Education and Information Technologies (2024) 29:18477–18499
https://doi.org/10.1007/s10639-024-12566-6

Assigning multiple labels of sustainable development goals to open educational resources for sustainability education

Rui Yao[1] · Meilin Tian[1] · Chi-Un Lei[1] · Dickson K. W. Chiu[1]

Received: 3 July 2023 / Accepted: 14 February 2024 / Published online: 9 March 2024

Abstract

Sustainable Development Goals (SDG) 4.7 aims to ensure learners acquire the knowledge and skills for promoting sustainable development by 2030. Yet, Open Educational Resources (OERs) that connect the public with SDGs are currently limitedly assigned and insufficient to promote SDG and sustainability education to support the achievement of SDG 4.7 and other SDGs by 2030, indicating a need for automatic classification of SDG-related OERs. However, most existing labeling systems can not support multiple labeling, tend to generate a large number of false positives, and have poor transferability within the OER domain. This research proposes a method to automatically assign SDGs based on AutoGluon, a machine-learning framework with powerful predictive capabilities, to allow multiple SDGs to be assigned to each OER. In the proposed framework, challenges of category imbalance and limited data availability are addressed, enhancing the precision and applicability of SDG integration in educational resources. To validate the transferability of model knowledge within the OER corpus, we used 900 lecture video descriptions from SDG Academy, forming the foundation for comparing our framework with existing labeling systems. According to the experiment results, our model demonstrates outstanding merits across various metrics, including precision, recall, F1, ACC, AUC, and AP.

Keywords Sustainable development goals · AutoGluon · Classification · Open educational resources · Machine learning

✉ Chi-Un Lei
culei@hku.hk

Rui Yao
yaorui@connect.hku.hk

Meilin Tian
meilin98@connect.hku.hk

Dickson K. W. Chiu
dicksonchiu@ieee.org

[1] The University of Hong Kong, Pokfulam, Hong Kong

1 Introduction

Aiming for our planet's healthy and productive future, governments, businesses, and other stakeholders need a long-term roadmap to ensure a peaceful, prosperous, sustainable, and habitable earth for all. Building on the achievements of the United Nations (UN) Millennium Development Goals, 17 Sustainable Development Goals (SDGs) were proposed in 2015 and adopted by all UN member states (Colglazier, 2015). These goals provided a roadmap to stimulate global economic and social development collaborations among governments and other stakeholders in public-private partnerships (Cf, 2015; Jomo et al., 2016). Furthermore, educating the public about the importance of SDGs is crucial. SDG 4.7 aims to ensure learners acquire the knowledge and skills needed to promote sustainable development by 2030 (Cf, 2015). Higher literacy and familiarity with SDGs help people be better prepared to collaborate in promoting global economic and social development through SDGs. As a result, various global SDG education initiatives have been proposed, including the Global Schools Program (Landorf, 2021) and the Global Education Innovation Initiative (Reimers & Chung, 2019). Besides, different research teams conducted reviews of the literature on SDG education. For example, Chiba et al. (2021) identified what is missing in the literature to understand effective curriculum development and implementation of SDG 4.7. Serafini et al. (2022) identified the main barriers that hinder the integration of university education with sustainable development guidelines, such as the alignment of course syllabi with SDGs and the alignment of courses with the external demands of SDGs. All these studies pointed out that more educational resources assigned to relevant SDGs are needed.

Meanwhile, Open Educational Resources (OERs) (Hannon et al., 2014; Sandanayake, 2019; Stagg, 2014) provide general accessibility for SDG education. UN's document "Recommendation on Open Educational Resources" advocates OERs to help all Member States create inclusive knowledge societies to achieve the 2030 Agenda for Sustainable Development (UNESCO, 2019). Since OERs are freely available for use and sharing, they support equity by providing free access to knowledge for everyone. Moreover, teachers may adapt OERs according to the needs of students and local communities for easier understanding and better relevance. For example, Lane (2017) introduced the application of OERs to support environmental science education.

Assigning SDG labels to learning resources and OERs helps utilize them for SDG education (Jha et al., 2020). In particular, SDG labels provide a common language for educators to share resources and best teaching practices on sustainability education and address global challenges collaboratively. Furthermore, labeling OERs can inspire more stakeholders to engage with SDG issues, cultivate global citizens dedicated to sustainable development, and support the achievement of SDG 4.7 and other SDGs by 2030. Different classification or assignment schemes have been proposed. For example, the UK Open University has collected their OERs aligning with SDGs and listed them for educators to find SDG-related resources more effectively. Since 2016, the SDG Academy has offered over 1,800 free and open educational videos on sustainable development to enrich the field for the 2030 Agenda. In addition, the SDG Knowledge Hub[1], launched in 2016, contains over 9,000 published news articles with SDG labels. These materials can also be curated for educational activities.

However, compared to the vast number of unclassified OERs, the limited number of OERs with SDG labels still needs to be increased to accelerate SDG education, indicating a need for automatic classification of SDG−related OERs. This approach is also endorsed by the UN's "Recommendation on Open Educational Resources" (UNESCO, 2019), which suggests utilizing open−license tools to help ensure that OERs can be easily found and accessed.

Although efforts have been made in modeling SDG labeling systems, there still exists the following research gap:

1. Few scattered studies focus on SDG auto-classification (Lei, 2022; Lei et al., 2022; Wang et al., 2022), but they mainly cover specific non-open educational resources. The transferability of cross-domain trained models needs further validation, as potential domain shifts may impact model performance (Ma et al., 2023).
2. Existing OER labeling systems like Aurora-SDG (Vanderfeesten et al., 2022) and Monash University (Monash University, 2017) have notable deficiencies, including inefficient multi-output classification and the use of limited annotated data (Wulff et al., 2023). Besides, many of them are keyword mapping-based methods, which leads to a number of false positive classifications.

Identified by other studies (Otto & Kerres, 2022; Armitage et al., 2020; Pukelis et al., 2020; Schmidt & Vanderfeesten, 2021), the challenges of handling the above gaps can be concluded as follows:

1. Developing a robust multi-output classification system that can accurately assign multiple SDG labels to OERs.
2. Mitigating the prevalent category imbalance in OERs to enhance the accuracy and utility of SDG-focused education.

Overcoming these challenges is crucial for the precision of SDG labeling, thereby amplifying the impact and accessibility of SDG-centric education.

Thus, this research aims to automatically assign SDGs based on AutoGluon, an advanced machine-learning framework with powerful predictive capabilities. The method is validated with a dataset from the SDG Academy, a platform selected for its alignment with global sustainability objectives, the nuanced diversity of content, and its suitability for a domain-specific evaluation.

To conclude, our contributions are as follows:

[1] https://sdg.iisd.org/

1. The propsoed automatic framework for assigning SDGs to OERs outperforms existing benchmark methods, which is crucial for expanding global sustainable development education.
2. Our method allows for assigning multiple SDGs to OERs, demonstrating the interconnections between SDGs in education. In contrast, benchmark methods typically assign only one SDG to OERs.
3. We uniquely address the challenges of category imbalance and limited data availability, enhancing the precision and applicability of SDG integration in educational resources.
4. We evaluate the proposed framework using OERs from the SDG Academy and compare it with other popular benchmark methods. Using a wide range of metrics, we are evaluating the transferability of existing labeling systems to the OER domain for the first time. Results demonstrate our method's outstanding merits.

