International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 10, October, 2025, pp. 2535-2546
Journal homepage: http://iieta.org/journals/isi

Text Augmentation Approaches to Enhance Traditional Machine Learning Performance for SDGs Classification of Indonesian News Articles

Aina Musdholifah[1*], Siti Zaiton Mohd Hashim[2], Sri Mulyana[1], Rifki Afina Putri[1], Faizah[1]

[1] Department of Computer Science and Electronics, Universitas Gadjah Mada (UGM), Sleman 55281, Indonesia
[2] Intelligent Informatics Research Group, Faculty of Computing, Universiti Teknologi Malaysia (UTM), Johor Bahru 81310, Malaysia

Corresponding Author Email: aina_m@ugm.ac.id

https://doi.org/10.18280/isi.301001

Received: 8 September 2025
Revised: 11 October 2025
Accepted: 25 October 2025
Available online: 31 October 2025

Keywords:
Sustainable Development Goals, multilabel text classification, data augmentation, imbalanced data, Support Vector Machine, Random Forest

ABSTRACT

The Sustainable Development Goals (SDGs) require effective monitoring, yet detecting SDG-related content in Indonesian texts is difficult due to limited resources, Code-Mixing, and the multilabel nature of the task. One article may correspond to several goals, creating imbalance and inter-label dependencies that complicate classification. This study applies multilabel classification for Indonesian SDG news using Naïve Bayes, Logistic Regression, Support Vector Machine, and Random Forest with TF-IDF features and 5-Fold Cross Validation. However, these approaches showed limited performance. To improve results, four data augmentation strategies were explored for oversampling: Code-Mixing with Back Translation, Code-Mixing with Paraphrased Back-Translation, Simple Back Translation, and Paraphrased Back Translation. From 4,195 original articles on Universitas Gadjah Mada websites, 5,105 augmented samples were generated, producing 9,300 documents. Experiments show that augmentation reduces imbalance and enhances classification. SVM and RF achieved the best results, with F1-Scores above 0.93 and Hamming Loss between 0.028 and 0.067, while LR was competitive with higher efficiency. Among augmentation methods, the most effective were Code-Mixing with Back Translation and Simple Back Translation without paraphrasing. Overall, this study demonstrates that augmentation can significantly improve traditional and lightweight classifiers, offering a practical and resource-efficient alternative for SDG multilabel classification in Indonesian news and other comparable low-resource text environments.

1. INTRODUCTION

The Sustainable Development Goals (SDGs) provide a globally recognized roadmap for sustainable development, adopted by the United Nations in 2015 to replace the Millennium Development Goals (MDGs). Comprising 17 interrelated goals, the SDGs cover poverty eradication, gender equality, climate action, education, health, and institutional equity, with measurable targets to be achieved by 2030 [1]. Monitoring the Sustainable Development Goals (SDGs) has become a global urgency, as reliable tracking mechanisms are crucial for policymakers, researchers, and stakeholders worldwide. Accurate monitoring of SDG progress also enables the Indonesian government to strategically align development priorities, for example through the integration of the SDGs into the National Medium-Term Development Plan, namely RPJMN, underscoring the strategic importance of the SDGs [2]. Much of the sustainability discourse in Indonesia is reflected in news articles, blogs, policy reports, and community reports, which are characterized by heterogeneous structures and informal language styles. This makes automated monitoring systems indispensable for improving SDG tracking in the era of big data [1, 3]. However, despite this importance, significant gaps remain in the availability of tools for resource-limited languages including Indonesian. These limitations pose challenges to inclusive and equitable monitoring of SDG discourse across linguistic contexts.

Despite advances, automatic classification of long-text articles remains a persistent challenge. News texts often contain redundant phrases, context shifts, and diverse narrative styles, complicating feature extraction and semantic representation [1, 2]. Traditional machine learning models such as Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) have shown promising results in structured corpora [3, 4], but they tend to underperform on noisy, heterogeneous news data. On the other hand, deep learning approaches such as BERT, RoBERTa, and DistilBERT offer improvements by capturing contextual nuances [5], but they are particularly resource-expensive and less suitable for low-resource contexts.

The multi-label nature of SDG texts adds further complexity. A single article may address multiple goals simultaneously (e.g., education, gender equality, and poverty), resulting in inter-label dependencies that traditional binary classifiers fail to capture [3]. While earlier methods such as Binary Relevance and Label Powerset remain widely used [6, 7], recent developments focus on correlation-aware and semantic fusion strategies to better model these dependencies [5]. However, these approaches often require large and balanced datasets—an issue for low-resource languages like Bahasa Indonesia [4].

An additional challenge is class imbalance. Goals such as SDG4 (Quality Education) and SDG5 (Gender Equality) dominate most datasets, while others like SDG6 (Clean Water) and SDG14 (Life Below Water) are rarely represented [8]. This long-tailed distribution often causes classifiers to be biased toward majority classes, lowering recall for minority labels [4, 9]. While strategies like resampling, cost-sensitive learning, and ensemble models exist [4], they are not without limitations—resampling may alter meaning, and cost-sensitive models may still fail on interdependent minority classes [5].

Text augmentation has emerged as a practical and effective way to address these limitations. Techniques like Back-Translation, synonym replacement, paraphrasing, and Code-Mixing enrich minority classes by generating new, semantically faithful samples [10-12]. These methods have demonstrated improvements in macro-F1 scores and reductions in Hamming Loss in multi-label classification tasks. Furthermore, combining surface-level and embedding-level augmentation enhances model robustness—even for traditional classifiers like SVM and RF [12].

The novelty of this research lies in systematically exploring text augmentation approaches to enhance the performance of traditional machine learning models (SVM, RF, LR, NBC) for SDG classification of Indonesian news articles. While most prior studies have emphasized transformer-based architectures, this work demonstrates that effective augmentation strategies can alleviate class imbalance, enrich linguistic diversity, and improve multilabel classification accuracy even with computationally efficient classifiers. By focusing on Indonesian-language news—characterized by long text, Code-Mixing, and label imbalance—this study offers new insights into scalable, resource-friendly methods for SDG monitoring in developing country contexts.

Practical implications of this research are twofold: (i) it enables policymakers, researchers, and civil society to more accurately track SDG discourse within Indonesian media ecosystems, thus supporting evidence-based decision-making; and (ii) it provides a cost-efficient alternative for organizations with limited computational resources, demonstrating that robust SDG classification can be achieved without fully relying on transformer-based architectures. This makes the proposed approach highly relevant for governments, NGOs, and academic institutions seeking practical tools for sustainable development monitoring in low-resource settings.

To distinguish the effects of class balancing from the effects of semantic diversification introduced by augmentation, an additional baseline experiment was included in this study. This baseline evaluates the same traditional classifiers on the original dataset after simple class-balancing via random resampling, without augmentation. The comparison allows us to determine whether performance improvements stem from the increased semantic variety of augmented samples or merely from rebalancing the distribution of SDG labels.

2. RELATED WORK

In recent years, there has been a growing interest in applying text classification methods to support the mapping of Sustainable Development Goals (SDG). Morales-Hernández et al. compared traditional classifiers such as Support Vector Machine (SVM) with transformer-based models like DistilBERT, showing that transformers consistently outperform in terms of accuracy and robustness [5]. Their extended study further confirmed that neural-based models scale more effectively with increasing complexity of SDG assignments [3]. Similarly, Yao et al. employed AutoGluon, an AutoML framework, to assign multiple SDG labels to research articles, obtaining strong F1-Scores and generalizability [13]. More recently, a comparative analysis of transformer models, including BERT and RoBERTa, demonstrated their superiority over SVM for SDG text classification tasks [14]. Complementary to these studies, experimental results reported in the previous study [15] show that the integration of preprocessing techniques including TF-IDF vectorization with conventional machine learning algorithms, such as XGBoost, Random Forest, and Decision Tree, can still produce competitive classification results, highlighting the value of methodological rigor even in non-transformer approaches. In addition, the study by Sutriawan et al. [16] showed that advanced contextual embeddings such as BERT and GPT consistently outperform classical embeddings such as TF-IDF and Word2Vec, reinforcing the importance of representation learning in text classification.

