 doi:10.20944/preprints202210.0381.v1

Article

A RoBERTa Approach for Automated Processing of Sustainability Reports

Merih Angin [1], Beyza Taşdemir [2,‡], Cenk Arda Yılmaz [2,‡], Gökcan Demiralp [2,‡], Mert Atay[2], Pelin Angin[2*], Gökhan Dikmener[3]

[1] Koc University; mangin@ku.edu.tr
[2] Middle East Technical University; {beyza.tasdemir,cenk.yilmaz,gokcan.demiralp, mert.atay, pangin}@metu.edu.tr
[3] United Nations Development Programme, SDG AI Lab; gokhan.dikmener@undp.org
* Correspondence: pangin@ceng.metu.edu.tr
‡ These authors contributed equally to this work.

Abstract: There is a strong need and demand from the United Nations, public institutions, and private sector for classifying government publications, policy briefs, academic literature, and corporate social responsibility reports according to their relevance to the Sustainable Development Goals (SDGs). It is well understood that the SDGs play a major role in the strategic objectives of various entities. However, linking projects and activities to the SDGs has not always been straightforward or possible with existing methodologies. Natural language processing (NLP) techniques offer a new avenue to identify linkages for SDGs from text data. This research examines various machine learning approaches optimized for NLP-based text classification tasks for their success in classifying reports according to their relevance to the SDGs. Extensive experiments have been performed with the recently released Open Source SDG (OSDG) Community Dataset, which contains texts with their related SDG label as validated by the community volunteers. Results demonstrate that especially RoBERTa achieves very high performance in the attempted task, which is promising for automated processing of large collections of sustainability reports for detection of relevance to SDGs.

Keywords: corporate social responsibility; natural language processing; RoBERTa; sustainable development goals

1. Introduction

Corporate social responsibility (CSR), which can be defined as international self-regulation by private companies that includes a political objective related to both positive and negative environmental and social aspects, is becoming increasingly important in today's business world. Despite its contribution to sustainability, under current market conditions, companies seem to be incapable of finding sustainable development solutions on their own. In addition to promote CSR and eco-efficiency, sustainability requires active participation and cooperation of governments, businesses, and citizens [1].

There is undoubtedly a need for effective CSR and sustainability regulations. To meet this need, both voluntary and mandatory initiatives have been launched, which have made sustainability a fundamental part of the corporate agenda. There are numerous voluntary initiatives, some of which are recognized globally, such as the United Nations Global Compact, AA1000 and Global Reporting Initiative (GRI). These mechanisms have improved business efficiency, but their applicability is rather limited. There are also Pollutant Release Transfer Register, Carbon Pricing Mechanisms and CSR/Sustainability and Integrated Reporting Requirements as mandatory initiatives and they have the potential to change the standard approach to CSR. Specific environmental regulations on pollutants and carbon emissions help companies manage their environmental impact and address climate change

 doi:10.20944/preprints202210.0381.v1

while mandatory reporting requirements ensure that a company's CSR activities are known to stakeholders and facilitate accountability [2].

The aforementioned developments and initiatives have led investors to start focusing more on corporate sustainability and environmental, social and governance (ESG) assessments in their investment decisions. As a result, more companies are now voluntarily issuing sustainability reports. Such an increase in sustainability reports is a promising indicator for the future of sustainability, however, we should also note that these reports significantly lack standardization. An analysis on the sustainability reports from the top 20 companies in the SP 500 [3] shows that the reports vary greatly in terms of length, word count and number count. It also shows that most of the reported figures are rounded. These findings emphasize the need to standardize sustainability reporting, and support such ongoing initiatives by regulators that can provide investors with better ESG information.

In this context, one of the most significant initiatives is the Sustainable Development Goals (SDGs) [4] approved by the United Nations General Assembly (UNGA) in 2015, which are intended to be achieved by 2030. SDGs play a key role in facilitating the integration of sustainability to ensure a better and more sustainable future, while responding to the current and future needs of stakeholders and balancing economic, social and environmental development. Table 1 provides a summary of the 17 SDGs. SDGs, in nature, are related to each other and existing research [5] clearly show the significant relationships and interlinkages between the 17 SDGs.

Table 1. Sustainable Development Goals [4]

Goal #	Description
1	End poverty in all its forms everywhere
2	End hunger, achieve food security and improved nutrition and promote sustainable agriculture
3	Ensure healthy lives and promote well-being for all at all ages
4	Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all
5	Achieve gender equality and empower all women and girls
6	Ensure availability and sustainable management of water and sanitation for all
7	Ensure access to affordable, reliable, sustainable and modern energy for all
8	Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all
9	Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation
10	Reduce inequality within and among countries
11	Make cities and human settlements inclusive, safe, resilient and sustainable
12	Ensure sustainable consumption and production patterns
13	Take urgent action to combat climate change and its impacts
14	Conserve and sustainably use the oceans, seas and marine resources for sustainable development
15	Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss
16	Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels
17	Strengthen the means of implementation and revitalize the global partnership for sustainable development

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2022 doi:10.20944/preprints202210.0381.v1

SDGs can be broadly categorized under three main topics, i.e., economy, society and environment, based on the relevance of their goals. Figure 1 demonstrates a visualization of this categorization.