2 Literature review

2.1 AutoGluon-based machine learning

AutoGluon Tabular (Erickson et al., 2020) is a high-performance machine-learning framework extensively validated in various real-world competition tasks. Its core mechanism uses an improved multi-layer stacking approach, aggregating multiple base classifiers for ensemble learning. Due to its outstanding performance and user-friendly design, researchers from various domains have explored the possibility of using this framework for machine learning modeling. For example, Liu et al. used the global burden of accidental carbon monoxide poisoning (ACOP) and the World Bank database to predict the epidemiology of ACOP using AutoGluon (Liu et al., 2022). In addition, Seo et al. (2021) evaluated various machine learning and deep learning classification models, including AutoGluon, for classifying walking assistive devices for cerebrovascular accident (CVA) patients. Besides, Blohm et al. (2020) compared four machine-learning tools across 13 text classification datasets and found AutoGluon performed the best in 7 out of 13 tasks.

Figure 1 illustrates the model architecture in AutoGluon for text classification. The framework generates features for textual data by employing n-grams, specifically unigrams, bigrams, and trigrams. The framework also calculates statistical attributes such as word count, character count, and the proportion of uppercase letters to enhance prediction accuracy. Another motivation for explicitly modeling the relationship between such lexical features and the corresponding labels is that some features, such as text length, significantly impact the performance of detections (Wulff et al., 2023). Furthermore, the framework has established a maximum limit of 10,000 features and eliminated the least frequent tokens to constrain the number of features, thereby streamlining the model and enhancing its efficiency. In correspondence with the feature extraction of tabular data, the framework employs LightGBM (Ke et al., 2017), CatBoost (Dorogush et al., 2018), XGBoost (Chen & Guestrin, 2016), and Vowpal-Wabbitv (Langford et al., 2007) as base classifiers for the learning process. Furthermore, advanced feature extraction through transformer models

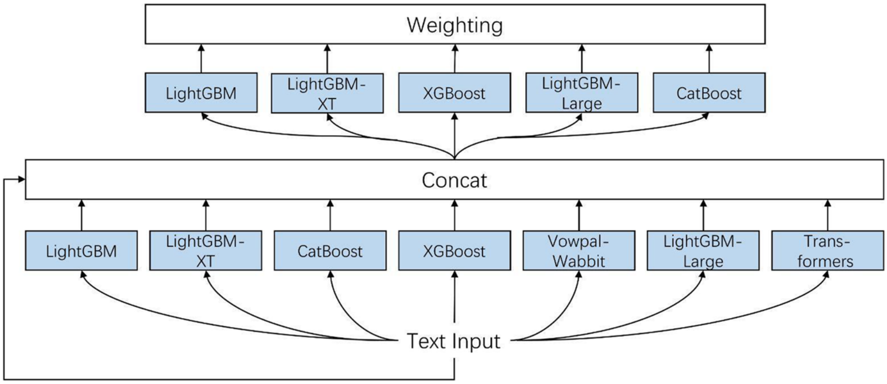

Fig. 1 Model architecture

(Vaswani et al., 2017) with the pre-trained and fine-tuning paradigm is incorporated into the overall ensemble framework (Shi et al., 2021), allowing us to benefit from both paradigms with valuable diversity for ensemble learning simultaneously.

Several features of the model architecture are as follows:

1. The original features and the output of previous layers' base learners are concatenated as input for subsequent layers to achieve multi-layer stacking. This design improves the original stacking scheme, allowing higher-level stackers to access initial data while aggregating the output of lower-level classifiers, thus reducing model bias.
2. The repetitive k-fold bagging method was proposed, as shown in Fig. 2. Each class of base learners has k copies (with k = 3 as an example in the figure), each of which is trained and evaluated on different data blocks, and this process can be repeated n times. The framework did not implement repetition in our algorithm's training and evaluation process on different data blocks. The algorithm stops automatically after five cycles without improvement, ensuring efficiency. This method allows higher-level models to be trained only on the predictions of lower-level models, thus mitigating the risk of overfitting due to different levels repeatedly learning the same data, ultimately reducing the model's variance.
3. The framework selects models based on their evaluation performance and inference time on hold-out datasets in the proposed approach. After the assessment, the framework decides to utilize all the base learners collectively, as they demonstrated satisfactory performance within the ensemble framework. However, the framework excludes the default option, KNN, as it is unsuited for large-scale NLP tasks. Ultimately, the output of the ensemble classifier is a linearly weighted combination of the final layer model outputs, with the weights obtained through learning.

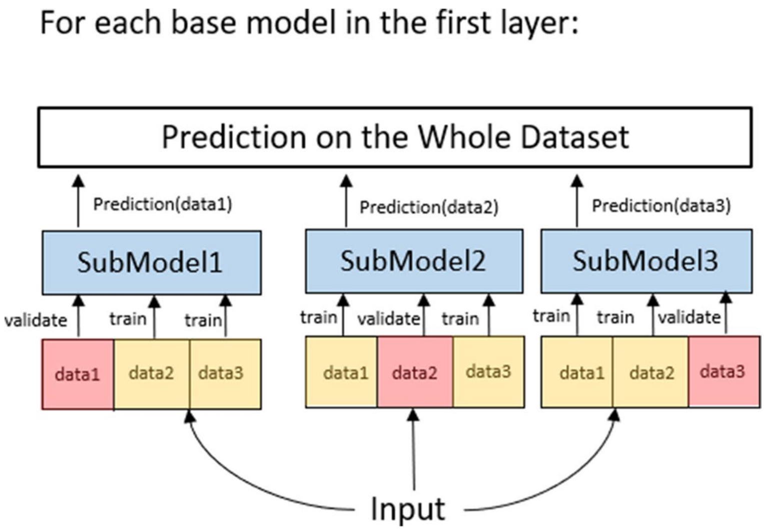

Fig. 2 K-fold cross bagging

2.2 SDG classification of educational resources

Assigning high-quality SDG labels to educational resources is a challenging task. The complex interrelationships between SDGs (Bowen et al., 2017) and the vast number of materials to be labeled make the manual assignment or classification of accurate labels difficult. As a result, various organizations are working toward achieving SDG classification to find a "common language" in this field. Researchers are increasingly exploring machine-learning techniques as alternative solutions (Kashnitsky et al., 2022; Pukelis et al., 2022; Vanderfeesten et al., 2020). The main development directions are text classification and SDG-related academic research keyword queries.

Vanderfeesten et al. (2020) introduced the Aurora dataset to validate their classification system, employing a survey-based approach where 244 respondents assessed the relevance of 100 research papers randomly selected by the Aurora system from a pool of SDG-relevant papers. Training data was annotated using keyword queries in their subsequent work (Vanderfeesten et al., 2022). Furthermore, the Aurora European Universities Alliance, the University of Southern Denmark, the University of Auckland, and Elsevier are collaborating to map research articles to SDGs (Kashnitsky et al., 2022). They adopted the Boolean query method used by Times Higher Education in Social Impact Rankings, using academic abstracts as the corpus. Further, they fine-tuned the pre-trained mBERT model based on the generated labels as an extension of the pre-defined dictionary. However, more than the Boolean query-based method is required to model the rich semantic information in the text and may result in many false-positive predictions.

The OSDG community is building machine-learning models and datsets, making significant contributions to SDG classification (Pukelis et al., 2020). University

College London (UCL) and York University in Canada have used the OSDG framework for the curriculum analysis. Lei (2022); Lei et al. (2022) used the OSDG dataset and logistic regression to analyze how SDG knowledge is taught and assessed in public K-12 curricula and university general education. Wang et al. (2022) also used the OSDG dataset and logistic regression to analyze the teaching of SDGs in Coursera MOOCs, providing an overview of the different proportions of SDGs in Coursera MOOCs from various universities. However, the performance of logistic regression is unsatisfactory, with low F1 scores.