Multilabel classification has become central in SDG-related tasks, since research articles often span multiple goals simultaneously. Traditional approaches such as Logistic Regression (LR) and SVM implemented in a One-vs-Rest (OvR) fashion remain competitive as baseline models [17]. Beyond these baselines, tools such as SDG-Meter [18] and corpora such as SDGi [8] have advanced multilingual and cross-domain multilabel classification. Moreover, recent studies have addressed the long-tailed distribution inherent in multilabel problems by introducing distribution-balanced loss functions that mitigate imbalance while preserving label correlations, a property particularly relevant to SDG datasets [19]. Complementing these findings, the research by Metlapalli et al. [20] shows that using traditional machine learning models such as Random Forest, NBC, and SVM, can handle classification tasks on social media documents with multiple labels and are reasonably balanced. Similarly, the study by Herrouz et al. [21] proposed a multilabel learning framework to predict citation categories based on textual features such as title, abstract, and keywords, thus demonstrating the broader utility of multilabel classification methods across scientific domains.

Another critical challenge in SDG classification lies in the imbalance of class distributions. Goals such as SDG4 (Quality Education) are frequently represented, whereas others such as SDG6 (Clean Water and Sanitation) and SDG14 (Life Below Water) are underrepresented. Prior studies have proposed ensemble and resampling strategies to mitigate this issue. Tahir et al. [22], suggested heterogeneous ensembles to address rare labels, while Tarekegn et al. [23] reviewed data-level and algorithm-level approaches, including cost-sensitive learning, to improve balancing of the data. More recently, Xiao et al. [24] introduced Pairwise Instance Relation Augmentation (PIRAN), which generates synthetic samples informed by label relationships, thereby improving robustness in long-tailed multilabel scenarios. In parallel, experiment result by Almamoori and Bhaya [25] were demonstrating that synthetic data can enhance classification performance. Although applied outside the textual domain, the conceptual approach remains relevant for mitigating imbalance in multilabel text classification [15, 25].

Finally, text augmentation has emerged as a practical solution for both enriching datasets and alleviating imbalance. In SDG-related tasks, methods such as simple Back-Translation, paraphrased Back-Translation, and Code-Mixing have been successfully employed to generate linguistic diversity while maintaining semantic fidelity. For instance, Zheng et al. [26] demonstrated that ensemble paraphrase generation and Back-Translation significantly improved performance on low-frequency labels by reducing Hamming Loss and increasing macro-F1. Likewise, insights from augmentation in non-text domains, particularly generative techniques such as CTGAN, further highlight the potential of data augmentation to reinforce classifier robustness and to enhance representation of minority labels in imbalanced datasets [25].

3. METHODOLOGY

This study focuses on the development of a multi-label classification model to identify Sustainable Development Goals (SDGs) in open-domain Indonesian-language articles. As illustrated in Figure 1, the methodology follows a structured pipeline comprising multiple stages.

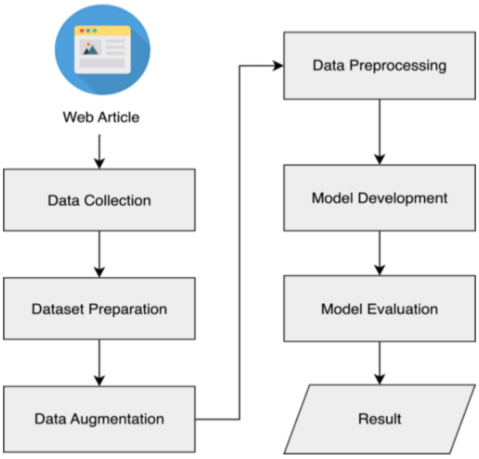


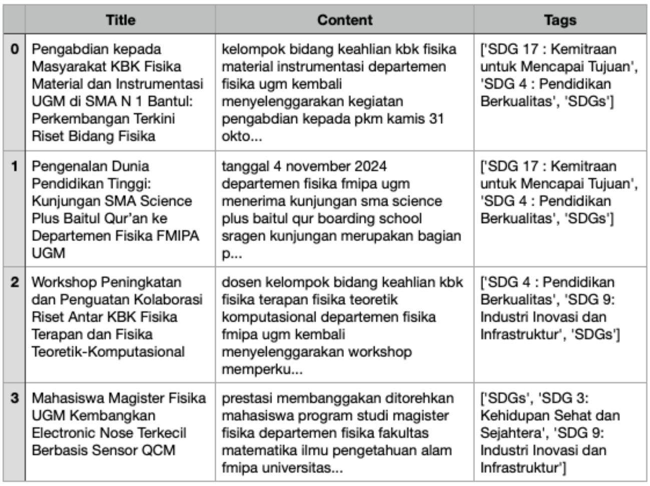
Figure 1. Workflow diagram of the proposed SDGs classification

3.1 Data collection and dataset preparation

The dataset used in this study was collected from several faculty websites under the Universitas Gadjah Mada (UGM) domain, i.e., ugm.ac.id. The data acquisition process was carried out from November 2023 to April 2025 by a web scraping approach implemented in Python using the BeautifulSoup library. This technique allowed us to automatically extract news articles published on these websites. An example of the collected SDG news articles can be seen in Figure 2.

In total, 4,195 news articles were successfully retrieved. Each article was associated with one or more Sustainable Development Goals (SDGs) based on the tags embedded within the original webpages. These tags were directly mapped to the SDG labels (SDG1–SDG17), enabling the dataset to be structured for multilabel classification.

Following data collection, a data preparation stage was carried out to transform the tags into a binary multilabel representation. For each article, the presence of a specific SDG tag was assigned a value of 1, while the absence of that tag was assigned a value of 0. As a result, each article was represented as a binary vector across the 17 SDG categories, providing the basis for multilabel classification in the subsequent stages of analysis.



	Title	Content	Tags
0	Pengabdian kepada Masyarakat KBK Fisika Material dan Instrumentasi UGM di SMA N 1 Bantul: Perkembangan Terkini Riset Bidang Fisika	kelompok bidang keahlian kbk fisika material instrumentasi departemen fisika ugm kembali menyelenggarakan kegiatan pengabdian kepada pkm kamis 31 okto...	['SDG 17 : Kemitraan untuk Mencapai Tujuan', 'SDG 4 : Pendidikan Berkualitas', 'SDGs']
1	Pengenalan Dunia Pendidikan Tinggi: Kunjungan SMA Science Plus Baitul Qur'an ke Departemen Fisika FMIPA UGM	tanggal 4 november 2024 departemen fisika fmipa ugm menerima kunjungan sma science plus baitul qur boarding school sragen kunjungan merupakan bagian p...	['SDG 17 : Kemitraan untuk Mencapai Tujuan', 'SDG 4 : Pendidikan Berkualitas', 'SDGs']
2	Workshop Peningkatan dan Penguatan Kolaborasi Riset Antar KBK Fisika Terapan dan Fisika Teoretik-Komputasional	dosen kelompok bidang keahlian kbk fisika terapan fisika teoretik komputasional departemen fisika fmipa ugm kembali menyelenggarakan workshop memperku...	['SDG 4 : Pendidikan Berkualitas', 'SDG 9: Industri Inovasi dan Infrastruktur', 'SDGs']
3	Mahasiswa Magister Fisika UGM Kembangkan Electronic Nose Terkecil Berbasis Sensor QCM	prestasi membanggakan ditorehkan mahasiswa program studi magister fisika departemen fisika fakultas matematika ilmu pengetahuan alam fmipa universitas...	['SDGs', 'SDG 3: Kehidupan Sehat dan Sejahtera', 'SDG 9: Industri Inovasi dan Infrastruktur']

Figure 2. Example of SDG news article data in Indonesian

3.2 SDGs data augmentation

To address the issue of label imbalance in the dataset, data augmentation was proposed. Several Sustainable Development Goal (SDG) categories were underrepresented, which could negatively affect the performance of the classification models. Therefore, applying augmentation techniques was suggested to improve label balance and enable the models to better capture the minority classes.