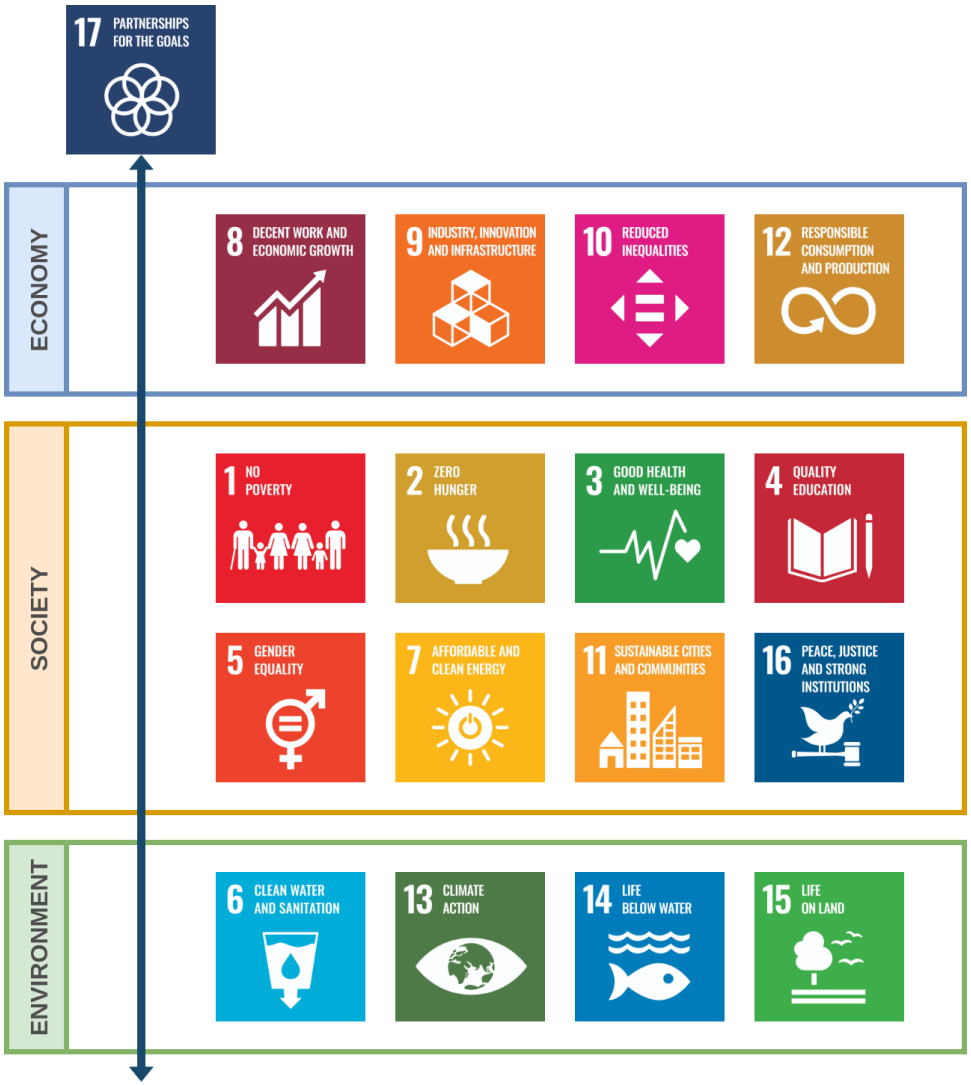

Figure 1. Sustainable Development Goals

Taking into account the notable increase in sustainability reports, approval of 17 SDGs and the presence of interlinkages between them, utilizing digital platforms and solutions to effectively classify reports according to their relevance to different SDGs is imperative [6]. In this work, we describe a natural language processing (NLP)-based framework for processing sustainability reports to identify sections relevant to SDGs. Our framework has been built and evaluated using the recently released OSDG Community Dataset containing textual data on the SDGs, annotated by community volunteers. We have performed extensive experiments to compare the performances of both classical machine learning (ML) models and deep learning (DL) models in binary and multi-class classification tasks and advocate the utilization of RoBERTa-based processing due to its superior performance in both tasks. To the best of our knowledge, this is the first work to achieve such high performance in automated classification of sustainability documents from a collection of this size.

