven to provide the best performance in the classification model [13]. In addition, both algorithms are used because they can handle high-dimensional data [14] according to the characteristics of the data used. The problem transformation method was obtained from the scikit-multilearn library, while the classification algorithm was obtained from scikit-learn.

In the construction of a level model with a hierarchical approach, the problem transformation method used was one of BR, LP, or CC according to the results of the multilevel approach grid search. The classification algorithm used also complied with the results of the multilevel approach grid search between LR and SVM.

The level model with a hierarchical approach arranged classes into levels by paying attention to the relationship between levels hierarchically. The model was built flat for each level, and the predictions at the top level were additional inputs for the next level. The process was repeated until the bottom level (top-down approach) [10], [11].

E. EVALUATION

The performance of the classification model is seen from the value of the evaluation size on the testing data. To produce a more consistent evaluation size value, the model was built using k-fold cross validation because the dataset has the same chance, which is 1/k for training data and testing data. This study evaluated the model using 5-fold cross-validation.

Filter model evaluation measured the accuracy, F1-score, and execution time. The evaluation measures of the level model, namely accuracy, F1 score, hamming loss, and execution time. Accuracy is the ratio of correctly labeled to total labels. The F1 score is the harmonic average between recall and precision. A recall is the ratio of the correct positive classification to the number of classifications that should be positive. Precision is the ratio of the correct positive classification to the amount of data predicted positively. Hamming loss is the ratio of incorrectly labeled oyang to total labels. The execution time in this study is the time needed for the algorithm to display the model training results [13]. The formula for accuracy, recall, precision, and F1 score is listed in the following (3)–(6).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

$$Recall \ = \frac{TP}{TP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$F1\ score = \frac{2 \times (recall \times precision)}{(recall + precision)}. \quad (6)$$

TP represents the model predicting positive classes correctly, according to actual conditions. TN represents the model that predicts negative classes correctly, according to actual conditions. FP represents the model predicting the class as positive, while the actual condition is negative. FN represents the model predicting the class as negative, while the actual condition is positive.

The best filter models are those with high accuracy and F1 scores, as well as optimal timing. The best level models are those with high accuracy and F1 scores, low hamming loss, and optimal timing [13].

F. DEPLOYMENT

The best model was implemented to categorize the research publications of Polstat STIS lecturers into the SDG levels. The results of the category were visualized to gain insight into the Polstat STIS which is detailed as follows.

1. Distribution of SDGs and non-SDGs publications that was visualized with a pie chart.
2. Distribution of SDGs publications based on language that was visualized with bar charts.
3. Distribution of publications in the level category of four pillars and SDG goals that was visualized with a bar chart.

III. RESULTS AND DISCUSSION

A. DATASET DEVELOPMENT

Three annotators labeled the dataset with SDG indicator labels representing the SDGs' four pillars, goals, and targets. After that, a labeling reliability test was carried out using Krippendorff's alpha coefficient, which was calculated for each level. The amount of data collected from keywords was 20,050 data. After that, the data were selected for research that included each indicator. The amount of data collected for model development was 8,090 data.

1) RELIABILITY TEST BETWEEN ASSESSORS

The alpha values for each level are shown in Table II. Generally, the alpha value is in the range of 0.60 – 0.80, so the

Binary Relevance (BR)

Research	Education	Health	Law
Research 1	1	1	0
Research 2	1	1	0
Research 3	1	0	0
Research 4	0	0	0

Research	Education
Research 1	1
Research 2	1
Research 3	1
Research 4	0

Research	Health
Research 1	1
Research 2	1
Research 3	0
Research 4	0

Research	Law
Research 1	0
Research 2	0
Research 3	0
Research 4	0

(a)

Label Powerset (LP)

Research	Education	Health	Law
Research 1	1	1	0
Research 2	1	1	0
Research 3	1	0	0
Research 4	0	0	0

Research	Label
Research1	110
Research2	110
Research3	100
Research4	000

(b)

Classifier Chain (CP)

Research	Education	Health	Law
Research 1	1	1	0
Research 2	1	1	0
Research 3	1	0	0
Research 4	0	0	0

Research	Education
Research 1	1
Research 2	1
Research 3	1
Research 4	0

Research	Education	Health
Research 1	1	1
Research 2	1	1
Research 3	1	0
Research 4	0	0

Research	Education	Health	Law
Research 1	1	1	0
Research 2	1	1	0
Research 3	1	0	0
Research 4	0	0	0

Description:
Input
Output

(c)

Figure 1. Illustration of the problem transformation method, (a) binary relevance (BR), (b) powerset label (LP), (c) classifier chain (CC).

data quality is relatively good. The alpha value decreased from level 1 to level 4 as the number of categories increased, so the possibility of annotators to label differently increased.