This study aims to explore four augmentation techniques that enhance linguistic diversity and strengthen the representation of minority SDG labels. These techniques preserve semantic meaning while introducing variations in vocabulary and sentence structure. The techniques include:

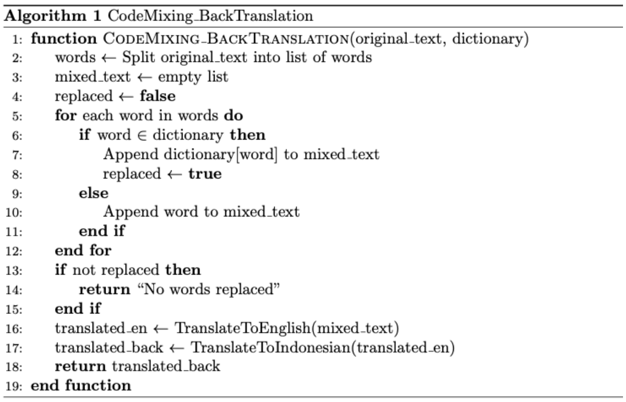
3.2.1 Code-Mixing with Back-Translation

In Code-Mixing with Back-Translation (CM+BT), the original Indonesian text is partially mixed with English words or phrases using a predefined bilingual dictionary consisting of 192 manually selected Indonesian–English phrase pairs. These phrases were chosen based on their frequency and relevance in SDG-related news articles, ensuring that the inserted English terms produce natural and domain-appropriate code-switching patterns. After the Code-Mixing step, the modified sentences are translated into English and then back into Indonesian using the GoogleTranslator interface from the deep_translator library, which serves as a simple wrapper around Google Translate and allows translation through a single function call. This process introduces natural lexical variation while maintaining semantic coherence. The algorithm tokenizes the input, applies word-level replacements based on the dictionary, ensures at least one substitution, and then performs the Back-Translation steps as illustrated in Figure 3.

3.2.2 Code-Mixing with Paraphrased Back-Translation

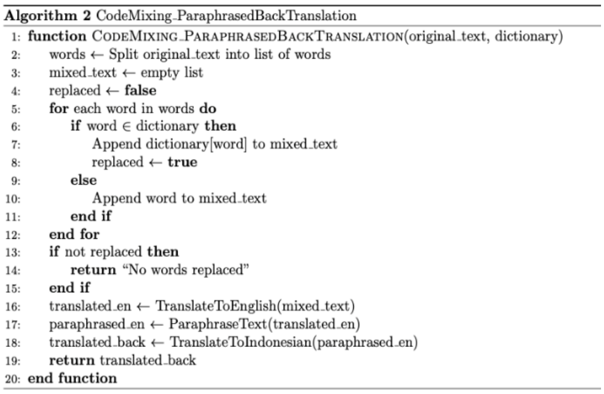
Code-Mixing with Paraphrased Back-Translation

(CM+PBT) method extends the previous strategy by adding a paraphrasing step after translating the code-mixed sentence into English. The paraphrasing is performed using a lightweight T5 (Text-to-Text Transfer Transformer) model fine-tuned specifically for paraphrase generation, namely "Vamsi/T5_Paraphrase_Paws"**,** which is available through the Hugging Face Model Hub. This model introduces syntactic variation while preserving the original meaning, enabling diverse sentence constructions before the final Back-Translation into Indonesian. The added variation enriches linguistic diversity while maintaining semantic fidelity and the Code-Mixing characteristics of the original text. The overall logic of this approach is illustrated in Figure 4.



```
Algorithm 1 CodeMixing_BackTranslation
function CodeMixing_BackTranslation(original_text, dictionary)
    words ← Split original_text into list of words
    mixed_text ← empty list
    replaced ← false
    for each word in words do
        if word ∈ dictionary then
            Append dictionary[word] to mixed_text
            replaced ← true
        else
            Append word to mixed_text
        end if
    end for
    if not replaced then
        return "No words replaced"
    end if
    translated_en ← TranslateToEnglish(mixed_text)
    translated_back ← TranslateToIndonesian(translated_en)
    return translated_back
end function
```

Figure 3. Pseudocode of CM+BT

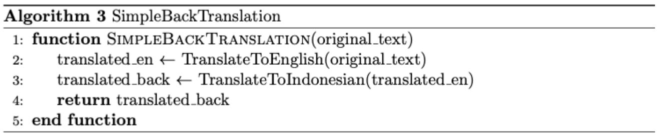


```
Algorithm 2 CodeMixing_ParaphrasedBackTranslation
function CodeMixing_ParaphrasedBackTranslation(original_text, dictionary)
    words ← Split original_text into list of words
    mixed_text ← empty list
    replaced ← false
    for each word in words do
        if word ∈ dictionary then
            Append dictionary[word] to mixed_text
            replaced ← true
        else
            Append word to mixed_text
        end if
    end for
    if not replaced then
        return "No words replaced"
    end if
    translated_en ← TranslateToEnglish(mixed_text)
    paraphrased_en ← ParaphraseText(translated_en)
    translated_back ← TranslateToIndonesian(paraphrased_en)
    return translated_back
end function
```

Figure 4. Pseudocode of CM+PBT

3.2.3 Simple Back-Translation

This baseline augmentation method, Simple Back-Translation (SBT), involves translating the original Indonesian text into English and then translating it back without any lexical or syntactic transformation in between. Despite its simplicity, this method introduces slight variations due to inconsistencies in machine translation systems, offering a low-cost way to enrich data. This lightweight method provides structural diversity useful for regularizing the model. The pseudocode for this approach is provided in Figure 5.

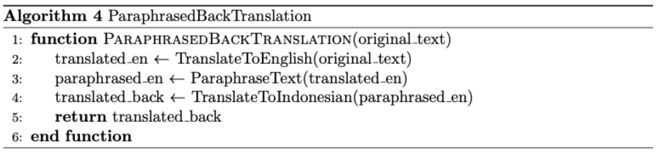


```
Algorithm 3 SimpleBackTranslation
function SimpleBackTranslation(original_text)
    translated_en ← TranslateToEnglish(original_text)
    translated_back ← TranslateToIndonesian(translated_en)
    return translated_back
end function
```

Figure 5. Pseudocode of SBT

3.2.4 Paraphrased Back-Translation

In Paraphrased Back-Translation (PBT), the Indonesian text is translated into English, paraphrased to alter its structure and wording, and then translated back. Unlike Code-Mixing methods, this technique focuses solely on restructuring the sentence, allowing the model to generalize better across a variety of expressions. This approach enhances syntactic flexibility without changing the vocabulary of the original sentence. The algorithmic flow is illustrated in Figure 6.



```
Algorithm 4 ParaphrasedBackTranslation
function ParaphrasedBackTranslation(original_text)
    translated_en ← TranslateToEnglish(original_text)
    paraphrased_en ← ParaphraseText(translated_en)
    translated_back ← TranslateToIndonesian(paraphrased_en)
    return translated_back
end function
```

Figure 6. Pseudocode of PBT

3.3 Data preprocessing and feature extraction

Before feature extraction, the collected news articles were processed through a comprehensive text preprocessing pipeline to clean and standardize the textual data, thereby reducing noise and improving model performance. The preprocessing stages included:

- Tokenization – splitting the text into individual tokens (words).
- Lowercasing – converting all tokens into lowercase to address case sensitivity.
- Stopword and Symbol Removal – eliminating semantically uninformative words, punctuation marks, and symbols to reduce noise and dimensionality [27].
- Word Normalization – reducing words to their base or canonical forms through stemming or lemmatization, which is essential to handle word variation [28, 29].

After the preprocessing, the cleaned text was transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) method. TF-IDF is widely used in text mining and information retrieval, as it reflects both the importance of terms within documents and their rarity across the corpus [30].

The TF-IDF score for a term *t* in document *d*. It is computed by Eq. (1). The resulting TF-IDF vectors were subsequently used as feature representations for training and evaluating the multilabel classification models in this study.