The SDG Knowledge Hub data curated by Wul and Meier (2023) encompasses news articles from the SDG Knowledge Hub website, managed by the International Institute for Sustainable Development (IISD). Comprising 9,172 articles, the dataset includes assigned SDG labels by subject experts, providing a valuable resource for studying SDG-related content and classification systems.

In conclusion, these frameworks, utilizing various approaches from single key matches to complex machine learning models, represent the diversity in automated SDG classification. Benchmarks like SDSN, SDGO, SIRIS, Elsevier, and Auckland systems illustrate the use of Boolean operations in their methodologies. Systems like Aurora (Vanderfeesten et al., 2020) employ a survey-based approach where 244 respondents assessed the relevance of 100 research papers randomly selected by the Aurora system from a pool of SDG-relevant papers. Training data was annotated using keyword queries in their subsequent work (Vanderfeesten et al., 2022). These selected systems demonstrate the range of methodologies in query system design pertinent to our study's focus on SDG research. The OSDG.ai system (Pukelis et al., 2022), however, showcases the application of machine learning in SDG classification utilizing datasets from the OSDG Community (OSDG et al., 2022).

2.3 SDG classification procedure

In our proposed classification framework, the OSDG dataset is used as the training dataset, and AutoGluon is employed as the classification framework. Furthermore, optimization with domain knowledge is used to tackle data imbalance issues and multi-label learning challenges. Technical details of the framework configuration are discussed in this Section.

2.4 Training dataset: OSDG dataset

This study uses the OSDG Community Dataset (OSDG-CD) as the training dataset because it provides an annotated dataset and an open-source SDG text classification API supporting multiple languages (OSDG, 2022). The dataset comprises 3 to 6 SDG-related sentences collected from UN-related libraries, like SDG-Pathfinder[2] and SDG Academy Library [3]. Over 1,000 volunteers participated in the

[2] https://sdg-pathfinder.org/

[3] https://sdgacademy.org/sdgacademy-library/

crowdsourced annotation work on the OSDG platform. The OSDG dataset is unique because it only applies a single SDG to each sample. Typically, each text may be related to multiple SDGs, as these texts are excerpted from United Nations documents. However, the interdependencies between different SDGs still need to be clarified and subjective. Therefore, using a dataset with single−labeled SDG can sometimes result in models with higher precision and accuracy.

We use the OSDG-CD 2022.10 version of the dataset to train the model. This CSV dataset file contains 37,575 text excerpts with 16 SDG labels (excluding SDG 17). Each sample provides a consistency score calculated from the annotation results of different volunteers. We selected samples with consistency greater than or equal to 0.6, retaining 21,758 samples. This dataset represents the most extensive collection of data that meets our highest requirements for annotation quality.

Figure 3 shows the distribution of labels for different SDGs in the data. Although the dataset has the best annotation quality known to us, imbalanced sampling is a drawback, as SDG10, SDG12, and SDG15 samples are scarce. This is inevitable considering the varying degrees of involvement in different SDGs in global UN documents. Thus, we train separate binary classifiers for each SDG and use Bagging sampling methods to mitigate data imbalance issues.

2.5 Classification through AutoGluon

In the SDG classification, several primary challenges could be often identified:

1. As in traditional methods, multi-output classification problems, i.e., allocating zero to multiple SDG category labels for each sample instead of assigning a single output value for each sample.
2. Category imbalance problems, i.e., the distribution of sample quantities across different categories, are significantly imbalanced, as shown in Fig. 3. This may pose challenges to the accurate identification of minority classes.
3. Lack of high-quality annotated data. After filtering samples, our training data contains only 22,096 annotated data. For a 16-class multi-label text classification

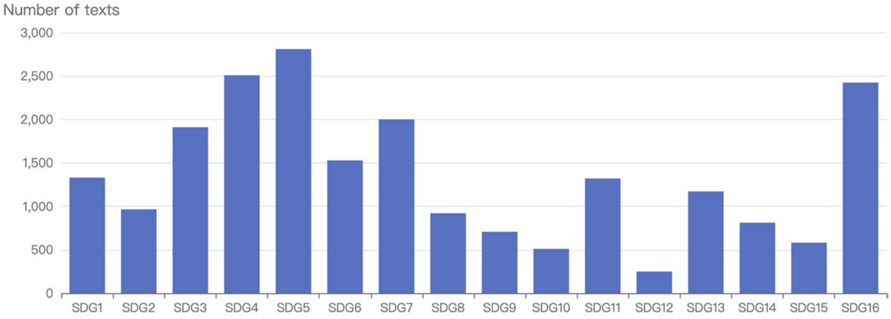

Fig. 3 Distribution of the samples with assigned SDGs in the OSDG database

problem, obtaining a well-trained model based solely on existing data is challenging.

Combining our domain knowledge and understanding of AutoGluon's underlying technology, the proposed classification framework has been configured as follows:

1. Following AutoGluon's implementation for multi-output problems, an independent binary classifier is trained for each label category. This results in 16 separate binary classifiers, each responsible for determining whether the input sentence corresponds to the label it is assigned to predict.
2. The following strategies were employed to alleviate data imbalance issues:

    a. Each training layer uses eight Bagging folds, as their documentation recommends (AutoGluon, 2022). This decision balances performance and computational cost, aligning with the model's robust nature, which is less sensitive to hyperparameter tuning. When combined with undersampling methods, Bagging ensemble methods can better address imbalanced problems (Roshan & Asadi, 2020). Each Bagging data fold uses a random undersampling subset of the majority class and standard bootstrapped samples of the minority class. This mitigates the imbalance problem at the level of base classifiers and prevents the drawback of traditional undersampling schemes that may discard valuable data.
    b. A comprehensive range of evaluation metrics is used to understand our classifiers' testing-time performance on this imbalanced dataset, including precision, recall, F1, Accuracy, AUC (Area under the ROC Curve), and AP (Average Precision). ROC and PR curves are plotted and categorized according to different SDG categories, providing an intuitive comparison between our classifiers and benchmark methods. Evaluation metrics like AUC and Average Precision (AP) were calculated directly from the models' probability distributions, bypassing the need for threshold-based classification. This dual approach in evaluation allows for a comprehensive and fair assessment of each model's capabilities.
    c. Higher weights are assigned to instances of minority classes.

3. To address the issue of insufficient data quantity, pre-trained transformers are used as one of our base learners, allowing us to utilize information collected from external texts. Additionally, fine-tuning the model to fit our problem domain better enables more effective feature extraction (Shi et al., 2021).
4. AutoGluon's design inherently provides stable performance across various tasks with little to no manual adjustment of parameters. This characteristic is particularly valuable in our study, which focuses on developing a reliable and effective SDG classification model without extensive hyperparameter optimization complexities.

We conducted model training on the Google Colab Pro + program using a Tesla T4 GPU, running on Ubuntu 20.04 with approximately 12.7GB of available RAM, 16GB of video memory, and AutoGluon version 0.6.0. In practice, we did not perform conventional preprocessing on the raw data, such as case conversion, punctuation removal, and stopword removal, to leave more suitable statistical and linguistic features for feature extractors, allowing the algorithm to determine signals important for prediction.