2) DATASET CHARACTERISTICS

Figure 2 shows that the distribution of SDGs and non-SDG research publications differs from each other. SDG achieved a percentage value of 66.5%, while non-SDG achieved 33.5%. These values indicate that the dataset is unbalanced.

Based on the language, the dataset was divided into Indonesian and English publications from the title or abstract section. The visualization of dataset distribution based on the language of the texts is presented in Figure 3. Indonesian language was more prevalent on titles, totaling 4,576 publications. Meanwhile, English was more dominant in abstracts, with 4,816 publications. It aligns with the keywords used for data collection, which included both Indonesian and English terms related to SDG indicator metadata. In addition, the difference in the number of Indonesian and English research publications based on titles and abstracts is relatively small.

B. CLASSIFICATION MODEL OF MULTILEVEL APPROACH

1) FILTER MODEL: RELATED TO SDGS

The filter model classified the data into the SDGs-non-SDG categories. Before the classification, a grid search was

TABLE II
KRIPPENDORFF'S ALPHA COEFFICIENT VALUE EACH LEVEL

No	Level	Alpha	No	Level	Alpha
1	Four pilar of SDGs	0.869	3	SDG targets	0.760
2	SDG goals	0.815	4	SDG indicators	0.734

conducted using 5-fold cross-validation to obtain the best parameters. Based on the evaluation results in Table III, the best filter model is the model with title inputs and SVM algorithm because the prediction results of the model achieve the highest accuracy and F1 scores. In addition, the best filter model also had a fairly optimal execution time.

Adding information to the model's inputs, such as abstracts, usually increases the size of the evaluation because the training data will delve more deeply into the true meaning of the text [26], [27]. This study differs, nevertheless, in that the model including the title has a superior assessment measure value than the title and abstract. It is in accordance with [28] showing that the addition of information leading to the word addition to the vocabulary does not guarantee the production of important word features with high accuracy values.

Prior research has also shown that the addition of information to the model input does not affect the increase in the value of the evaluation measure [29]. The object of this research was research publications from the Journal of Universal Computer Science (J.UCS) and the Association of Computing Machinery (ACM). In the J.UCS data, the model with title input alone produced accuracy and F1 scores of 0.85 and 0.91, respectively. Meanwhile, the model with title and keyword input produced accuracy and F1 scores of 0.65 and 0.71, respectively. In ACM data, the model with title input alone produced accuracy and F1 scores of 0.85 and 0.92, respectively. Meanwhile, the model with title and keyword input yielded accuracy and F1 scores of 0.79 and 0.86 [29].

Based on this, it is possible that the model utilizing title input is sufficient to capture the essential words needed to construct the model. In addition, the model in this study also processed multiple languages at once, namely Indonesian and English. As a result, a single word with the same meaning from both languages is considered two different important words. For example, the words "poverty" and "poor" are two different important words if the model processes two languages at once and uses TF-IDF feature extraction. Based on this, models with abstract inputs as well as titles and abstracts had lower performance values because of the large number of different important words that had the same possible meaning. Therefore, further research can recite the development of models with only one language.

2) LEVEL 1 MODEL: 4 PILAR SDGS

The level 1 model classified data into four pillar categories of the SDGs. Before classification, a grid search was performed with 5-fold cross-validation to obtain the best parameters.

Based on the evaluation results in Table IV, the best level 1 model is the model with title input, LP method, and SVM algorithm. It is because the model had the highest accuracy and F1 scores, with 0.8067 and 0.8353, respectively. It also had the lowest hamming loss, with 0.0830. The best level 1 models also had quite optimal execution times.

LP was the best method in this study due to its ability to consider the intercategory linkages that the BR method did not overcome by making labels a separate unique combination. The LP method also did not classify labels in a chain from previous inputs like the CC method. The CC method accumulated previous classifier errors, so the category results tended to be inaccurate [25]. Meanwhile, SVM was the best algorithm in this study due to its ability to maximize margins [30].

3) LEVEL 2 MODEL: SDGS GOALS

The level 2 model classified data into SDG goal categories. Before classification, a grid search was carried out with 5-fold cross-validation to get the best parameters.

Based on the results of the evaluation in Table IV, the best level 2 model is the model with title input, LP method, and SVM algorithm. It is because the model achieved the highest accuracy and F1 scores of 0.7501 and 0.7922, respectively. In addition, it had the lowest hamming loss, with 0.0267. The best level 2 model also had a fairly optimal execution time.

The best models selected at this level were in line with [13]. The best model for classifying articles into the 17 SDG goals is a combination of models with LP methods and SVM algorithms [13].

4) LEVEL 3 MODEL: SDGS TARGETS

The level 3 model classified the data into SDGs target categories. Before classification, a grid search was carried out with 5-fold cross-validation to obtain the best parameters.