The remainder of this paper is organized as follows: Section 2 summarizes related work in the field of automated text processing for sustainability and other similar domains. Section 3 provides details of our framework for classification of sustainability reports based

 doi:10.20944/preprints202210.0381.v1

on SDGs. Section 4 provides performance analysis of the framework for different ML and DL models and Section 5 concludes the paper with future work directions.

2. Related Work

With the increasing number of documents and texts created by international organizations and companies each day, new research efforts have been dedicated to automate the time-consuming process of reading such documents and identifying texts related to target topics such as SDGs.

NLP methods have long been used to automatically process large document collections produced by international organizations. Deniz et al. [7] used NLP to automatically classify sentiments in the large document collection of the International Monetary Fund Executive Board meeting minutes achieving high accuracy when model training was performed with domain-specific data. Sovrano et al. [8] proposed an ensemble method for multi-label text classification of UNGA Resolutions, combining non domain-specific deep learning based document similarities with domain-specific Term Frequency-Inverse Document Frequency (TF-IDF) document similarities. Their proposed method achieved modest performance, but their domain-specific similarity addition improved the baseline without any transfer learning or re-training. Kim and LaFleur [9] proposed a proof-of-concept classifier for analyzing UNGA resolutions adopting the dictionary method and supervised learning. Their proposed classifier achieved 94% test accuracy and their analysis showed how NLP techniques can be used to identify trends and provide insight on the UN's work reflected in UNGA Resolutions.

After the introduction of the 17 SDGs by the UN, many research efforts were also dedicated to identifying and linking SDGs by automatically processing data from various sources. Yeh et al. [10] proposed "SUSTAINBENCH", a benchmark and a public leaderboard website for multiple SDG related datasets with standard train-test splits and well-defined performance metrics. They also provided baseline models and their evaluation results for each dataset. Matsui et al. [11] proposed an NLP model for supporting sustainable development goals, which involves translating semantics, visualizing nexus, and connecting stakeholders. Nilsson et al. [12] developed a framework for mapping interactions between the SDGs. They focused on modeling interactions through important factors such as geographical context, resource endowments, time horizon and governance. Smith et al. [13] proposed an approach to quantify the network of SDG interdependencies using policy and scientific documents. Their proposed method combined NLP methods and network analysis to provide a mapping of SDGs' relationships. Toetzke et al. [14] proposed a machine learning framework for categorization of global development aid activities based on textual descriptions. Their framework utilized document embeddings and clustering and generated activity clusters representing the topics of underlying aid activities many of which were yet to be analyzed empirically. While these works are all based on processing SDG-related data, they focused on different problems than automatically detecting the relevancy of texts in large document collections to one or more specific SDGs.

A limited number of recent works have focused on utilization of machine learning and deep learning techniques to develop text classification systems capable of identifying, with high accuracy, the related SDGs in a document collection. Pukelis et al. [15] proposed Open Source SDG (OSDG) project and tool to integrate data from multiple sources into a single framework for SDG classification. The integration aimed linking features from previous approaches and research (e.g., ontology items, keywords, features from machine learning models) to the topics in the Microsoft Academic Graph. Amel-Zadeh et al. [16] provided a proof of concept for the use of ML and NLP to detect companies' alignment with SDGs based on their CSR reports. Their proposed method with binary outcomes used Word2Vec [17] and Doc2Vec models for training a logistic regression classifier, a fully-connected neural network and an SVM which, with a Doc2Vec [18] embedding, achieved the highest average accuracy of 83.5% for predicting alignment. Guisiano et al. [19,20] proposed a multi-label classification system using BERT and an online tool "SDG-meter"

 doi:10.20944/preprints202210.0381.v1

to automate this task. Their proposed BERT model achieved an accuracy of 98%. Despite the high accuracy achieved, the system was only tested on a collection of 400 texts, which is quite limited. Hajikhani and Suominen [21] proposed an ML model to automate the detection of SDG relevancy in patent documents. The authors also presented relatedness between different SDG categories using their highest performing model, which was the logistic regression classifier utilizing Word2Vec. The ML models they utilized above 60% accuracy for most SDGs. While some of the mentioned works have achieved successful results, to the best of our knowledge, none of the existing works evaluated the performance of their models on both binary and multi-class classification tasks for SDG relevance. Also RoBERTa, which we demonstrate to outperform all other models in this work, has not been utilized in any of the existing frameworks.

3. Automated Report Processing Framework

In this section we provide details on our automated report processing framework. We have developed both classical machine learning (ML) and deep learning (DL) based models for processing of reports to identify relevancy of text blocks in them to SDGs. Furthermore, we have developed both binary and multi-class classification models. Binary classification models have been built to identify whether a text block is relevant to each SDG, whereas multi-class classification models indicate the most relevant SDG for the given text block.