3 Performance evaluation

3.1 Testing dataset: Metadata from SDG Academy

For the testing dataset, given our focus on the OER domain, we employed annotated data collected from the SDG Academy. The SDG Academy is a program of the Sustainable Development Solutions Network, a global initiative for the United Nations supporting the Sustainable Development Goals. We utilized data from the SDG Academy library, consisting of over 1800 lecture video descriptions and metadata. Each video page displays related topics, such as the associated SDGs. Furthermore, every video is associated with at least one SDG and a maximum of four SDGs. We utilize the descriptions of each lecture as features and the SDG labels as targets for our validation dataset.

This dataset was specifically chosen for its relevance to global sustainability objectives, the diversity and complexity of its content, and its suitability for a practical, domain-specific evaluation of classification frameworks. The dataset encapsulated a diverse array of SDG-related topics and served as the dataset for gauging the performance of the various frameworks. The descriptions from these lecture videos offer a text-rich dataset, challenging the benchmarks with natural language complexities and nuances reflective of the multifaceted nature of SDG-related content. This approach ensures that the performance assessment of the classification frameworks is rooted in a practical and domain-specific context, providing insights into their operational efficacy.

In the preprocessing stage, we removed samples with languages other than English, such as Spain. Subsequently, we selected texts with video descriptions longer than 50 words and eliminated those with similar contexts to avoid repetition within the dataset. The final dataset comprises 900 samples with multiple SDG targets and detailed descriptions for classification purposes.

Figure 4 illustrates the label distribution in the SDG Academy for evaluations. Similar to the OSDG community dataset, the distribution of videos among the SDGs is uneven. For example, more OERs are related to SDG 13, highlighting the importance and prevalence of addressing climate change. Table 1 displays the distribution of video objects in the SDG Academy allocated to a specific number of SDGs, indicating more resources allocated to a single SDG. However, over 40% of resources are assigned to multiple SDGs. For instance, there are resources classified with (i) SDGs 9 and 12 (50 videos) and (ii) SDGs 5, 10, and 16 (44 videos). This implies a simultaneous focus on promoting (i) industry, innovation, infrastructure, consumption, and production and (ii) gender equality, reduced inequalities, and justice.

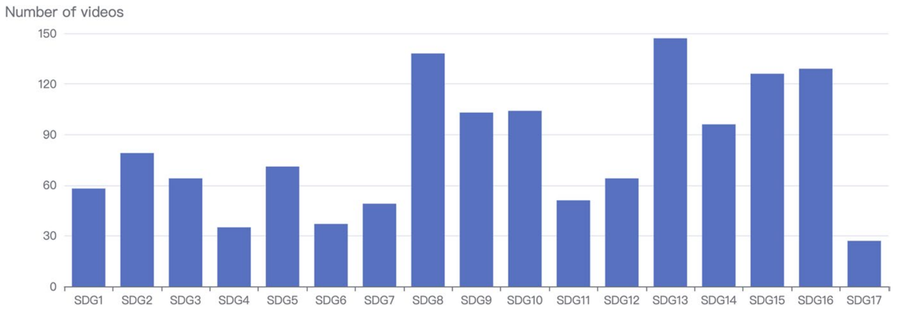

Fig. 4 Distribution of the samples with assigned SDGs in the SDG Academy

Table 1 Distribution of the video objects with the number of SDGs assigned in the SDG Academy

Number of SDGs assigned in the course	Number of courses
1	530
2	272
3	88
4	10

Furthermore, the video "Sustainable Food and Land Use" discusses CO_2 emissions caused by agriculture and the impact of agriculture and food production on SDGs 2, 6, 14, and 15, demonstrating the interconnectivity among SDGs within the OER and the need for assigning multiple SDGs to educational resources.

Testing the model with the SDG Academy dataset while using the knowledge derived from the OSDG dataset is challenging. From an algorithmic perspective, introducing multi-output predictions is necessary; from a text classification standpoint, the OSDG and SDG Academy have vastly different classification criteria. Since the OSDG dataset consists of volunteer-annotated data, crowdsourcing annotation consistency with multiple labels is challenging. On the other hand, the lectures in the SDG Academy may encompass multiple SDG themes. Unlike other SDG datasets derived from the UN's articles and books, the SDG Academy's text data consists of lecture video introductions rather than descriptions of SDG indicators. Therefore, the SDG Academy dataset tests the transferability between different sources of SDG-related content and aids SDG classification in education, particularly course categorization. For models from different organizations, the SDG Academy serves as a fair comparison of our algorithm and other APIs using a new dataset.

3.2 Comparison of classification performance between frameworks

In this study, we compare our AutoGluon-based methodology against a broad spectrum of established frameworks to provide a comprehensive backdrop for our study. In selecting these frameworks for comparative analysis, we prioritized those most

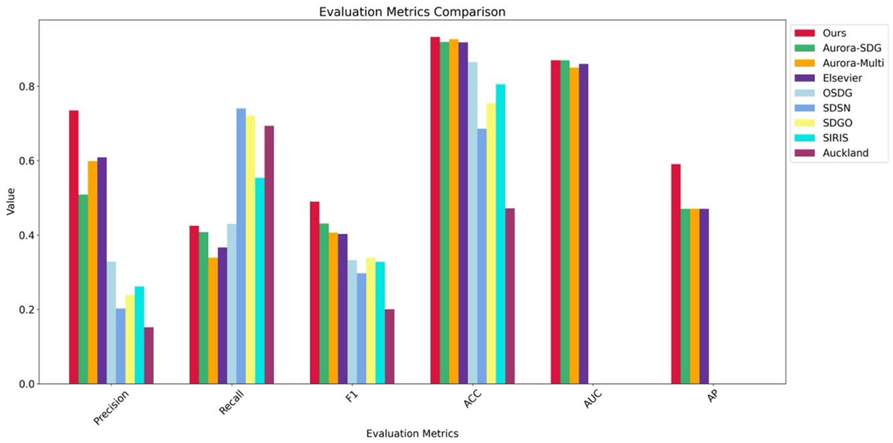

Fig. 5 Comparison of classification performance of the proposed framework and other existing frameworks

widely recognized for their relevance to our domain and methodological alignment with our objectives, including Aurora-SDG, Aurora-Multi-SDG[4], Elsevier−Multi−SDG[5], the OSDG model[6], SDSN, SDGO7[7], SIRIS[8], and Auckland[9]. When binarizing probability−based predictions into labels, we used a threshold of 0.98 for Aurora−SDG, as recommended in their documentation (Vanderfeesten et al., 2022). A threshold of 0.35 was used for Aurora−Multi−SDG and Elsevier−Multi−SDG to provide the optimal performance.

Figure 5 shows the algorithm’s performance. Focusing on evaluating imbalanced datasets, we calculated AUC and AP (metrics) and assessed the performance differences between our method and others across different SDG groupings, plotting ROC and PR curves (see Figs. 7 and 8 in the Appendix). Note that the OSDG, SDSN, SDGO, SIRIS, and Auckland provide binary predictions rather than probability-based predictions, making them unsuitable for AUC and AP score comparisons.

Our approach demonstrated competitive performance to existing benchmarks (in recall, ACC, and AUC scores) and maintained a clear advantage in other metrics (precision, F1, and AP). As a reference, the Aurora-SDG, SDSN, SDGO, SIRIS, and Auckland were characterized by high recall rates achieved by sacrificing precision. Such a tendency suggests these frameworks are inclined to overgeneralize, leading to a higher proportion of false positives and, thereby, less accurate classifiers. In the educational context, teachers and administrators prefer fewer false positives than false negatives. Aurora-Multi-SDG and

[4] https://aurora-sdg.labs.vu.nl/classifier/classify/aurora-multi-sdg.