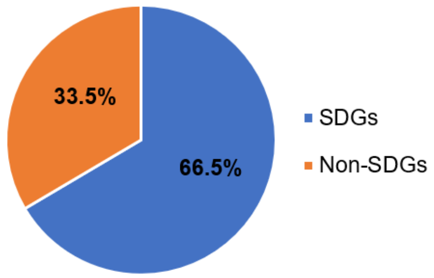

Figure 2. Distribution of SDGs and non-SDGs research publications.

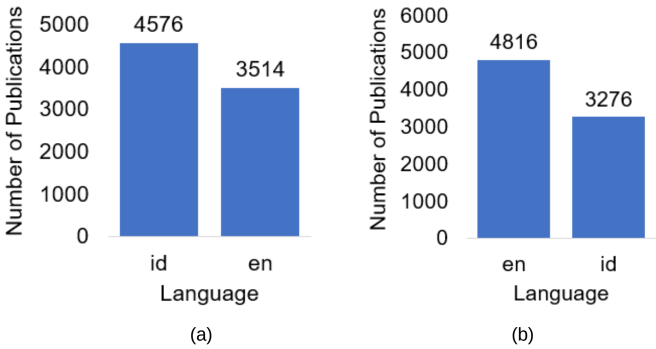

Figure 3. Distribution of research publications based on language and metadata, (a) title, (b) abstract.

TABLE III
FILTER MODEL CLASSIFICATION PERFORMANCE EVALUATION

Evaluation Metrics	LR			SVM		
	Title	*Abstract*	*Title + Abstract*	*Title*	*Abstract*	*Title + Abstract*
Accuracy	0.8476	0.8408	0.8435	0.8634	0.8410	0.8516
F1 Score	0.8204	0.8102	0.8136	0.8382	0.8065	0.8184
Time (s)	0.3123	15.390	0.8025	14.902	65.564	37.591

Based on the evaluation results in Table IV, the best level 3 model is the model with title input, LP method, and SVM algorithm because it achieved the highest accuracy and F1 scores of 0.6792 and 0.6829, respectively. The value of the hamming loss was the fourth smallest at 0.0045. The best level 3 model also had a fairly optimal execution time.

5) LEVEL 4 MODEL: SDGS INDICATORS

The level 4 model classified data into categories of SDGs indicators. At this level, a grid search was performed with 5-fold cross-validation to get the best parameters.

Based on the results of the evaluation in Table IV, the best level 4 model is the model with the input of the title, LP method, and SVM algorithm because it had the highest accuracy and F1 scores of 0.6194 and 0.6158, respectively. The value of the hamming loss was the third smallest at 0.0029. The best level 4 model also had a fairly optimal execution time.

C. CLASSIFICATION MODEL OF HIERARCHY APPROACH

The level model with a hierarchical approach was developed using the LP and SVM methods according to the best model in the multilevel approach. Level models with a hierarchical approach also used parameters from the multilevel.

In Table V, the best hierarchical model for each level is the model with title inputs because it had the highest accuracy and F1 scores and the smallest hamming loss at levels 3–4. At levels 1–2, the highest accuracy and F1 scores and the smallest hamming loss was achieved by model with abstract inputs as