ML and DL-based models involve different processing steps during model building and execution. In the subsections below, we describe the overall operation of each.

3.1. ML-based Processing

The automated text processing pipeline for the ML-based methods consists of 4 stages: Text Pre-processing, Vectorization, Model Training, and Model Execution. We start with a training dataset, where each text block is labelled by human annotators as relevant to one of the SDGs.

Pre-processing

In the **text pre-processing** stage, we first filter out the records with a low agreement score from the dataset. Having a low agreement score means that multiple annotators had different views about the particular text block's being relevant to a specific SDG. Removal of such instances from the data allows us to achieve higher quality in model training, as we will only be learning from instances that we are more certain about. Then, we filter out stopwords such as *and*, *the*, *is* etc., as well as punctuation marks, one-letter words and numbers, as these do not contribute to the meaning of sentences. To eliminate these, we utilize regular expressions and the relevant methods of the Natural Language Toolkit (NLTK) [22] for Python.

Lemmatization is an operation that converts words into their simplest form. It is widely used in pre-processing of raw text data in NLP tasks. By lemmatizing words, we aim to reduce potential confusions that different representations of the same base word could create, although they add the similar or the same meaning to the context. We use NLTK WordNet Lemmatizer for lemmatization of the text. WordNet [23] is a large lexical database consisting of English words, allowing the analysis and processing of English sentences. We first tokenize sentences and find the POS (part-of-speech) tag for each token, which is basically the function of the word in a sentence (noun, verb, adjective, etc.). Then, according to their POS, we find the correct base-form of the words using NLTK WordNet Lemmatizer. An example lemmatization is shown in 2.

Vectorization

In order to convert the data into a form that can be processed by ML algorithms, numeric matrices need to be constructed for the data instances. This is achieved through **vectorization**, which maps each word in the complete dataset to a unique number and keeps counts of the occurrences of each word in the data instances (i.e. text blocks). The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2022 doi:10.20944/preprints202210.0381.v1

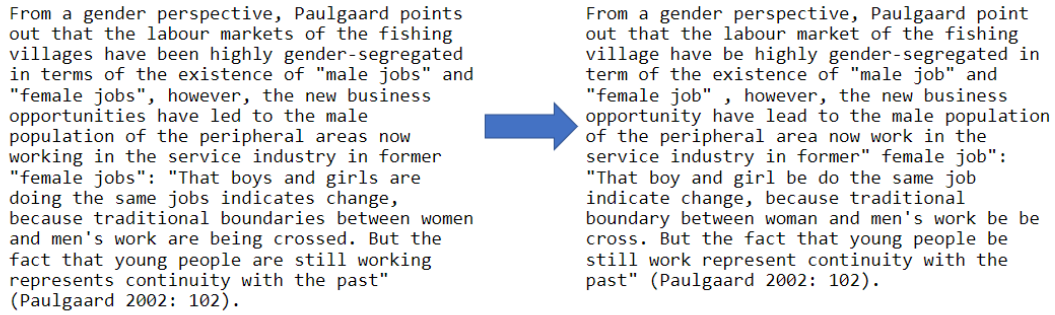

Figure 2. Lemmatization of text

importance of a specific word for a particular class should not only be determined by how frequently that word occurs in the data instances of that class, but also by how infrequently it occurs in instances of other classes, i.e., if a word occurs too often in the whole dataset, it carries less information. In order to account for this rationale, we use a TF-IDF vectorizer [24], which calculates a score for each word in a text block by multiplying the frequency score of that word in the text block by the inverse of the frequency of that word for the whole dataset.

Model Training

In this stage, we train different ML models with the dataset formed in the previous stage and optimize the models. In this work, we have utilized 4 well-known classification algorithms, namely Gaussian Naive Bayes, Decision Tree, Logistic Regression, and Linear Support Vector Machines.

3.2. Deep Learning-based Processing

DL algorithms have achieved significant success in many learning tasks in the past decade with the developments in computing infrastructures and the availability of big data, and the field of NLP is no exception. For the DL-based models, we did not apply any pre-processing to the data, because the whole information in a sentence can be helpful when using pre-trained language representation models like BERT and RoBERTa. They both need special encoding and tokenization methods of the inputs so that they can be trained. For that purpose, we used the dedicated methods in the *transformers* library of the HuggingFace [25] AI community. Below we provide a short explanation of both DL models utilized.

3.2.1. BERT

BERT [26], which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, is an open-source, pre-trained deep learning algorithm developed by Google. It is proficient at a variety of NLP tasks, including sequence-to-sequence based language generation tasks such as question answering and sentence prediction, and natural language understanding tasks such as sentiment classification and word sense disambiguation.