[5] https://www.elsevier.com/about/partnerships/sdg-research-mapping-initiative

[6] https://osdg.ai/

[7] https://figshare.com/articles/dataset/SDG_ontology/11106113/1

[8] https://zenodo.org/records/3567769

[9] https://www.sdgmapping.auckland.ac.nz/

Elsevier-Multi-SDG also exhibited inferior overall performance, partly due to their ability to provide only probability output with a summation equal to 1 and assignment/classification of single labels for predictions.

Furthermore, Fig. 5 indicates our proposed framework tends to provide conservative estimates, achieving the highest precision and Accuracy scores, which is considered more objective in this domain than the recall rate. The result of the conservative estimation can be interpreted as a combined effect of various features. Firstly, our Bagging sampling and Stacking ensemble strategies allow our final estimates to be formed based on a weighted ensemble of decisions from different base classifiers. As each base classifier represents a separate observation of different training subsets and pattern recognition consistent with individualized characteristics, inferences that achieve consensus across all classifiers can be considered close to the actual situation underlying the respective sample, thus a primary advantage of applying ensemble predictions. Moreover, the OSDG training set used for training underwent strict community review; high-consistency filtering can be considered to result in high-quality and low-confusion annotations, making it difficult for intrinsically ambiguous or deviating test samples to achieve consensus from different classifiers, thus tending to reject the corresponding predictions in our labeling system.

Figure 6 displays the performance of different algorithms at different thresholds. Higher thresholds generally provide classification results with higher precision and lower recall rates. Balancing precision and recall is challenging due to the ambiguity of multi-label classification. Therefore, the F1 score can be selected as a criterion for weighing the pros and cons, as both low precision and recall rates would affect the overall F1 score. However, different choices are feasible depending on the type of OERs. A high-precision classification method with a higher threshold can be chosen for an OER explicitly involving SDGs, while a high-recall classification method can be selected with a lower threshold for an OER only implicitly related to SDGs.

3.3 OER multi-SDG classification: A case study

In practice, training with multi-label datasets may not be more accurate than using single-label datasets, as the intersections between SDGs can be challenging to define through explicit criteria. For instance, as shown in Table 2, the OER "Introduction to the Ocean & Climate" discusses how the ocean drives Earth's climate, which undoubtedly covers two SDGs (SDG13 and SDG14). The proposed method has successfully categorized these SDGs. On the other hand, most benchmark methods can only classify one SDG. Furthermore, Table 3 also presents another example of assigning multiple SDGs to an OER. These examples demonstrate the ability of the suggested framework to allocate multiple SDGs to OERs.

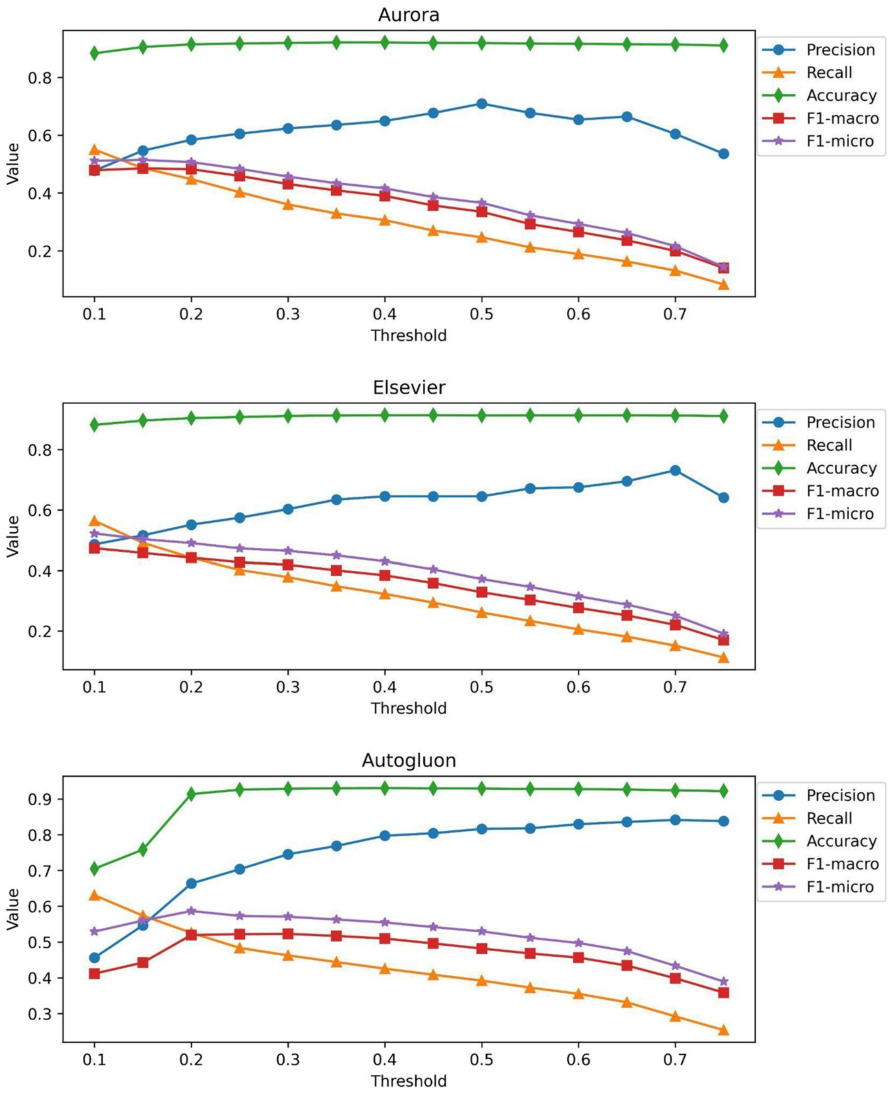

Fig. 6 Performance of different algorithms with different threshold values

Table 2 Multi-SDG classifications: Introduction to the ocean & climate

Classification framework	Assigned SDGs
SDG Academy (the official assignment)	13, 14
Our proposed framework	13, 14
OSDG	11, 13, 15
Aurora-SDG	14
Aurora-Multi-SDG	14
Elsevier-Multi-SDG	13

Table 3 Multi-SDG classifications: Climate change adaptation and mitigation

Classification framework	Assigned SDGs
SDG Academy (the official assignment)	2, 13
Our proposed framework	2, 13
OSDG	1, 3, 9
Aurora-SDG	2
Aurora-Multi-SDG	N/A
Elsevier-Multi-SDG	N/A

4 Discussion

4.1 Difference between the proposed framework and other frameworks using ensemble techniques

Wulff et al. (2023) systematically evaluated seven existing SDG assignment frameworks using different metrics and expert-annotated datasets covering various text sources. Their results show that different systems exhibit different biases for different SDG categories, and existing systems are prone to false positives, meaning that they assign SDG labels to samples that do not belong to any SDG category. Finally, they discovered that text length significantly affected the performance of current labeling systems and proposed an ensemble learning approach that leverages different labeling systems to provide joint support for SDG label assignment. The ensemble model demonstrated improved accuracy while reducing the number of false positives. Overall, that research emphasizes the need to avoid treating labeling systems as interchangeable and suggests that ensemble models may be a viable alternative to existing SDG labeling systems. Unlike their work, our approach does not simply integrate existing systems but focuses on early training. The OSDG dataset trains entirely new models in our proposed framework, incorporating various best practices to provide a better labeling system in this domain.