TABLE IV
PERFORMANCE EVALUATION OF CLASSIFICATION MODELS LEVEL 1-4 MULTILEVEL APPROACH

Method	Algorithm	Accuracy	F1 Score	Hamming Loss	Time (s)	Accuracy	F1 Score	Hamming Loss	Time (s)	Accuracy	F1 Score	Hamming Loss	Time (s)
		Title				*Abstract*				*Title and Abstract*			
Level 1													
BR	LR	0.7291	0.8094	0.0898	0.8517	0.6889	0.7799	0.1011	284.61	0.7136	0.7992	0.0923	40.388
	SVM	0.7365	0.8217	0.0863	1000.9	0.7153	0.8055	0.0946	1158.0	0.7380	0.8235	0.0863	959.44
LP	LR	0.8058	0.8330	0.0840	0.6421	0.7724	0.7996	0.0983	23.201	0.7941	0.8218	0.0877	28.715
	SVM	0.8067	0.8353	0.0830	453.10	0.7789	0.8083	0.0953	507.82	0.8010	0.8296	0.0845	513.68
CC	LR	0.7744	0.8027	0.0978	14.548	0.7527	0.7812	0.1073	57.685	0.7722	0.7999	0.0977	86.564
	SVM	0.7846	0.8120	0.0936	909.98	0.7642	0.7941	0.1014	912.88	0.7826	0.8122	0.0931	739.34
Level 2													
BR	LR	0.6040	0.7500	0.0297	43.674	0.5072	0.6678	0.0372	20.208	0.5408	0.7003	0.0342	35.509
	SVM	0.6135	0.7482	0.0287	1335.4	0.5395	0.6830	0.0344	2081.8	0.5802	0.7195	0.0312	2471.5
LP	LR	0.7439	0.7818	0.0285	63.880	0.6651	0.7049	0.0375	10.959	0.7055	0.7459	0.0324	13.276
	SVM	0.7501	0.7922	0.0267	605.10	0.6915	0.7307	0.0335	675.93	0.7225	0.7640	0.0296	1125.4
CC	LR	0.6683	0.7446	0.0362	60.343	0.5867	0.6685	0.0450	36.306	0.6168	0.6990	0.0412	97.273
	SVM	0.6845	0.7446	0.0343	1054.2	0.6326	0.6921	0.0399	1471.8	0.6670	0.7273	0.0357	1921.7
Level 3													
BR	LR	0.5395	0.6395	0.0044	41.125	0.3719	0.4706	0.0059	108.69	0.4154	0.5157	0.0054	224.91
	SVM	0.5300	0.6249	0.0045	393.31	0.3938	0.4763	0.0055	2463.8	0.4507	0.5377	0.0050	2088.2
LP	LR	0.6670	0.6572	0.0050	83.534	0.5458	0.5323	0.0067	19.284	0.6003	0.5883	0.0058	60.395
	SVM	0.6792	0.6829	0.0045	109.53	0.5713	0.5551	0.0060	545.89	0.6202	0.6099	0.0053	476.22
CC	LR	0.5642	0.6525	0.0043	71.177	0.3949	0.4763	0.0059	169.90	0.4358	0.5224	0.0054	410.90
	SVM	0.5681	0.6469	0.0044	485.04	0.4375	0.5031	0.0055	2810.3	0.4902	0.5623	0.0049	2425.4
Level 4													
BR	LR	0.4841	0.5679	0.0027	63.391	0.2964	0.3912	0.0036	267.52	0.3288	0.4268	0.0034	466.51
	SVM	0.4670	0.5555	0.0027	998.18	0.3031	0.3905	0.0034	3731.7	0.3585	0.4465	0.0031	3022.8
LP	LR	0.6142	0.5846	0.0032	96.483	0.4598	0.4504	0.0044	41.329	0.5161	0.4990	0.0039	50.279
	SVM	0.6194	0.6158	0.0029	137.67	0.4953	0.4868	0.0038	533.03	0.5462	0.5378	0.0034	382.91
CC	LR	0.5064	0.5778	0.0026	104.14	0.3115	0.3973	0.0036	424.26	0.3524	0.4357	0.0034	1105.8
	SVM	0.5036	0.5715	0.0026	826.44	0.3445	0.4178	0.0034	4243.7	0.3986	0.4763	0.0031	3916.3

well as titles and abstracts. Models with title inputs also had the most optimal execution times for all levels. The best level model with this approach was the one with title inputs, LP methods, and SVM.

At levels 1 and 2, level models with hierarchical approaches and abstract inputs and headings and abstracts performed better than heading inputs. At levels 3 and 4, level models with a hierarchical approach and title inputs performed better than other inputs. The results showed an accumulation of errors at the previous level if the model inputs were more diverse. In addition, models with high levels and dimensions were unsuitable for complete inputs, such as abstracts or headings and abstracts, because of many factors the model must consider in classifying data. In the title input, the factors the model must consider in classifying the data were not as much as the abstract or title and abstract inputs. Based on this explanation, it is suggested to use title input if the categories used in classifying data are quite large.

D. SELECTED MODEL

The best filter model was the one with titles and SVM inputs. The best level model for each approach was the one with title inputs, LP methods, and SVM algorithms. A summary of the best models is shown in Table VI.

In Table V, the level model with a multilevel approach has the highest accuracy and F1 scores and the smallest hamming loss. However, the most optimal execution time was a level model with a hierarchical approach. The selected level model was the model with title inputs, LP methods, and SVM with a multilevel approach.

Level models with a multilevel approach had better performance than hierarchical approaches because the model stood alone for each level in building the model. The level model with a hierarchical approach must be aware of the previous level's results so that the previous level's errors were accumulated for the subsequent level. Although, a level model with a hierarchical approach could explain the relationship of categories for each level.

Model users, namely PPPM Polstat STIS, remain able to use the level model with a multilevel approach according to their needs. If the user needs the SDG goal category, the user can use the level 2 model that predicts the category at level 1 as well. However, if users need to see the categories of SDG indicators, they can use a level 4 model that predicts categories at levels 1–3 as well. Consequently, the results of the categories are not as accurate as the higher-level models. However, the model has categorized research publications into SDG levels automatically.