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \qquad (1)$$

In this context, $\text{TF}(t, d)$ is the term frequency that indicates the proportion of occurrences of term *t* in document *d*, and defined by Eq. (2):

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \qquad (2)$$

Whereas $\text{IDF}(t)$ is the inverse document frequency and achieved by Eq. (3):

$$\text{IDF}(t) = \log\left(\frac{N}{n_t}\right) \qquad (3)$$

In Eq. (3), *N* denotes the total number of documents in the corpus and n_t is the number of documents containing term *t*, thus down-weighting common terms and emphasizing more informative words [31].

3.4 Classification algorithms

In this study, four machine learning algorithms were selected for multilabel classification: Naïve Bayes (NBC), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF). These algorithms were chosen because they represent different paradigms of classification: probabilistic, linear, margin-based, and ensemble learning.

3.4.1 Naïve Bayes Classifier (NBC)

NBC is a probabilistic generative model that applies Bayes' theorem under the assumption of conditional independence between features given the class label. Despite the independence assumption being often unrealistic in natural language data, NBC has demonstrated strong performance in text classification due to the sparsity and high dimensionality of TF-IDF features [32].

The decision rule is based on maximizing the posterior probability as shown in Eq. (4):

$$\hat{y} = \arg \underset{y\in Y}{max} P(y) \prod_{i=1}^{n} P(x_i \mid y) \quad (4)$$

where, *P(y)* is the prior probability of class *y*, and $P(x_i|y)$ is the likelihood of observing feature x_i given *y*. NBC is computationally efficient, requires little training data, and often serves as a robust baseline in text classification tasks.

3.4.2 Logistic Regression (LR)

LR is a linear discriminative model that estimates the probability of a label by applying the logistic (sigmoid) function over a linear combination of input features. Unlike NBC, LR directly models the conditional probability of the class given the input features without assuming independence.

The probability function is defined by Eq. (5):

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^{\mathrm{T}}\mathbf{x}+b)}} \quad (5)$$

Classification is performed by applying a threshold, typically 0.5. For multilabel settings, LR is extended using One-vs-Rest (OvR), where one classifier is trained independently for each label [3, 32]. LR is known for its interpretability, efficiency in training, and effectiveness with high-dimensional sparse vectors such as TF-IDF.

3.4.3 Support Vector Machine (SVM)

SVM is a margin-based classifier that seeks to find the optimal hyperplane that maximizes the separation (margin) between classes. This makes SVM robust to high-dimensional feature spaces and effective when the number of features exceeds the number of samples, as is common in text classification [3].

The optimization problem is defined by Eq. (6):

$$\underset{\mathbf{w},b}{min} \frac{1}{2} \|\mathbf{w}\|^2 \text{s.t. } y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1, \forall i \quad (6)$$

SVM can also incorporate kernel functions to capture nonlinear relationships, though in text classification the linear kernel is often sufficient. Like LR, multilabel classification is handled using the One-vs-Rest scheme. SVM tends to achieve high accuracy but can be computationally expensive for very large datasets.

3.4.4 Random Forest (RF)

RF is an ensemble method that combines multiple decision trees trained on bootstrapped subsets of the dataset, utilizing random feature selection. Each tree produces a classification, and the final prediction is determined by majority voting, as shown in Eq. (7).

$$\hat{y} = \text{mode}\{h_t(\mathbf{x}) \mid t = 1,2,\dots,T\} \quad (7)$$

where, $h_t(\mathbf{x})$ is the prediction from tree t. RF reduces overfitting compared to a single decision tree and is well-suited for imbalanced datasets [33]. Its ability to capture nonlinear feature interactions makes it particularly robust for diverse label distributions, such as the SDG dataset used in this study.

3.5 Evaluation strategy

The evaluation of multilabel classifiers requires a strategy that accounts for both data imbalance and the multi-output nature of predictions. In this study, we used 5-Fold Cross Validation to ensure robust estimation, and multiple evaluation metrics tailored to multilabel classification.

3.5.1 5-Fold Cross Validation

Cross-validation reduces overfitting by dividing the dataset into five folds (k = 5). For each fold, four partitions are allocated for training and one for validation. The overall performance is calculated as the average across all folds, as shown in Eq. (8):

$$CV_{\text{score}} = \frac{1}{k}\sum_{i=1}^{k} \text{score}_i, k = 5 \quad (8)$$

This method ensures that every instance is used for both training and validation, producing a more reliable estimate of model generalization [34].

3.5.2 Precision, Recall, and F1-Score

These three metrics capture different perspectives of performance in imbalanced multilabel settings. For each label *l*, these three metrics are calculated using Eq. (9), Eq. (10), and Eq. (11).

$$\text{Precision}_l = \frac{TP_l}{TP_l + FP_l} \quad (9)$$

$$\text{Recall}_l = \frac{TP_l}{TP_l + FN_l} \quad (10)$$

$$F1_l = 2 \times \frac{\text{Precision}_l \cdot \text{Recall}_l}{\text{Precision}_l + \text{Recall}_l} \quad (11)$$

Precision is defined by Eq. (9) as the proportion of predicted positives that are correct, and a lower false-positive rate is better. Recall, as defined in Eq. (10), measures the proportion of actual positives correctly identified, and a lower false-negative rate is better. Whereas the F1-Score, which is the harmonic mean of Precision and Recall, balancing the Precision and Recall metrics, is defined by Eq. (11). For multilabel classification, micro-averaging aggregates contributions across labels, while macro-averaging gives equal weight to each label [3, 35].

3.5.3 Hamming Loss

Hamming Loss quantifies the proportion of misclassified labels, treating both false positives and false negatives equally, as depicted in Eq. (12):

$$HL = \frac{1}{N \times L}\sum_{i=1}^{N} \sum_{l=1}^{L} I(y_{i,l} \neq \hat{y}_{i,l}) \tag{12}$$

In this context, N represents the number of samples, L denotes the number of labels, and I indicates the indicator function. This approach is particularly suitable for multilabel problems, as it assesses correctness at the individual label level [36].

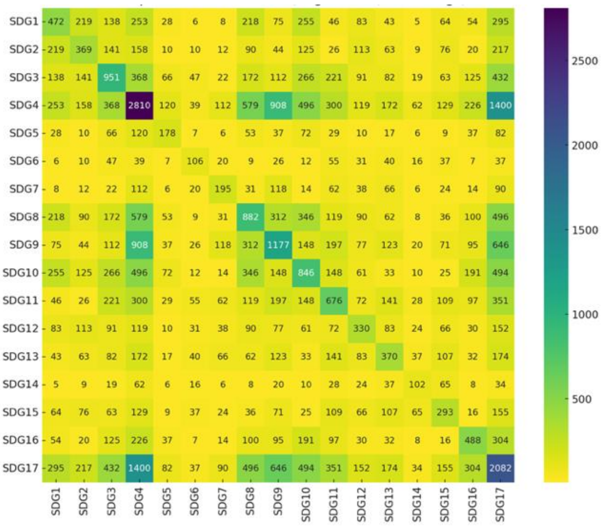
3.5.4 Imbalance analysis

Since the dataset contains minority SDG labels, imbalance significantly impacts the performance of classifiers, especially for recall. To address this, the models were evaluated both before and after data augmentation, allowing analysis of how augmentation techniques improved representation of minority classes. Previous studies show that augmentation and rebalancing strategies substantially reduce Hamming Loss and improve F1-Score for low-frequency labels [33, 36].