BERT uses an attention mechanism called *Transformer* that recognizes contextual relationships between words (or subwords) in a text by using the surrounding text to establish context. Transformer's basic design consists of two independent mechanisms: an encoder that reads the text input, and a decoder that generates a word prediction. BERT architecture consists of several Transformer encoder stacks, which have a bidirectional

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2022 doi:10.20944/preprints202210.0381.v1

nature. This means that BERT learns information from a sequence of words from both directions.

3.2.2. RoBERTa

RoBERTa [27], which stands for **R**obustly **O**ptimized **BERT** Pre-training **A**pproach, is a variation of BERT, proposed by researchers at Facebook and Washington University. It not only optimizes the training of BERT architecture during pre-training but also is superior at predicting intentionally concealed sections of a text. RoBERTa has a similar architecture to BERT with a few alterations in the architecture and the training procedure. RoBERTa alters key hyperparameters in BERT, such as BERT's next-sentence pre-training intent. In addition to dynamically modifying the masking pattern as opposed to a single static mask in BERT, using larger mini-batches, higher learning rates, and longer sequences during training allows RoBERTa to be more successful and more performance-oriented than BERT at masked language modeling.

4. Evaluation

4.1. SDG Dataset

OSDG Community Dataset (OSDG-CD) [28], first published by the OSDG team on October 1st, 2021, is a public dataset that aims to support NLP research and studies on deriving insights into the nature of SDGs. OSDG-CD contains texts with their related SDG labels validated by more than a thousand volunteers from over 100 countries via the OSDG Community Platform (OSDG-CP). In our experiments, we used version 04.2022, which was the latest version of the dataset when this research was conducted. In this section, we provide details of the dataset and present our analysis.

The dataset contains texts from various public documents such as reports, policy documents, and publication abstracts; moreover, each text excerpt is nearly a paragraph in length. We validated this information by analyzing the word numbers of texts. The longest text consists of 226 tokens (words), while the shortest consists of 16 tokens. Furthermore, the average size of documents is approximately 89.79 tokens per text. Considering these, the dataset can be safely used, without truncating, to fine-tune most of the transformer models as they generally consume 512 tokens at maximum.

Although there are 17 Sustainable Development Goals defined by UNDP, the dataset includes the first 15 SDGs. Our observations on the distribution of texts over those SDGs show that the dataset is quite unbalanced. Figure 3 shows the distribution of the SDGs. Therefore, we split the data in a stratified manner so that the ratio between test and train data for each SDG category was preserved.

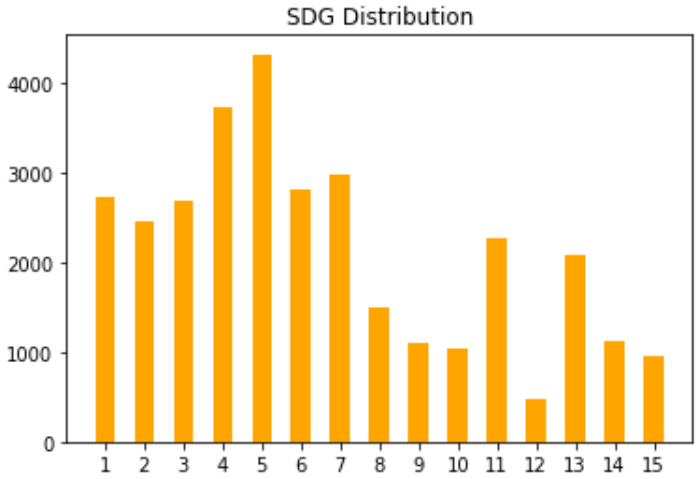

Figure 3. SDG Distribution Histogram

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2022 doi:10.20944/preprints202210.0381.v1

The volunteers from OSDG-CP contributed to the annotation of the dataset by completing some labeling exercises. Each text was validated by at least three different volunteers and up to 9 different volunteers. All the labeling exercises were binary decision problems, meaning that each volunteer could accept or reject a suggested label. This information was embedded into the dataset as *'labels_negative'*, *'labels_positive'* and *'agreement'* columns, where *'agreement'* represented the agreement score based on the formula below:

$$agreement = \frac{|labels_{positive} - labels_{negative}|}{labels_{positive} + labels_{negative}}$$

When we analyzed the dataset, we observed that some records had low agreement scores. Also, in some records, the number of volunteers who rejected the suggested label was more than those who accepted it. These 'low-quality' records can potentially reduce the accuracy of classifier models; therefore, it would be appropriate to filter them out in the pre-processing stages of the experiments. Figure 4 shows the distribution of quality and poor quality records in each SDG, where quality records are defined as the records having an agreement score greater than or equal to 0.6 and having *'labels_positive'* greater than *'labels_negative'*. In contrast, the poor quality records are the remaining ones.