Wulff et al. also discussed the issue of domain shift, which means if samples not following the same distribution are used for testing, the model may produce inaccurate results. But in our framework, the outputs of different classifiers are integrated into a unified decision, and the agreement or disagreement between classifiers is directly modeled as the probability of the ensemble classifier's output. Compared to the results generated by individual classifiers, the ensemble approach has better robustness in decision-making for out-of-distribution samples.

4.2 Limitations in the Aurora-SDG framework

The Aurora-SDG framework (Vanderfeesten et al., 2022) fine-tunes a pre-trained mBERT model based on academic publication abstracts, trains separate binary classifiers for each SDG to achieve multi-label classification, and attempts to support multiple languages. However, there are several drawbacks to this approach:

1. The training data is annotated through keyword queries rather than by human experts on a case-by-case basis. This query method may not be sufficient to model the rich semantic information present in the text. For example, when identifying materials related to SDG Goal 3, "Good Health and Well-being," a commercial advertisement for medical beauty treatments would be labeled with a 100% probability because it contains the term "healthy skin." This idea can also be confirmed by the authors' experimental results in their article: They used an expert-annotated dataset for testing, and out of the 97 manually labeled English publications, the method produced 258 predictions with a very high threshold (≥ 0.99), of which only 46 were correct, resulting in an accuracy of about 17.8% and a recall rate of about 47.42%. This deficiency is prevalent in all keyword-based annotation systems, as illustrated in our experimental results in Fig. 5. A considerable number of models exhibit high recall and low precision, impacting the effectiveness of the models.
2. The authors selected all positive samples from the training set for each SDG classifier and randomly sampled an equal number of other categories as negative samples. Although this approach can address the data imbalance, the simple downsampling method can cause valuable information loss.
3. Some SDG categories have fewer positive samples, and fine-tuning large models on small datasets may pose a potential risk of overfitting or biasing.

5 Future works

Similar to other research on classifying SDG for educational materials, the proposed framework can be applied to materials in K12, higher education, and MOOCs with better classification performance. The large-scale classification of educational resources can help stakeholders understand sustainability education from K12 to continuing education. The framework can also be used to classify SDGs taught across multiple courses, helping stakeholders understand the interconnections between courses and the areas of focus or learning gaps within the courses related to SDGs. Moreover, following the recent developments in the Aurora framework (Vanderfeesten et al., 2022), our framework can integrate BERT multilingual models to become an OER text classifier for other languages, facilitating the adoption of SDG education in local communities. Recognizing the potential breadth and depth that could be achieved by incorporating both the OSDG and SDG Knowledge Hub datasets, we suggest that future work should explore this avenue.

Furthermore, we propose that future research endeavors consider the following objectives:

1. Facilitating the assessment of the quality of current datasets and incorporating a universally applicable, high-quality, large-scale dataset in the SDG sphere or providing new datasets that would prove beneficial;
2. Examining the inductive potential of few-shot learning, transfer learning, and large language models in this area;

3. Exploring the implementation of reliable labeling systems within this domain. Existing SDG labeling systems may possess biases, potentially leading to inequitable investments in specific regions compared to others (Wulff et al., 2023). Consequently, establishing an interpretable and impartial SDG labeling system bears substantial importance;
4. Using social media to aid and promote education (Cheung et al., 2023; Jiang et al., 2023; Leung et al., 2022; Wang et al., 2021), especially concerning sustainability and environmental issues (Chung et al., 2020; Ho et al., 2023); and.
5. Exploring the interdisciplinary area of SDG education with digital literacy and competencies to ensure inclusive and quality education, such as the effectiveness of online learning, the importance of digital literacy in higher education, and improving parental digital literacy for safer online engagement (Oyewola et al., 2022; Tokovska et al., 2022; Nurhayati et al., 2022).

6 Conclusions

SDG text classification can help stakeholders find a "common language" to classify contributions towards achieving the SDGs. In this study, we aim to classify OERs so that more resources can be discovered and searched to expand sustainability education globally. Our proposed classification framework uses OSDG and AutoGluon as the training dataset and classification algorithm, respectively. The trained model is then evaluated using video metadata from the SDG Academy Library. The proposed training model is compared with eight benchmark frameworks using various metrics. Overall, our model demonstrates competitive performance across different metrics and allows assigning multiple SDGs to OERs, which is naturally common in SDG-related corpus. In addition, our work has provided new insights into this field by summarizing existing systems and datasets, introducing improved sampling methods and classification algorithms, and facilitating education on SDG themes.

A lower recall rate is the primary drawback of this study. Yet, low recall scores are quite common with existing labeling systems. No single approach has achieved an ideal recall rate, implying that the general lack of performance may be attributed to the complexity of the task or the small scale of existing annotated datasets.

This research substantially benefits sustainable education technology by providing an advanced, automatic framework for SDG classifications in OERs. Our innovative method, which allows for multiple SDG labels per resource, enriches educational content and promotes a more integrated understanding of sustainability issues. This advancement in resource classification methodology extends beyond OERs, contributing valuable insights to the broader field of educational resource management. Ultimately, our study propels forward the integration of sustainability into educational resources, marking a significant leap in both the technology and pedagogy of sustainability education.

Appendix

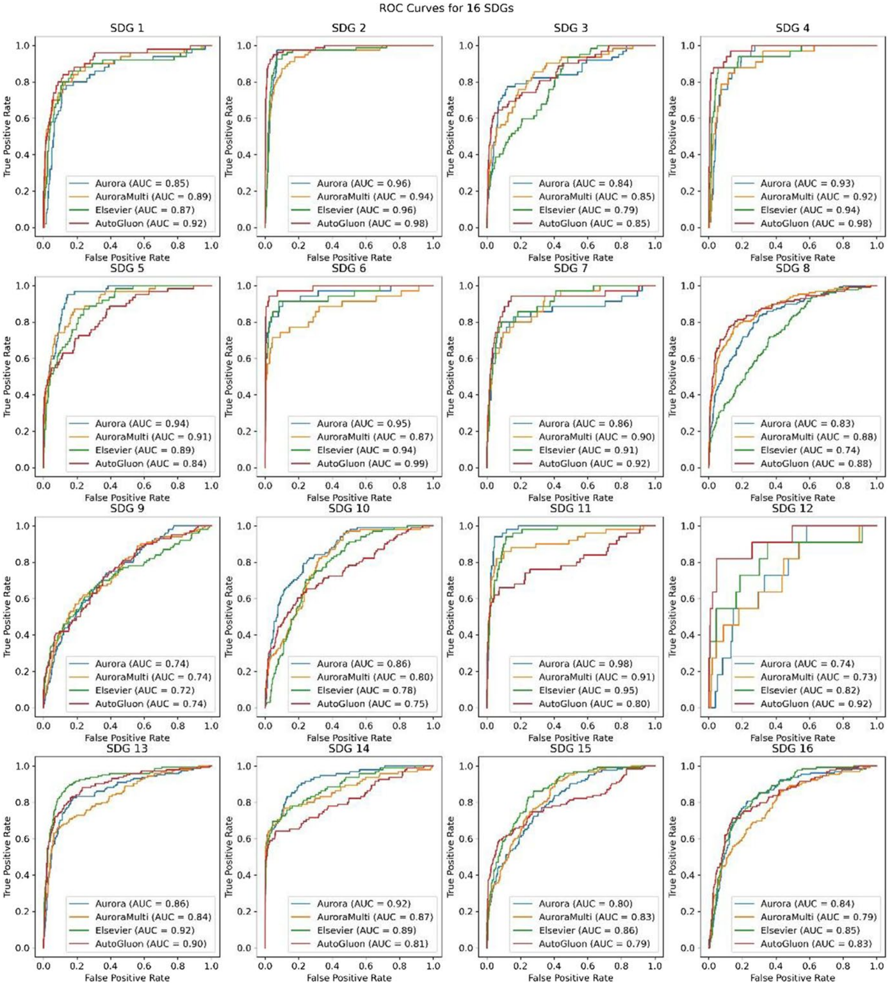

Fig. 7 ROC curves of different methods (AutoGluon: The proposed method)