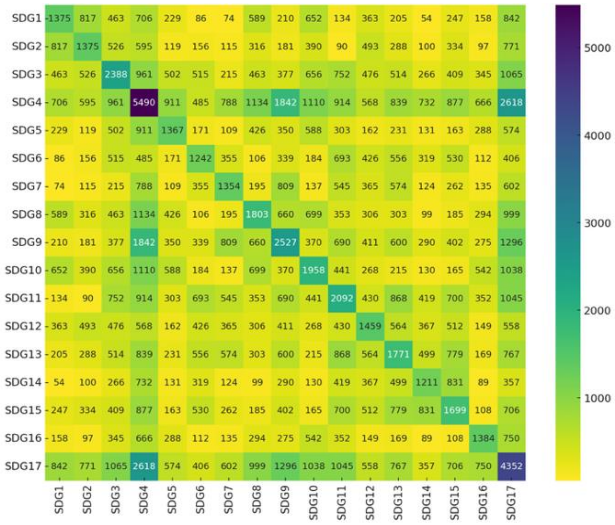
4. RESULT AND DISCUSSION

4.1 Data statistics

The co-occurrence heatmaps illustrate how frequently pairs of SDG labels appear together within the same article, where higher values indicate stronger co-occurrence relationships and lower values suggest weaker or rare co-occurrence. The comparison highlights differences between the original dataset and the augmented dataset generated through Code-Mixing and Back-Translation, as shown in Figure 7(a) and Figure 7(b).



(a) Original data



(b) Augmentation data

Figure 7. Correlation matrix

The comparison highlights a strong imbalance in the original dataset, where co-occurrence values are relatively low and dominated by a few pairs, such as SDG4–SDG4 (2810), SDG9–SDG11 (1177), and SDG3–SDG4 (951), while most other pairs occur rarely (fewer than 200). This indicates a bias toward a limited set of SDGs.

After augmentation, co-occurrence values increase significantly across nearly all label pairs, resulting in a more balanced distribution. For example, SDG4–SDG4 reaches 5,490, SDG1–SDG17 reaches 4,352, SDG9–SDG10 reaches 2,527, and SDG8–SDG12 reaches 1,803. While dominant patterns remain evident, underrepresented relationships are reinforced without introducing noise, thereby enriching the contextual overlap among labels.

Overall, augmentation improves linguistic variety, balances SDG representation, and provides stronger foundations for multi-label classification and thematic analysis. As shown in Figure 8, SDG4 appears as the most frequent label with more than 5,400 occurrences, followed by SDG17 with approximately 4,352 and SDG9 with approximately 2,527, indicating that the dataset is mainly focused on education, partnerships, and infrastructure.

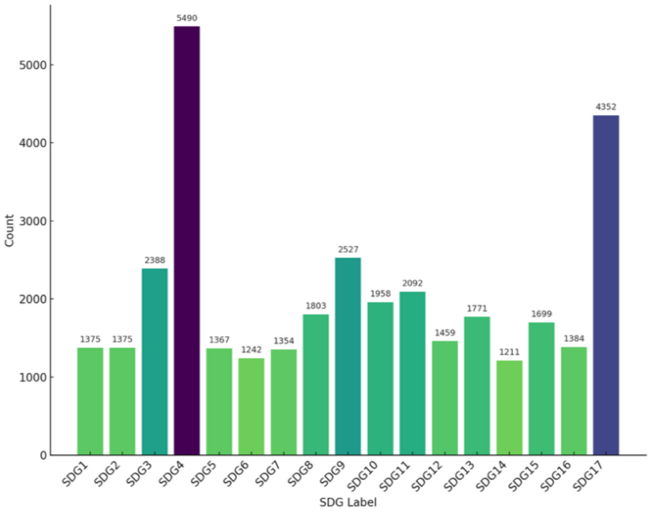


Figure 8. Label cardinality of SDG categories

Some SDGs, such as SDG5 (Gender Equality, 1,367), SDG6 (Clean Water and Sanitation, 1,242), and SDG14 (Life Below Water, 1,211), occur much less frequently, resulting in a long-tailed distribution where minority labels are underrepresented. This imbalance presents challenges for multi-label classification, as models tend to perform poorly on rare classes. To mitigate this issue, augmentation was applied, substantially increasing the frequency of minority SDGs and reducing the disparity with majority classes. As a result, the dataset becomes more balanced, model robustness is improved, and a more equitable representation across all 17 goals is achieved.

4.2 Augmentation impact

This study employed 4,195 original and 5,105 augmented Indonesian news articles, resulting in a total of 9,300 documents for SDG classification. The original dataset was highly imbalanced, with SDG4 accounting for 67% and SDG17 for 50%, while minority labels such as SDG5, SDG6, and SDG14 appeared in less than 5%. To mitigate this issue, augmentation was applied to selectively enrich minority labels while considering the multi-label nature of the dataset. After augmentation, minority labels increased substantially, with SDG5 rising from 4% to 15%, SDG6 from 3% to 13%, and SDG14 from 2% to 13%, while the dominance of SDG4 and SDG17 decreased proportionally to 59% and 47%, respectively. These results, summarized in Table 1, demonstrate that augmentation effectively rebalanced the dataset and established a more equitable foundation for model training.

From the final dataset of 9,300 articles, 7,723 were used for training and 1,577 were reserved for testing. The testing set was exclusively drawn from the original 4,195 articles and included only entries with at least one SDG label, thereby ensuring that evaluation relied on labels already assigned in the original data. Based on distribution analysis, the number of augmented samples required for each minority label was determined to achieve proportional balance. The augmentation process was then conducted using four techniques designed to strengthen the representation of minority labels: (1) Code-Mixing with Back-Translation, (2) Code-Mixing with Paraphrased Back-Translation, (3) Simple Back-Translation, and (4) Paraphrased Back-Translation.

Table 1. SDG label distribution before and after data augmentation

No.	SDG Label	Count		Proportion	
		Before	After	Before (%)	After (%)
0	SDG1	473	1375	0.11	0.15
1	SDG2	369	1375	0.09	0.15
2	SDG3	951	2388	0.23	0.26
3	SDG4	2812	5490	0.67	0.59
4	SDG5	178	1367	0.04	0.15
5	SDG6	107	1242	0.03	0.13
6	SDG7	196	1354	0.05	0.15
7	SDG8	882	1803	0.21	0.19
8	SDG9	1177	2527	0.28	0.27
9	SDG10	846	1958	0.20	0.21
10	SDG11	678	2092	0.16	0.22
11	SDG12	330	1459	0.08	0.16
12	SDG13	371	1771	0.09	0.19
13	SDG14	103	1211	0.02	0.13
14	SDG15	293	1699	0.07	0.18
15	SDG16	488	1384	0.12	0.15
16	SDG17	2084	4352	0.50	0.47

4.3 Model performance

4.3.1 Validation result

The average performance of Naïve Bayes (NBC), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) was evaluated across four augmentation methods using 5-Fold Cross Validation (CV). The reported metrics include Precision, Recall, F1-Score, Hamming Loss, and computational time, as summarized in Table 2.

Table 2. Average performance of models on SDG augmentation (5-Fold CV)

Algorithm	Precision	Recall	F1-Score	Hamming Loss	Time (s)
Code-Mixing with Back-Translation (CM+BT)					
NBC	0.9010	0.3915	0.5458	0.1426	0.3291
LR	0.9532	0.8103	0.8759	0.0502	5.7909
RF	0.9768	0.9190	0.9470	0.0225	122.3948
SVM	0.9744	0.9221	0.9475	0.0223	853.9230
Code-Mixing with Paraphrased Back-Translation (CM+PBT)					
NBC	0.9025	0.3896	0.5442	0.1429	0.3362
LR	0.9520	0.8087	0.8745	0.0508	6.4773
RF	0.9765	0.9178	0.9462	0.0228	120.4344
SVM	0.9740	0.9210	0.9468	0.0227	823.4964
Simple Back-Translation (SBT)					
NBC	0.8995	0.3836	0.5378	0.1443	0.2802
LR	0.9528	0.8103	0.8758	0.0503	8.8281
RF	0.9775	0.9186	0.9471	0.0224	121.4008
SVM	0.9741	0.9216	0.9472	0.0225	849.0328
Paraphrased Back-Translation (PBT)					
NBC	0.9038	0.3882	0.5430	0.1430	0.2726
LR	0.9527	0.8091	0.8751	0.0506	6.7884
RF	0.9776	0.9180	0.9469	0.0226	120.6313
SVM	0.9746	0.9210	0.9470	0.0226	827.0066

The experimental results indicate that SVM achieved the highest performance, with an F1-Score of 0.9475 and a Hamming Loss of 0.0223, confirming its robustness for SDG multilabel classification. Random Forest (RF) followed closely with an F1-Score of 0.9471 and a Hamming Loss of 0.224, demonstrating comparable effectiveness. Logistic

Regression (LR) produced a slightly lower F1-Score of 0.8759, but its lower computational demand makes it an attractive option in resource-constrained environments. In contrast, Naïve Bayes Classifier (NBC) recorded the weakest performance, with an F1-Score of 0.5458 and a Recall of 0.3915, highlighting the limitations of its independence assumption.