Table VI shows that there is a negative relationship between the number of dimensions and the value of the evaluation measure. An increase in the number influenced the decrease in the performance of evaluation measures, namely a decrease in accuracy and F1 scores and an increase in hamming loss values. It is because the model had a greater task in predicting data. In addition, an increase in the number of dimensions (features) in

TABLE V
EVALUATE MODEL CLASSIFICATION PERFORMANCE WITH A HIERARCHICAL APPROACH

Evaluation Metrics	Level 1	Level 2	Level 3	Level 4	Level 1	Level 2	Level 3	Level 4	Level 1	Level 2	Level 3	Level 4
	Title				*Abstract*				*Title and Abstract*			
Accuracy	0.8041	0.7274	0.6411	0.5874	0.8067	0.7289	0.6294	0.5722	0.8067	0.7289	0.6294	0.5722
F1 score	0.8349	0.7699	0.6483	0.5844	0.8353	0.7725	0.6390	0.5692	0.8353	0.7725	0.6390	0.5692
Hamming loss	0.0842	0.0298	0.0051	0.0031	0.0830	0.0296	0.0053	0.0033	0.0830	0.0296	0.0053	0.0033
Time (s)	46.143	61.735	57.802	58.557	211.36	289.61	257.45	179.88	176.89	243.44	216.04	151.51

TABLE VI
SUMMARY OF BEST MODELS

Model	Dimensions Number	Best Model	Accuracy	F1 Score	Hamming Loss	Time (s)
Multilevel Approach						
Filter	1	Title, SVM	0.8634	0.8382	-	14.902
Level 1	4	Title, LP, SVM	0.8067	0.8353	0.0830	453.10
Level 2	17		0.7501	0.7922	0.0267	605.10
Level 3	143		0.6792	0.6829	0.0045	109.54
Level 4	289		0.6194	0.6158	0.0029	137.67
Hierarchical Approach						
Level 1	4	Title, LP, SVM	0.8041	0.8349	0.0842	46.143
Level 2	17		0.7274	0.7699	0.0298	61.735
Level 3	143		0.6411	0.6483	0.0051	57.802
Level 4	289		0.5874	0.5844	0.0031	58.557

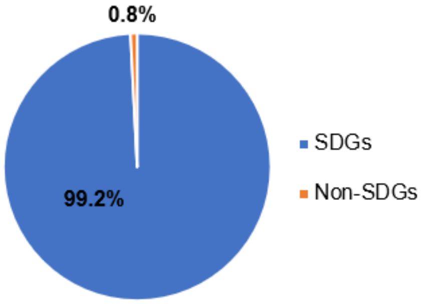

Figure 4. Distribution of SDGs and non-SDG research publications.

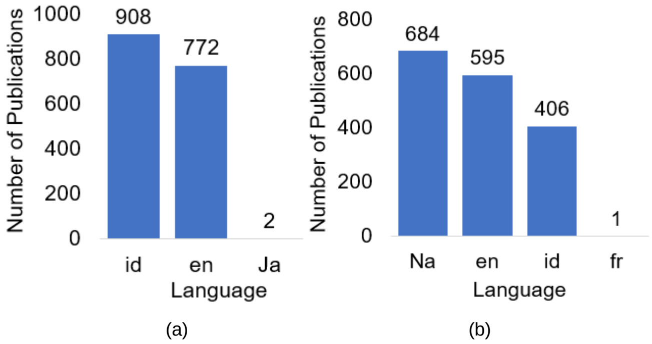

Figure 5. Distribution of research publications by language and metadata, title and (b) abstract.

a model had a positive relationship with the occurrence of overfitting [31].

Execution time was also affected by the number of vocabularies used in the model. Level 2 models using both multilevel and hierarchical approaches had the longest execution time compared to other models. In fact, the number of dimensions was less than that of the level 3 and 4 models. It is in accordance with the results of the grid search in the max_features section which affected the number of vocabularies used to build the model. Max_features at levels 3 and 4 was 3,117 words, while the max_features at level 2 was 6,235 words.

E. CATEGORY RESULT

The data of Polstat STIS lecturers from 1983 until 2024 were successfully collected, amounting to 1,687 data. Data were collected through the PPPM website, Sinta Research Data menu, on May 28, 2024. There was no selection process for these data. All lecturer research publication data were predicted to use the best model that has been built, namely a model with title input and SVM algorithm for filter models as well as models with title input, SVM methods, and SVM algorithms with a multilevel approach.

In Figure 4, the distribution of SDGs and non-SDG research publications is different from each other. The percentages are 99.2% for SDGs and 0.8% for non-SDG. These results show that the majority of lecturers' research has contributed to the progress of the SDGs.

Based on the title, the publication languages consisted of Indonesian, English, and Japanese. Based on the abstract, the publication languages comprised Indonesian, English, and French. In addition, there were publications that were not accompanied by abstracts, which were encoded with “Na” (not available). In the distribution of data based on languages in Figure 5, Indonesian dominates data based on the title of 908 data. Meanwhile, English dominates the data based on abstract, with 595 data.

Based on the prediction results, the distribution of research publication data of Polstat STIS lecturers on the four pillars of SDGs is visualized in Figure 6. The economic category had the highest number of publications, with 1232 publications. There were significant differences between the socio-economic and environmental legal and governance categories. Based on this, the majority of Polstat STIS lecturers conducted research related to social and economic. Environmental topics as well as law and governance could be some of the proposed topics for lecturers for research.