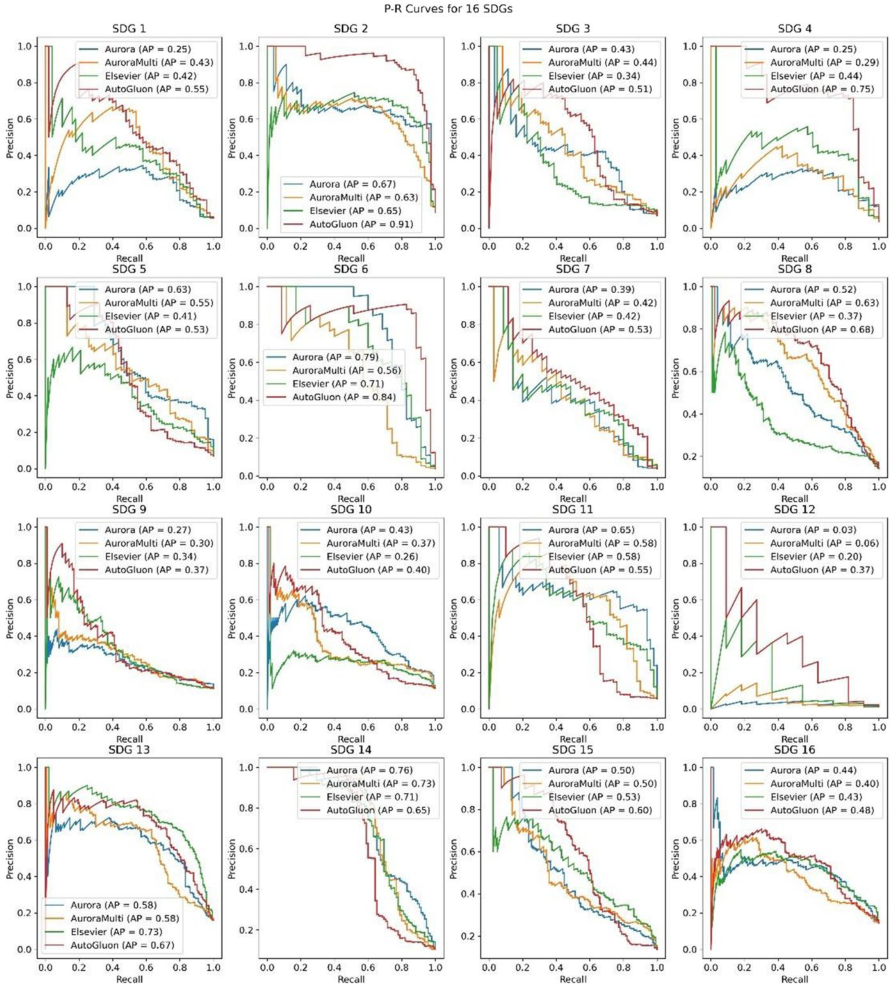

Fig. 8 PR curves of different methods (AutoGluon: The proposed method)

Acknowledgements We thank Mr. Hongren Wang implementing the SDG classification through existing classification models, and supporting the revision.

Authors' contributions Rui Yao trained and tested the SDG multi-labeled classification model with AutoGluon. Meilin Tian collected the SDG Academy dataset and calculated the performance of different models. Chi-Un Lei guided the project and was a major contributor to writing the manuscript. Dickson K.W. Chiu helped proofread and revise the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the Hong Kong University Grants Committee under the teaching development grant "Building a Collaborative Platform for Enhancing Artificial Intelligence" and "Applications in Education and Empowering University Teachers for the Next-Stage Virtual Teaching" on recruiting researchers conducting the study.

Data availability The OSDG datasets analysed during the current study are available in the OSDG repository https://doi.org/10.5281/zenodo.7136826. Other datasets generated during the current study are available in our GitHub repository,https://github.com/HKU-SDG-Classification/SDG-Curriculum-Analysis.

Declarations

Competing interests The authors declare that they have no competing interests.

References

Armitage, C. S., Lorenz, M., & Mikki, S. (2020). Mapping scholarly publications related to the sustainable development goals: Do independent bibliometric approaches get the same results? *Quantitative Science Studies, 1*(3), 1092–1108.

AutoGluon (2022). AutoGluon tasks. In *AutoGluon Documentation 0.4.0*. Retrieved December 7, 2023, from https://auto.gluon.ai/stable/tutorials/tabular/tabular-indepth.html

Blohm, M., Hanussek, M., & Kintz, M. (2020). Leveraging automated machine learning for text classification: Evaluation of AutoML tools and comparison with human performance. arXiv preprint. Retrieved from https://arxiv.org/abs/2012.03575

Bowen, K. J., Cradock-Henry, N. A., Koch, F., Patterson, J., Häyhä, T., Vogt, J., & Barbi, F. (2017). Implementing the sustainable development goals: Towards addressing three key governance challenges—collective action, trade-offs, and accountability. *Current Opinion in Environmental Sustainability, 26*, 90–96.

Cf, O. (2015). *Transforming our world: The 2030 agenda for Sustainable Development*. United Nations

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*.

Cheung, K., Lam, A. H., & Chiu, D. K. (2023). Using YouTube and Facebook as German language learning aids: A pilot study in Hong Kong. *German as a Foreign Language*, 1, 146—168.

Chiba, M., Sustarsic, M., Perriton, S., & Edwards, D. B., Jr. (2021). Investigating effective teaching and learning for sustainable development and global citizenship: Implications from a systematic review of the literature. *International Journal of Educational Development, 81*, 102337.

Chung, C. H., Chiu, D. K., Ho, K. K., & Au, C. H. (2020). Applying social media to environmental education: Is it more impactful than traditional media? *Information Discovery and Delivery, 48*(4), 255–266.

Colglazier, W. (2015). Sustainable development agenda: 2030. *Science, 349*(6252), 1048–1050.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *arXiv Preprint*. Retrieved from https://arxiv.org/abs/1810.11363

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). Autogluon-tabular: Robust and accurate automl for structured data. arXiv preprint. Retrieved from https://arxiv.org/abs/2003.06505

Hannon, J., Huggard, S., Orchard, A., & Stone, N. (2014). OER in practice: Organisational change by bootstrapping. *RUSC Universities and Knowledge Society Journal, 11*(3), 134–150.

Ho, C. Y., Chiu, D. K., & Ho, K. K. (2023). Green space development in academic libraries: A case study in Hong Kong. *Global perspectives on sustainable Library practices* (pp. 142–156). IGI Global.

Jha, R. K., Ganguly, S., & Mishra, S. (2020). Alignment of OER platforms with SDGs: An exploratory study. *Handbook of Research on Emerging Trends and Technologies in Library and Information Science* (pp. 77–96). IGI Global.