Minor variations were observed across augmentation methods; however, Code-Mixing with Back-Translation yielded the best performance for SVM (F1 = 0.9475, Precision = 0.9744, Recall = 0.9221, Hamming Loss = 0.0223), while Simple Back-Translation produced the best results for RF (F1 = 0.9471, Precision = 0.9775, Recall = 0.9186, Hamming Loss = 0.0224). LR remained stable across all augmentations, with F1-Scores ranging from 0.8745 to 0.8759 and Hamming Loss around 0.050. NBC obtained its highest F1-Score (0.5458) under Code-Mixing with Back-Translation.

Overall, these findings demonstrate that while all augmentation methods preserved stable performance, CM+BT method is most effective for SVM, whereas SBT method provides optimal performance for RF.

4.3.2 Evaluation result

The comparative evaluation between the original and augmented datasets underscores the substantial role of augmentation in improving multilabel classification performance. Using the original dataset of 4,195 articles (2,618 training and 1,577 testing), all models were constrained by class imbalance, which led to suppressed recall and modest F1-Scores. As seen in Table 3, among baseline models, SVM achieved the best performance with an F1-Score of 0.6058 and a Hamming Loss of 0.1403, indicating its relative robustness under imbalanced conditions. Random Forest (RF) followed with an F1-Score of 0.5363 and a Hamming Loss of 0.1573, showing comparable precision but reduced sensitivity compared to SVM. Logistic Regression (LR) yielded an F1-Score of 0.4473, reflecting its limited ability to capture label correlations in the presence of imbalance. Naïve Bayes Classifier (NBC) exhibited the weakest performance, with an F1-Score of 0.3572 and a Hamming Loss of 0.2038, further confirming the restrictive nature of its independence assumption.

These findings establish that, without augmentation, the models are unable to generalize effectively across minority SDG labels, resulting in high error rates and skewed predictions. Consequently, dataset augmentation becomes essential to alleviate imbalance, improve recall, and achieve more equitable performance across labels.

After augmentation, as shown in Table 4, the training set expanded to 7,723 samples while the testing set remained fixed at 1,577 original articles to ensure fair evaluation. Performance improved substantially across all algorithms: SVM reached an F1-Score of 0.9349 with Hamming Loss reduced to 0.0276, RF advanced to F1-Scores between 0.8152 and 0.8285 with Hamming Loss around 0.0668, LR improved to approximately 0.7271, nearly doubling its original performance, and NBC increased to 0.4476.

These results, summarized in Tables 3 and 4, are further visualized in Figure 9, which clearly illustrates the F1-Score improvements across all models, particularly the dramatic gains achieved by SVM and RF. This confirms that augmentation effectively mitigated label imbalance and provided a more equitable foundation for SDG multilabel classification.

Table 3. Evaluation of original data

Algorithm	Precision	Recall	F1-Score	Hamming Loss	Time (s)
NBC	0.6050	0.2534	0.3572	0.2038	0.0675
LR	0.7566	0.3175	0.4473	0.1754	0.0196
RF	0.7855	0.4071	0.5363	0.1573	1.9311
SVM	**0.8141**	**0.4824**	**0.6058**	**0.1403**	0.0218

Table 4. Evaluation of augmentation data

Algorithm	Precision	Recall	F1-Score	Hamming Loss	Time (s)
Code-Mixing with Back-Translation (SM+BT)					
NBC	0.8352	0.3036	0.4453	0.1690	0.0486
LR	0.9394	0.5896	0.7244	0.1002	0.0237
RF	0.9729	0.7189	0.8268	0.0673	3.1671
SVM	**0.9895**	**0.8842**	**0.9339**	**0.0280**	**0.0180**
Code-Mixing with Paraphrased Back-Translation (CM+PBT)					
NBC	0.8340	0.3028	0.4443	0.1693	0.0516
LR	0.9390	0.5912	0.7256	0.0999	0.0249
RF	0.9659	0.7052	0.8152	0.0714	2.8407
SVM	**0.9882**	**0.8843**	**0.9334**	**0.0282**	**0.0188**
Simple Back-Translation (SBT)					
NBC	0.8386	0.3053	0.4476	0.1684	0.0668
LR	0.9416	0.5922	0.7271	0.0993	0.0334
RF	0.9708	0.7226	0.8285	0.0668	3.2907
SVM	**0.9897**	**0.8858**	**0.9349**	**0.0276**	**0.0192**
Paraphrased Back-Translation (PBT)					
NBC	0.8329	0.3011	0.4423	0.1697	0.0524
LR	0.9421	0.5892	0.7250	0.0999	0.0242
RF	0.9682	0.7114	0.8202	0.0697	2.8958
SVM	**0.9875**	**0.8838**	**0.9328**	**0.0285**	**0.0184**

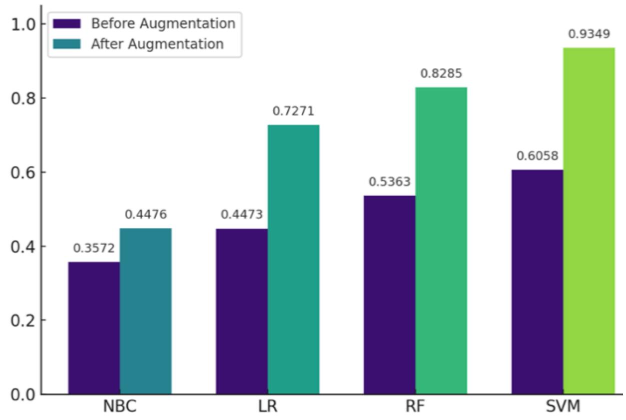


Figure 9. F1-Score before vs after augmentation

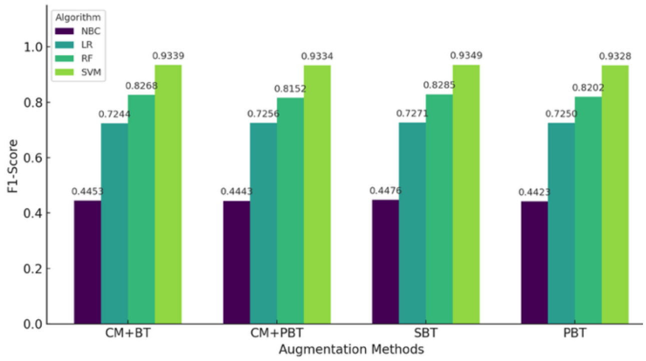


Figure 10. F1-Score across augmentation methods

Beyond the overall improvements, further analysis was conducted to assess the effectiveness of individual augmentation methods. As summarized in Table 4 and illustrated in Figure 10, SVM consistently achieved the highest performance across all methods, reaching an F1-Score of 0.9349. Among the augmentation techniques, Code-Mixing with Back Translation (CM+BT) produced the best overall result for SVM. RF also demonstrated strong performance, attaining its peak F1-Score of 0.8268 under Simple Back Translation (SBT). LR remained stable across augmentation methods, with F1-Scores ranging between 0.7244 and 0.7272, while NBC achieved only marginal improvements, with F1-Scores between 0.4423 and 0.4576. These findings demonstrate that although all augmentation methods contributed positively to classification robustness, CM+BT was most effective for SVM, whereas SBT provided the most favorable results for RF.