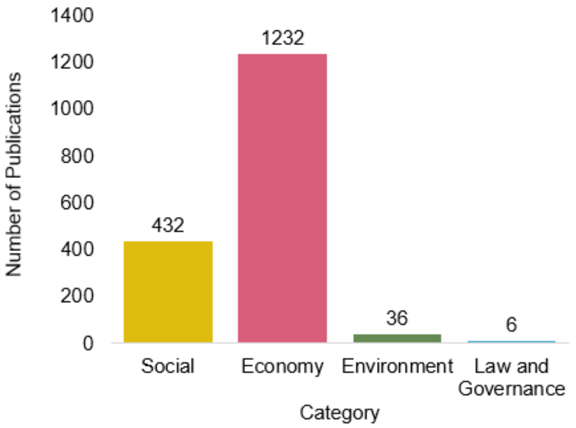

Figure 6. Distribution of SDG research publications based on 4 pillars of SDGs.

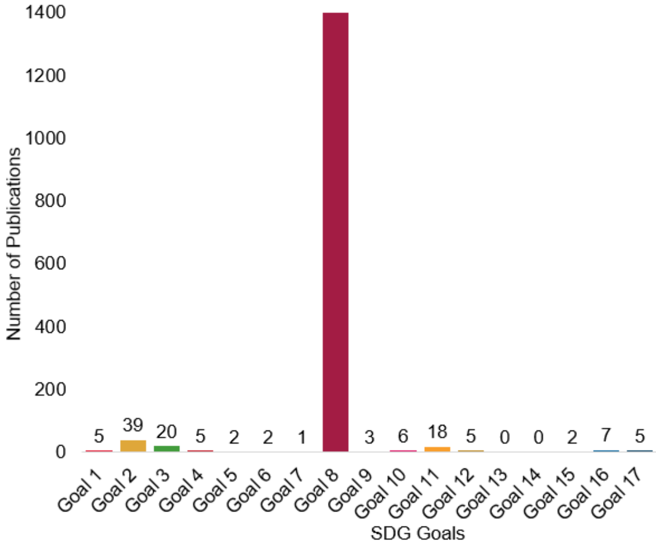

Figure 7. Distribution of SDG research publications based on SDG objectives.

Based on the prediction results, the distribution of publication data at the level of SDG objectives is visualized in Figure 7. The destination category 8 had the highest number of publications, totaling 1,495 publications, followed by the destination category 17, with 100 publications. Goal 8 discusses decent work and economic growth, while Goal 17 discusses partnerships to achieve the goals. The categories with the least number of publications were goals 13 and 14. Goal 13 addresses climate change, while Goal 14 addresses ocean ecosystems. Therefore, topics 13 and 14 can be some of the proposed topics for lecturers in conducting research.

Table VII shows the model's ability to correctly classify research publication data for each level. Table VIII shows the model's ability to correctly classify data for some levels.

The model correctly classified the data for each level listed in Table VII because the word "stunting" was clearly stated in the title. The word "stunting" was the keyword of indicator 2.2.1. This is one of the allegations related to the model easily classifying data correctly.

The model correctly classified the data for some of the levels shown in Table VIII due to the combination of words in the title, namely the words "transportation" and "growth." The word "transportation" contained the keyword indicator 11.2.1. (a) and 11.2.1. (b), that was "convenient access to public transportation" and "residents are served by public transportation." The word "growth" contained the keyword

TABLE VII
EXAMPLE OF CORRECT PREDICTIONS FOR ALL LEVELS

Research Publication					
Analysis of Factors Affecting Stunting in Indonesia in 2021 Stunting is a problem that arises due to a lack of nutritional intake and can interfere with the growth …					
	Filter Label	***4 Pilar Label***	***Goal Label***	***Target Label***	***Indicator Label***
Actual	1	Social	2	2.2	2.2.1
Predicted	1	Social	2	2.2	2.2.1

TABLE VIII
EXAMPLE OF CORRECT PREDICTIONS FOR PARTIAL LEVELS

Research Publications					
Nowcasting the Transportation and Accommodation Sectors Growth using the Google Trends Index This research aims to assess the possibility of the daily and weekly Google Trends Index (GTI) to …					
	Filter Label	***4 Pilar Label***	***Goal Label***	***Target Label***	***Indicator Label***
Actual	1	Economy, environment	8, 11	8.1, 11.2	8.1.1, 11.2.1.(a), 11.2.1.(b)
Predicted	1	Economy	8	11.3	11.3.1.(a)

indicators 8.1.1, 8.2.1, 8.9.1, 9.2.1, and 10.1.1. (e), that was "GDP growth rate," "GDP growth rate per worker," "growth rate of tourism contribution to GDP," "Manufacturing industry GDP growth rate," and "economic growth in underdeveloped areas." Indicators 11.2.1. (a) and 11.2.1. (b) were goal 11, which addressed environmental development. On the other hand, indicators 8.1.1, 8.2.1, 8.9.1, 9.2.1, and 10.1.1. (e) were goals 8, 9, and 11 that addressed economic development. These are some allegations related to the level 1 and 2 models classifying the data correctly. Likewise, some conjectures related to the level 3 and 4 models classified the data incorrectly due to the model's confusion in determining important words.