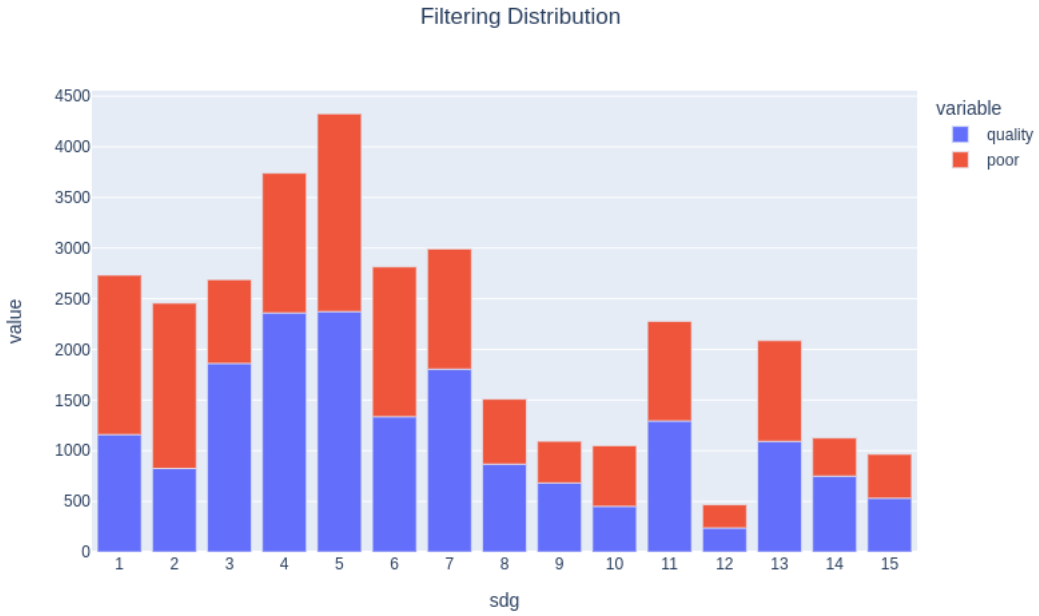

Figure 4. Quality and Poor records

4.2. Evaluation Results

In this section, we report the results of both binary and multi-class classification for the ML and DL-based models. For evaluating the performances of the different algorithms, we utilized commonly used metrics from the ML literature: accuracy, precision, recall and F1 score. These metrics are calculated using the formulae below, where TP is the number of positive instances classified as positive, TN is the number of negative instances classified as negative, FP is the number of negative instances classified as positive and FN is the number of positive instances classified as negative by the algorithm:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2022 doi:10.20944/preprints202210.0381.v1

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

4.2.1. Binary Classification

Foremost, we performed binary classification for all SDGs separately, and while doing this, we created a balanced dataset by undersampling for each SDG group in each trial by taking equal-sized random samples from the ones that do not belong to that SDG. Our goal was to observe what the results would look like if we had a balanced dataset.

Finally, we created a matrix representing each sentence and the words contained in it using the vectorizer we just obtained. Each row of the matrix is a sentence, and each cell is a word that is tokenized. At the model training stage for ML algorithms, we trained four models for all SDGs:

- Linear Support Vector Machines (SVM)
- Gaussian Naive Bayes
- Decision Tree
- Logistic Regression (LR)

Table 2 summarizes the F1 score results of the four ML algorithms for each SDG.

Table 2. Binary classification results (F1 scores) for machine learning models

SDG#	LR	SVM	Naive Bayes	Decision Tree
1	0.92	0.93	0.79	0.86
2	0.93	0.94	0.81	0.90
3	0.96	0.97	0.85	0.91
4	0.98	0.98	0.82	0.95
5	0.97	0.98	0.81	0.96
6	0.96	0.96	0.79	0.95
7	0.95	0.96	0.80	0.93
8	0.88	0.87	0.75	0.75
9	0.91	0.93	0.81	0.81
10	0.89	0.90	0.79	0.73
11	0.92	0.94	0.76	0.86
12	0.91	0.89	0.81	0.81
13	0.93	0.95	0.83	0.92
14	0.96	0.97	0.85	0.96
15	0.92	0.92	0.81	0.88

At the model evaluation stage for ML models, we concluded that SVM and Logistic Regression models produce the best results. Although these models achieved relatively better results than the other traditional ML algorithms, we applied Deep Learning based NLP methods for further comparison. First, we trained BERT and RoBERTa models with the same train and test datasets, and then evaluated their binary classification performance. As expected, these models outperformed the traditional ML algorithms. The results can be seen in Tables 3 and 4 for BERT and RoBERTa respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2022 doi:10.20944/preprints202210.0381.v1