4.3.3 Robustness evaluation on augmented test samples

To evaluate whether augmentation improves not only training performance but also model robustness to newly introduced linguistic patterns, we expanded the test set from 1,577 original Indonesian articles to 2,500 samples by adding 923 augmented sentences generated using our two augmentation strategies: CM+BT and CM+PBT. These additional samples introduce linguistic features absent from the original distribution—such as English–Indonesian code-mixed tokens and paraphrased syntactic structures—allowing us to examine model generalization beyond the original human-written news domain. While the original test set maintains label reliability, the augmented subset serves as a controlled-challenge set for assessing robustness to distributional shifts.

Table 5. Evaluation on expanded test data

Algorithm	Precision	Recall	F1-Score	Hamming Loss	Time (s)
Code-Mixing with Back-Translation (Added Test)					
NBC	0.8520	0.3355	0.4815	0.1343	0.0540
LR	0.9309	0.5529	0.6937	0.0907	0.0251
RF	0.6129	0.5600	0.5853	0.1474	3.3881
SVM	0.9868	0.6712	0.7989	0.0628	0.0295
Code-Mixing with Paraphrased Back-Translation (Added Test)					
NBC	0.8679	0.3977	0.5455	0.1231	0.0683
LR	0.9469	0.6619	0.7792	0.0697	0.0325
RF	0.9737	0.7830	0.8680	0.0442	4.4571
SVM	0.9892	0.9125	0.9493	0.0181	0.0290

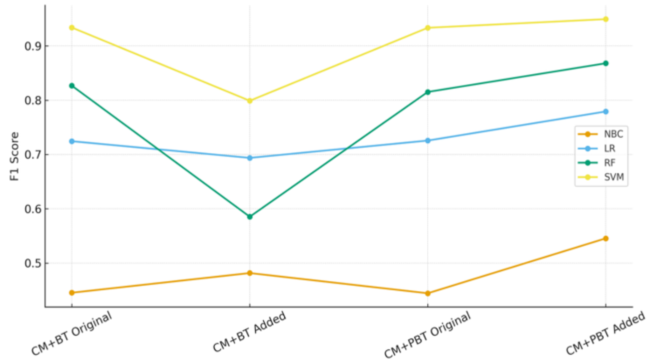


Figure 11. F1-Score comparison: Original vs. augmented test sets (CM+BT & CM+PBT)

As presented in Table 5 and Figure 11, CM+BT substantially reduces performance for several models, particularly RF and SVM, indicating that Code-Mixing introduces a distributional shift that certain architectures struggle to handle. In contrast, CM+PBT yields consistent improvements across all classifiers, suggesting that paraphrased sentences provide syntactically coherent variations that are more readily generalized by the models. These findings demonstrate that different augmentation strategies can exert distinct effects when incorporated into the test distribution and highlight the importance of robustness evaluation when augmentation is used in the training pipeline.

4.4 Analysis

The experimental results clearly demonstrate the significant impact of data augmentation in enhancing multilabel SDG classification performance. Substantial improvements in F1-

Score were observed, with SVM increasing from 0.6058 to 0.9349 and RF from 0.5363 to approximately 0.8285, indicating that augmenting underrepresented classes is essential for addressing skewed label distributions. Hamming Loss also decreased sharply, with SVM dropping from 0.1403 to 0.0276 and RF from 0.1573 to approximately 0.0673, confirming improved multilabel prediction accuracy after augmentation. These findings are consistent with the broader literature on long-tailed multilabel classification. Huang et al. [19] introduced distribution-balanced loss functions that mitigate class imbalance while preserving label dependency, demonstrating that addressing imbalance at the loss level substantially improves classifier performance in long-tailed settings. On the augmentation front, Xiao et al. [24] proposed the PIRAN framework—Pairwise Instance Relation Augmentation Network—which enhances classification of tail labels by generating synthetic samples based on feature relationships in head-label pairs, yielding notable performance gains.

Our results further show that augmentation improves recall, nearly doubling it for SVM (from 0.4824 to approximately 0.8858) and RF (from 0.4071 to approximately 0.7226), thereby enhancing the detection of minority SDG labels that were previously underrepresented. Performance consistency across augmentation methods also confirms the effectiveness of semantic-preserving strategies, with Code-Mixing with Back Translation slightly favoring SVM and Simple Back Translation benefiting RF, while overall results remained robust across methods. These findings corroborate that data augmentation, together with distribution-aware loss strategies, provides an effective solution for alleviating class imbalance in multilabel classification tasks, particularly when employing traditional classifiers such as SVM and RF. Logistic Regression also demonstrated significant improvements, confirming that even simpler models can benefit from improved label representation, while Naïve Bayes, though showing the smallest gains, still improved—suggesting that lexical diversity from augmentation partially offsets the limitations of its independence assumption.

The superior performance of Code-Mixing with Back-Translation for SVM can be attributed to the model's sensitivity to margin expansion resulting from lexical diversity. SVM benefits from having semantically similar but lexically varied samples, which better shape hyperplane boundaries. Conversely, RF performs best with Simple Back-Translation because RF relies on decision splits that benefit more from structural consistency than lexical randomness. Paraphrasing introduces syntactic variation that can increase tree branching complexity, making RF slightly less stable across paraphrased samples.

5. CONCLUSIONS

This study evaluated multilabel classification of Indonesian SDG-related news articles using Naïve Bayes (NBC), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) with TF-IDF features and 5-Fold Cross Validation. From 4,195 original and 5,105 augmented articles, a dataset of 9,300 samples was constructed through four augmentation methods. The results demonstrate that augmentation had a significant impact on model performance. SVM and RF achieved the highest performance with F1-Scores exceeding 0.93 and low Hamming Loss, LR provided competitive results with greater computational efficiency, while NBC, although weaker, also improved after augmentation. Furthermore, recall for SVM and RF nearly doubled, confirming that augmentation enhanced the ability of models to detect minority SDG labels. Overall, augmentation proved to be an effective strategy in mitigating class imbalance and significantly improving multilabel SDG classification performance.

These findings establish augmentation as a practical and scalable strategy for SDG text classification in low-resource settings. By showing that traditional models can achieve competitive performance when combined with effective augmentation, this work offers a cost-efficient alternative to transformer-based architectures, enabling broader adoption of SDG monitoring tools. For future work, further research may explore transformer-based approaches to capture deeper semantic features and improve performance in more complex datasets while balancing computational cost. This direction may extend the applicability of the proposed framework while maintaining scalability for SDG monitoring. In addition, future research will include benchmarking against lightweight Transformer models—such as multilingual DistilBERT—to provide a more comprehensive comparison and further validate the positioning of augmented traditional classifiers in low-resource scenarios.

ACKNOWLEDGMENT

This work is supported by the Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Schema C Research Grant No. 4488/UN1/FMIPA.1.3/KP/PT.01.03/2025.

REFERENCES

[1] Pawar, S., Kudage, O., Dalavi, S., Baraskar, M., Vanave, M., Bhosal, S., Jadhav, S.S. (2024). Text classification for news article. International Journal of Research in Applied Science and Engineering Technology, 12(5): 4464-4467. https://doi.org/10.22214/IJRASET.2024.62610

[2] Fitri, H., Widyawan, W., Soesanti, I. (2021). Topic modeling in the news document on Sustainable Development Goals. International Journal of Information Technology and Electrical Engineering, 5(3): 82-89. https://doi.org/10.22146/IJITEE.67467

[3] Morales-Hernández, R.C., Juagüey, J.G., Becerra-Alonso, D. (2022). A comparison of multi-label text classification models in research articles labeled with sustainable development goals. IEEE Access, 10: 123534-123548. https://doi.org/10.1109/ACCESS.2022.3223094

[4] Yohannes, H.M., Amagasa, T. (2022). A scheme for news article classification in a low-resource language. In Information Integration and Web Intelligence. iiWAS 2022. Lecture Notes in Computer Science, pp. 519-530. https://doi.org/10.1007/978-3-031-21047-1_47

[5] Morales-Hernández, R.C., Becerra-Alonso, D., Vivas, E.R., Gutiérrez, J. (2022). Comparison between SVM and DistilBERT for multi-label text classification of scientific articles related to the sustainable development goals. In Advances in Computational Intelligence.