The researchers propose several suggestions to improve the model's performance. First, the best model can continue to be retrained using the latest data. Second, the model is combined with specific keywords in each category at the level of SDG goals with TF-IDF to improve feature representation. Third, the model is modified with more reliable feature extraction techniques, such as pretrained word embedding. Fourth, the model is modified with multilingual transfer learning, such as multilingual bidirectional encoder representations from transformers (mBERT) which has been previously trained on large corpus.

IV. CONCLUSION

A total of 8,090 datasets have been successfully built, with an SDG-non-SDG distribution of 66.5% and 33.5%, respectively. The number of datasets that was successfully built with SDG-non-SDG distribution was tested for reliability to determine the quality of the data. The data quality was classified as good because the alpha value was in the range of $0.60 < \alpha \leq 0.80$.

Based on the evaluation size, the best filter model was the one with title inputs and SVM algorithms. The best level model was a model with title inputs, LP methods, and SVM algorithms built with a multilevel approach. The level model classified data into four pillars, goals, targets, and SDG indicators.

Some of the proposed topics for Polstat STIS lecturers that have not been widely researched are the environment, law, and governance, especially related to handling climate change, ocean ecosystems, justice, and resilient institutions. The results of this research in the form of the best model can be used particularly in the Polstat STIS or other universities in general to monitor the contribution of the SDGs by lecturers through research and the scope of research that can be carried out further.

The study results indicate that the title input achieved the best accuracy compared to other inputs with more complete information. This was due to the simultaneous use of two languages, namely English and Indonesian. Therefore, future research may modify the model with the input of only one language to optimize the TF-IDF process so that the word meanings of each other language are not considered to be different important words.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHORS' CONTRIBUTIONS

Conceptualization, Berliana Sugiarti Putri and Lya Hulliyyatus Suadaa; methodology, Berliana Sugiarti Putri and Lya Hulliyyatus Suadaa; writing – original draf preparation, Berliana Sugiarti Putri; writing – review & editing, Berliana Sugiarti Putri, Lya Hulliyyatus Suadaa, and Efri Diah Utami; validation, Lya Hulliyyatus Suadaa and Efri Diah Utami; supervision, Lya Hulliyyatus Suadaa.

ACKNOWLEDGMENT

The author would like to thank colleagues who have provided support during this research. The authors also extend their gratitude to the Polstat STIS, especially the Center for Research and Community Service for the facilities and guidance provided.

REFERENCES

[1] Bappenas, "Sekilas SDGs." Access date: 10-Oct-2023. [Online]. Available: https://sdgs.bappenas.go.id/sekilas-sdgs/

[2] Badan Pusat Statistik (BPS), "Persentase penduduk miskin (P0) menurut provinsi dan daerah, 2007-2023." Access date: 01-Oct-2023. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/MTkyIzI=/persentase-penduduk-miskin--p0--menurut-provinsi-dan-daerah--persen-.html

[3] Badan Pusat Statistik (BPS), "Angka partisipasi kasar (APK) menurut provinsi dan jenjang pendidikan, 2003-2022." Access date: 01-Oct-2023. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/MzAzIzI=/angka-partisipasi-kasar---a-p-k--.html

[4] Badan Pusat Statistik (BPS), "Angka partisipasi murni (APM) menurut provinsi dan jenjang pendidikan, 2003-2022." Access date: 01-Oct-2023. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/MzA0IzI=/angka-partisipasi-murni---a-p-m--.html

[5] "Pendidikan Tinggi," Law of the Republic of Indonesia, No. 12, 2012.

[6] "Jabatan Fungsional Dosen dan Angka Kreditnya," Regulation of the Minister of Administrative and Bureaucratic Reform, No. 17, 2013.

[7] Ministry of Education and Culture, "WCU Analysis," 2016. Access date: Oct. 01, 2023. [Online]. Available: https://sinta.kemdikbud.go.id/wcu

[8] Politeknik Statistika STIS, "Call for Paper Seminar Nasional Official Statistics 2023." Access date: 1-Oct-2023. [Online]. Available: https://semnas.stis.ac.id/call-for-paper

[9] C. Vens *et al*., "Decision trees for hierarchical multi-label classification," *Mach. Learn.*, vol. 73, no. 2, pp. 185–214, Nov. 2008, doi: 10.1007/s10994-008-5077-3.