Table 3. Binary classification results for BERT

SDG#	TP#	TN#	FP#	FN#	Accuracy	Precision	Recall	F1
1	363	362	13	28	0.946	0.965	0.928	0.947
2	268	260	13	4	0.969	0.954	0.985	0.969
3	623	585	16	5	0.983	0.975	0.992	0.983
4	788	749	14	8	0.986	0.983	0.990	0.986
5	798	746	18	5	0.985	0.978	0.994	0.986
6	433	429	10	10	0.977	0.977	0.977	0.977
7	606	556	20	10	0.975	0.968	0.984	0.976
8	256	263	29	25	0.906	0.898	0.911	0.905
9	215	206	1	19	0.955	0.995	0.919	0.956
10	141	125	20	12	0.893	0.876	0.922	0.898
11	412	414	12	17	0.966	0.972	0.960	0.966
12	71	73	3	11	0.911	0.959	0.866	0.910
13	364	332	13	13	0.964	0.966	0.966	0.966
14	244	249	1	1	0.996	0.996	0.996	0.996
15	182	156	8	5	0.963	0.958	0.973	0.966

Table 4. Binary classification results for RoBERTa

SDG#	TP#	TN#	FP#	FN#	Accuracy	Precision	Recall	F1
1	363	362	13	28	0.946	0.965	0.928	0.947
2	268	264	9	4	0.976	0.968	0.985	0.976
3	618	591	10	10	0.984	0.984	0.984	0.984
4	789	753	10	7	0.989	0.987	0.991	0.989
5	797	746	18	6	0.985	0.978	0.993	0.985
6	437	430	9	6	0.983	0.980	0.986	0.983
7	608	556	20	8	0.977	0.968	0.987	0.977
8	257	263	29	24	0.908	0.899	0.915	0.907
9	219	207	10	15	0.945	0.956	0.936	0.946
10	141	129	16	12	0.906	0.898	0.922	0.910
11	410	414	12	19	0.964	0.972	0.956	0.964
12	74	72	4	8	0.924	0.949	0.902	0.925
13	371	331	14	6	0.972	0.964	0.984	0.974
14	244	247	3	1	0.992	0.988	0.996	0.992
15	183	157	7	4	0.969	0.963	0.979	0.971

4.2.2. Multi-class Classification

In this section, we present the results of our multi-class classification experiments. As before, these experiments consist of 2 sub-experiments, the first is using supervised ML methods, and the second is fine-tuning BERT and RoBERTa models.

For multi-class classification models, in addition to F1 score, we also used confusion matrices to demonstrate the classification algorithms' performance. Since we have more than 2 classes in these classification tasks, by plotting a confusion matrix, we can better see which SDGs our models are confusing the other SDGs with. A confusion matrix is used to demonstrate how many test samples of a specific class were classified as instances of all classes in our task. For example, in the Gaussian Naive Bayes confusion matrix in Figure 5, by looking at the first row, we see that for SDG1, the classifier correctly predicted 161 instances as belonging to class SDG1, while it incorrectly predicted 5 samples as belonging to class SDG2.

The accuracies, F1 scores and confusion matrices based on the performances of the mentioned ML models on the test dataset can be seen in Figures 5, 6, 7 and 8. As seen in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2022 doi:10.20944/preprints202210.0381.v1

the confusion matrices, the lack of text data for SDG 10: Reduced Inequalities and SDG 12: Responsible Consumption and Production have affected the models' performance negatively.