MICAI 2022. Lecture Notes in Computer Science(), pp. 57-68. https://doi.org/10.1007/978-3-031-19496-2_5

[6] Pant, P., Sabitha, A.S., Choudhury, T., Dhingra, P. (2019). Multi-label Classification Trending Challenges and Approaches. Springer, Singapore, pp. 433-444. https://doi.org/10.1007/978-981-13-2285-3_51

[7] Liu, J.Z., Chang, W.C., Wu, Y.X., Yang, Y.M. (2017). Deep learning for extreme multi-label text classification. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, pp. 115-124. https://doi.org/10.1145/3077136.3080834

[8] Skrynnyk, M., Disassa, G., Krachkov, A., DeVera, J. (2024). SDGi corpus: A comprehensive multilingual dataset for sustainable development goals research. CEUR Workshop Proceedings, 3105: 23-34.

[9] Chernyshova, G., Taran, E., Firsova, A., Vavilina, A. (2025). Monitoring of sustainable development trends: Text mining in regional media. Sustainability, 17(7): 3122. https://doi.org/10.3390/SU17073122

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

[11] Pavlyshenko, B., Stasiuk, M. (2024). Data augmentation in text classification with multiple categories. Electronics and Information Technologies, (25): 67-80. https://doi.org/10.30970/ELI.25.6

[12] Freitas, L.J.G., Rodrigues, T., Rodrigues, G., Edokawa, P., Farias, A. (2024). Text clustering applied to data augmentation in legal contexts. arXiv preprint arXiv:2404.08683. https://doi.org/10.48550/arXiv.2404.08683

[13] Yao, R., Tian, M.L., Lei, C.U., Chiu, D.K.W. (2024). Assigning multiple labels of sustainable development goals to open educational resources for sustainability education. Education and Information Technologies, 29: 18477-18499. https://doi.org/10.1007/s10639-024-12566-6

[14] Fernández-González, D., Gómez-Rodríguez, C. (2023). Discontinuous grammar as a foreign language. Neurocomputing, 524: 43-58. https://doi.org/10.1016/J.NEUCOM.2022.12.045

[15] Ahmed, A.S., Haddad, A.A.A., Hameed, R.S., Taha, M.S. (2025). An accurate model for text document classification using machine learning techniques. Ingénierie des Systèmes d'Information, 30(4): 913-921. https://doi.org/10.18280/isi.300408

[16] Sutriawan, Rustad, S., Shidik, G.F., Pujiono. (2025). Performance evaluation of text embedding models for ambiguity classification in indonesian news corpus: A comparative study of TF-IDF, Word2Vec, FastText BERT, and GPT. Ingénierie des Systèmes d'Information, 30(6): 1469-1482. https://doi.org/10.18280/isi.300606

[17] Christ, K.L., Burritt, R.L. (2019). Implementation of sustainable development goals: The role for business academics. Australian Journal of Management, 44(4): 571-593. https://doi.org/10.1177/0312896219870575

[18] Guisiano, J.E., Chiky, R., De Mello, J. (2022). SDG-Meter: A deep learning based tool for automatic text classification of the Sustainable Development Goals. In Asian Conference on Intelligent Information and Database Systems, Ho Chi Minh City, Vietnam, pp. 259-271. https://doi.org/10.1007/978-3-031-21743-2_21

[19] Huang, Y., Giledereli, B., Köksal, A., Özgür, A., Ozkirimli, E. (2021). Balancing methods for multi-label text classification with long-tailed class distribution. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language, Punta Cana, Dominican Republic, pp. 8153-8161. https://doi.org/10.18653/v1/2021.emnlp-main.643

[20] Metlapalli, A.C., Muthusamy, T., Battula, B.P. (2020). Classification of social media text spam using VAE-CNN and LSTM model. Ingénierie des Systèmes d'Information, 25(6): 747-753. https://doi.org/10.18280/ISI.250605

[21] Herrouz, A., Djoudi, M., Degha, H.E., Boukanoun, B. (2023). An autonomous multi-agent system for customized scientific literature recommendation: A tool for researchers and students. Ingénierie des Systèmes d'Information, 28(4): 799-814. https://doi.org/10.18280/ISI.280401

[22] Tahir, M.A., Kittler, J., Bouridane, A. (2012). Multilabel classification using heterogeneous ensemble of multi-label classifiers. Pattern Recognition Letters, 33(5): 513-523. https://doi.org/10.1016/j.patrec.2011.10.019

[23] Tarekegn, A.N., Giacobini, M., Michalak, K. (2021). A review of methods for imbalanced multi-label classification. Pattern Recognition, 118: 107965. https://doi.org/10.1016/j.patcog.2021.107965

[24] Xiao, L., Xu, P.Y., Jing, L.P., Zhang, X.L. (2022). Pairwise instance relation augmentation for long-tailed multi-label text classification. arXiv preprint arXiv:2211.10685. https://doi.org/10.48550/arXiv.2211.10685

[25] Almamoori, A.A., Bhaya, W.S. (2023). Hybrid deep learning approach utilizing RNN and LSTM for the detection of DDoS attacks within the Bitcoin ecosystem. Ingénierie des Systèmes d'Information, 28(4): 931-937. https://doi.org/10.18280/ISI.280413

[26] Zheng, D., Kong, T., Jing, Y., Wang, J.A., Wang, X.J. (2022). Towards unifying reference expression generation and comprehension. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language, Abu Dhabi, United Arab Emirates, pp. 6598-6611. https://doi.org/10.18653/v1/2022.emnlp-main.442

[27] Chai, C.P. (2023). Comparison of text preprocessing methods. Natural Language Engineering, 29(3): 509-553. https://doi.org/10.1017/S1351324922000213

[28] Aliero, A.A., Adebayo, B.S., Aliyu, H.O., Tafida, A.G., Kangiwa, B.U., Dankolo, N.M. (2023). Systematic review on text normalization techniques and its approach to non-standard words. International Journal of Computer Applications, 185(33): 44-55. https://doi.org/10.5120/IJCA2023923106

[29] Nesca, M., Katz, A., Leung, C., Lix, L. (2022). A scoping review of preprocessing methods for unstructured text data to assess data quality. International Journal of Population Data Science, 7(1). https://doi.org/10.23889/ijpds.v7i1.1757

[30] Das, M., Selvakumar, K., Alphonse, P.J.A. (2023). A comparative study on TF-IDF feature weighting method and its analysis using unstructured dataset. arXiv preprint arXiv:2308.04037. https://doi.org/10.48550/arXiv.2308.04037

[31] Setiawan, Y., Maulidevi, N.U., Surendro, K. (2024). The

optimization of n-Gram feature extraction based on term occurrence for cyberbullying classification. Data Science Journal, 23(1): 31. https://doi.org/10.5334/DSJ-2024-031

[32] Naulak, C., Jindal, P. (2022). A comparative study of Naive Bayes classifiers with improved technique on text classification. TechRxiv. https://doi.org/10.36227/techrxiv.19918360.v1

[33] Wu, X., Gao, Y.C., Jiao, D. (2019). Multi-label classification based on random forest algorithm for non-intrusive load monitoring system. Processes, 7(6): 337. https://doi.org/10.3390/PR7060337

[34] Rainio, O., Teuho, J., Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14: 6086. https://doi.org/10.1038/S41598-024-56706-X

[35] Kadam, K., Peerzada, N., Karbhal, R., Sawant, S., Valadi, J., Kulkarni-Kale, U. (2021). Antibody Class(es) Predictor for Epitopes (AbCPE): A multi-label classification algorithm. Frontiers in Bioinformatics, 1: 709951. https://doi.org/10.3389/FBINF.2021.709951

[36] Joe, H., Kim, H.G. (2024). Multi-label classification with XGBoost for metabolic pathway prediction. BMC Bioinformatics, 25: 52. https://doi.org/10.1186/S12859-024-05666-0