Jiang, M., Lam, A. H. C., Chiu, D. K. W., & Ho, K. K. W. (2023). Social media aids for business learning: A quantitative evaluation with the 5E instructional model. *Education and Information Technologies, 28*(9), 12269–12291. https://doi.org/10.1007/s10639-023-11690-z

Jomo, K. S., Chowdhury, A., Sharma, K., & Platz, D. (2016). *Public-private partnerships and the 2030 agenda for Sustainable Development*. fit for purpose? *UN Department of Economic and Social Affairs (DESA) Working Paper,* No. 148 ST/ESA/2016/DWP/148. https://doi.org/10.18356/f42bd4bb-en

Kashnitsky, Y., Roberge, G., Mu, J., Kang, K., Wang, W., Vanderfeesten, M., Rivest, M., Keßler, L., Jaworek, R., & Vignes, M. (2022). Identifying research supporting the United Nations Sustainable Development Goals. arXiv preprint. Retrieved from https://doi.org/10.48550/arXiv.2209.07285

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Landorf, H. (2021). Illuminating equity through global learning: Jan L. Tucker memorial lecture. *Journal of International Social Studies, 11*(2), 1–10.

Lane, A. (2017). Open education and the sustainable development goals: Making change happen. *Journal of Learning for Development, 4*(3), 275–286.

Langford, J., Li, L., & Strehl, A. (2007). Vowpal wabbit online learning project. In *Technical report*. Retrieved December 7, 2023, from http://hunch.net

Lei, C. U. (2022). Analysis of the Connection of United Nations Sustainable Development Goals with the Hong Kong High School Technology Curriculum. *Proceedings of the APSCSE International Conference on Computers in Education 2*, 61–64.

Lei, C. U., Cham, C. Y., Liang, X., Qian, X., & Hu, X. (2022). Assessing the integration of United Nations sustainable development goals in a university general education curriculum. *Proceedings of the International Conference on Learning Analytics & Knowledge,* pp. 42–44.

Leung, T. N., Hui, Y. M., Luk, C. K., Chiu, D. K., & Ho, K. K. (2022). *Evaluating Facebook as aids for learning Japanese: learners' perspectives*. Library Hi Tech (ahead-of-print).

Liu, F., Jiang, X., & Zhang, M. (2022). Global burden analysis and AutoGluon prediction of accidental carbon monoxide poisoning by Global Burden of Disease Study 2019. *Environmental Science and Pollution Research, 29*(5), 6911–6928.

Ma, S., Yuan, Z., Wu, Q., Huang, Y., Hu, X., Leung, C. H., Wang, D., & Huang, Z. (2023). Deep Into the Domain Shift: Transfer Learning Through Dependence Regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15. https://doi.org/10.1109/TNNLS.2023.3279099

Monash University. (2017). Compiled-Keywords-for-SDG-Mapping_Final_17-05-10, Australia/Pacific Sustainable Development Solutions Network (SDSN). 2017. Available online: http://ap-unsdsn.org/wp-content/uploads/2017/04/Compiled-Keywords-for-SDG-Mapping_Final_17-05-10.xlsx. Accessed 8 Dec 2023.

Nurhayati, S., Noor, A. H., Musa, S., Jabar, R., & Abdu, W. J. (2022). A digital literacy workshop training model for child parenting in a fourth industrial era. *HighTech and Innovation Journal, 3*(3), 297–305.

OSDG, UNDP IICPSD SDG AI Lab, & PPMI. (2022). OSDG Community Dataset (OSDG-CD) (2022.10) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7136826

Otto, D., & Kerres, M. (2022). Increasing sustainability in open learning: Prospects of a distributed learning ecosystem for open educational resources. *Front Educ, 7*, 866917. https://doi.org/10.3389/feduc.2022.866917

Oyewola, O. M., Ajide, O. O., Osunbunmi, I. S., & Oyewola, Y. V. (2022). Examination of students' academic performance in selected mechanical engineering courses prior-to-and-during COVID-19 era. *Emerging Science Journal, 6*, 247–261.

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.

Pukelis, L., Puig, N. B., Skrynik, M., & Stanciauskas, V. (2020). OSDG–open-source approach to classify text data by UN Sustainable Development Goals (SDGs). arXiv preprint. Retrieved from https://arxiv.org/abs/2005.14569

Reimers, F. M., & Chung, C. K. (2019). *Teaching and learning for the twenty-first century: Educational goals, policies, and curricula from six nations*. Harvard Education.

Roshan, S. E., & Asadi, S. (2020). Improvement of bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization. *Engineering Applications of Artificial Intelligence, 87*, 103319.

Sandanayake, T. C. (2019). Promoting open educational resources-based blended learning. *International Journal of Educational Technology in Higher Education, 16*(1), 1–16.

Schmidt, F., & Vanderfeesten, M. (2021). Evaluation on accuracy of mapping science to the United Nations' Sustainable Development Goals (SDGs) of the Aurora SDG queries. 4964606. https://doi.org/10.5281/zenodo

Seo, K., Chung, B., Panchaseelan, H. P., Kim, T., Park, H., Oh, B., Chun, M., Won, S., Kim, D., & Beom, J. (2021). Forecasting the walking assistance rehabilitation level of stroke patients using artificial intelligence. *Diagnostics, 11*(6), 1096.

Serafini, P. G., Moura, J. M. d., Almeida, M. R. d., & Rezende, J. F. D. D. (2022). Sustainable Development goals in higher education institutions: A systematic literature review. *Journal of Cleaner Production*, *370*, 133–473. https://doi.org/10.1016/j.jclepro.2022.133473

Shi, X., Mueller, J., Erickson, N., Li, M., & Smola, A. (2021). Multimodal automl on structured tables with text fields. *8th ICML Workshop on Automated Machine Learning (AutoML).*

Stagg, A. (2014). OER adoption: A continuum for practice. *International Journal of Educational Technology in Higher Education, 11*(3), 151–165.

Tokovska, M., Zaťková, T., & Jamborová, Ľ. (2022). Digital competencies development in higher education institutions: A mixed methods research study. *Emerging Science Journal, 6*, 150–165.

UNESCO. (2019). Recommendation on Open Educational Resources (OER). Retrieved June 16, 2023, from https://www.unesco.org/en/legal-affairs/recommendation-open-educational-resources-oer

Vanderfeesten, M., Otten, R., & Spielberg, E. (2020). *Search queries for mapping research output to the sustainable development goals (SDGs)* v5.0 (Version 5.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3817445

Vanderfeesten, M., Jaworek, R., & Keßler, L. (2022). AI for mapping multilingual academic papers to the United Nations' Sustainable Development Goals (SDGs). https://zenodo.org/record/6487606#.ZBaotXZBw7c. Accessed 19 Mar 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wang, W., Lam, E. T. H., Chiu, D. K., Lung, MMw., & Ho, K. K. (2021). Supporting higher education with social networks: Trust and privacy vs perceived effectiveness. *Online Information Review, 45*(1), 207–219.

Wang, Y., Li, Y., Hu, X., Xu, Y., Liang, X., & Lei, C. U. (2022). Massive open online courses role in promoting United Nations sustainable development goals. *2022 IEEE Learning with MOOCS (LWMOOCS).*

Wul, D. U., & Meier, D. S. (2023). SDG knowledge hub dataset of SDG-labeled news articles. *Zenodo*. https://doi.org/10.5281/zenodo.7523032

Wulff, D. U., Meier, D. S., & Mata, R. (2023). Using novel data and ensemble models to improve automated labeling of Sustainable Development Goals. arXiv preprint. Retrieved from https://arxiv.org/abs/2301.11353