[10] J. Hernández, L.E. Sucar, and E.F. Morales, "Multidimensional hierarchical classification," *Expert Syst. Appl*., vol. 41, no. 17, pp. 7671–7677, Dec. 2014, doi: 10.1016/j.eswa.2014.05.054.

[11] H.S. Oh and Y. Jung, "External methods to address limitations of using global information on the narrow-down approach for hierarchical text classification," *J. Inf. Sci.*, vol. 40, no. 5, pp. 688–708, Oct. 2014, doi: 10.1177/0165551514544626.

[12] "Koordinasi, Perencanaan, Pemantauan, Evaluasi, dan Pelaporan Pelaksanaan Tujuan Pembangunan Berkelanjutan," Regulation of the Minister of National Development Planning/Head of the National Development Planning Agency of the Republic of Indonesia, No. 7, 2018.

[13] R.C. Morales-Hernández, J.G. Juagüey, 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, Nov. 2022, doi: 10.1109/ACCESS.2022.3223094.

[14] I.H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, May 2021, doi: 10.1007/s42979-021-00592-x.

[15] J. Alzubi, A. Nayyar, and A. Kumar, "Machine learning from theory to algorithms: An overview," *J. Phys. Conf. Ser*., vol. 1142, pp. 1–15, Dec. 2018, doi: 10.1088/1742-6596/1142/1/012012.

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

[17] S.L. Octaria, "Analisa Integrasi Data SINTA (Science and Technology Index)," Palembang, Indonesia, Universitas Sriwijaya, 2018. [Online]. Available: http://edocs.ilkom.unsri.ac.id/2906/1/TUGAS%202MTI%20_0903118%201621128_SITI%20LARISTA%20OCTARIA.pdf

[18] J. Hughes, "krippendorffsalpha: An R package for measuring agreement using Krippendorff's alpha coefficient," *R J*., vol. 13, no. 1, pp. 413–425, Jun. 2021, doi: 10.32614/RJ-2021-046.

[19] K. Krippendorff and R. Craggs, "The reliability of multi-valued coding of data," *Commun. Methods Meas.*, vol. 10, no. 4, pp. 181–198, Oct. 2016, doi: 10.1080/19312458.2016.1228863.

[20] M. Sighn, "Stop the stopwords using different Python libraries." Toward AI. Access date: 1-Oct-2023. [Online]. Available: https://towardsai.net/p/l/stop-the-stopwords-using-different-python-libraries

[21] F.Z. Tala, "A Study of stemming effects on information retrieval in bahasa Indonesia," Amsterdam, Netherlands, Universiteti van Amsterdam, 2003. [Online]. Available: https://eprints.illc.uva.nl/id/eprint/740/1/MoL-2003-02.text.pdf

[22] C. Toraman, E.H. Yilmaz, F. Şahinuç, and O. Ozcelik, "Impact of tokenization on language models: An analysis for Turkish," *ACM Trans. Asian Low-Resour. Lang. Inf. Proc.*, vol. 22, no. 4, pp. 1–21, Apr. 2023, doi: 10.1145/3578707.

[23] K. Kowsari *et al*., "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, pp. 1–68, Apr. 2019, doi: 10.3390/info10040150.

[24] Z. Abdallah, A. El Zaart, and M. Oueidat, "Experimental analysis and comparison of multilabel problem transformation methods for multimedia domain," in *2015 Int. Conf. Appl. Res. Comput. Sci. Eng. (ICAR)*, 2015, pp. 1–8. doi: 10.1109/ARCSE.2015.7338147.

[25] O. Ramadhani, "Klasifikasi multi-label dengan problem transformation menggunakan Python," Undergraduate thesis, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, 2020.

[26] G. Mustafa *et al*., "Multi-label classification of research articles using Word2Vec and identification of similarity threshold," *Sci. Rep*., vol. 11, Nov. 2021, Art no. 21900, doi: 10.1038/s41598-021-01460-7.

[27] M.D. Turner *et al*., "Automated annotation of functional imaging experiments via multi-label classification," *Front Neurosci*., vol. 7, pp. 1-13, Dec. 2013, doi: 10.3389/fnins.2013.00240.

[28] B.J. Hashimoto, "Is frequency enough?: The frequency model in vocabulary size testing," *Lang. Assess. Quart.*, vol. 18, no. 2, pp. 171–187, 2021, doi: 10.1080/15434303.2020.1860058.

[29] N.A. Sajid *et al*., "A novel metadata based multi-label document classification technique," *Comput. Syst. Sci. Eng.*, vol. 46, no. 2, pp. 2195–2214, Feb. 2023, doi: 10.32604/csse.2023.033844.

[30] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[31] M.A. Salam, A.T. Azar, M.S. Elgendy, and K.M. Fouad, "The effect of different dimensionality reduction techniques on machine learning overfitting problem," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, pp. 641–655, Apr. 2021, doi: 10.14569/IJACSA.2021.0120480.