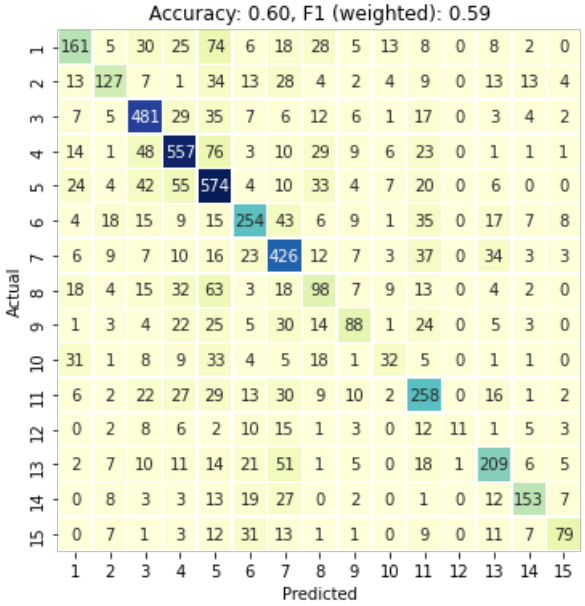

Figure 5. Gaussian Naive Bayes Confusion Matrix

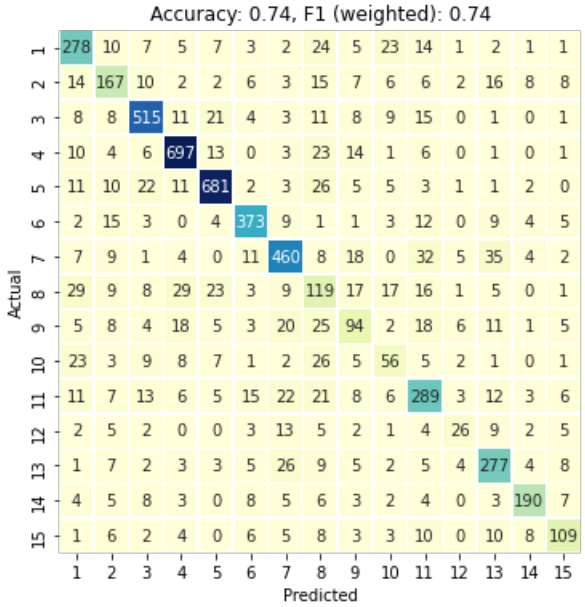

Figure 6. Decision Tree Confusion Matrix

After training the models and evaluating their performances, we found that BERT and RoBERTa outperformed the traditional ML methods that we tried in the first sub-experiment. Their F1 scores and confusion matrices can be seen in Figures 9 and 10. We have achieved an F1 score of 91 percent with BERT and 92 percent with RoBERTa, which are quite high. These results demonstrate that especially deep learning-based NLP models achieve significant success in the attempted task, which is promising for automated processing of large collections of sustainability reports for detection of relevance to SDGs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2022 doi:10.20944/preprints202210.0381.v1

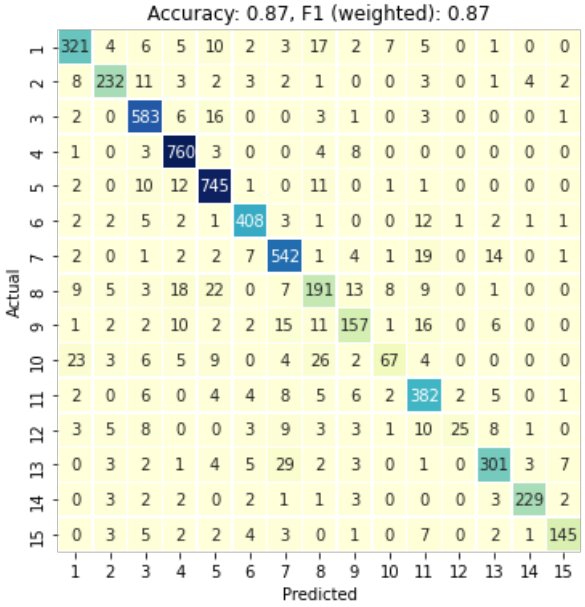

Figure 7. Logistic Regression Confusion Matrix

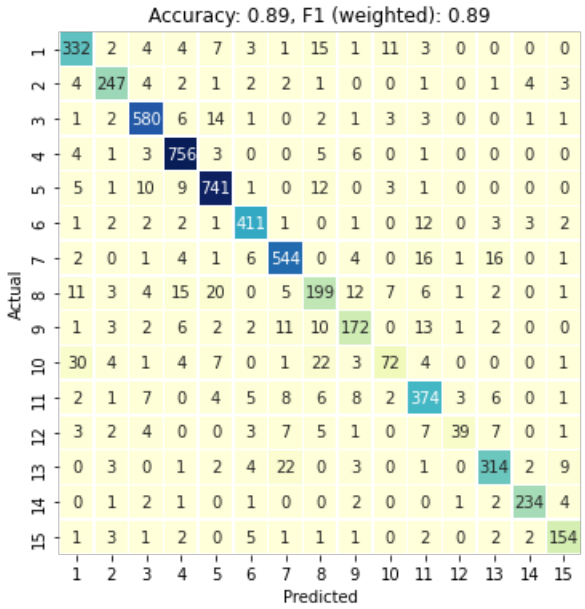

Figure 8. Linear Support Vector Machines Confusion Matrix

5. Conclusion

It is well established that the SDGs play a key role in the strategic objectives of diverse entities. Nevertheless, connecting projects and activities to the SDGs has been rather complicated and not always possible with existing methods. NLP provides a novel way to classify linkages for SDGs from text data. This research examined various machine learning and deep learning approaches optimized for NLP text classification tasks for their success in classifying textual data according to their relevance to SDGs. Extensive experiments have been performed with the recently released OSDG Community Dataset. Results demonstrate that especially RoBERTa achieves significant success in the attempted task, which is promising for automated processing of large document collections for detection of relevance to SDGs. The framework we have developed in this work can be readily used by the community for processing sustainability reports with high SDG detection/identification accuracy. In our future work we aim to use the same methodology to classify national

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 October 2022 doi:10.20944/preprints202210.0381.v1

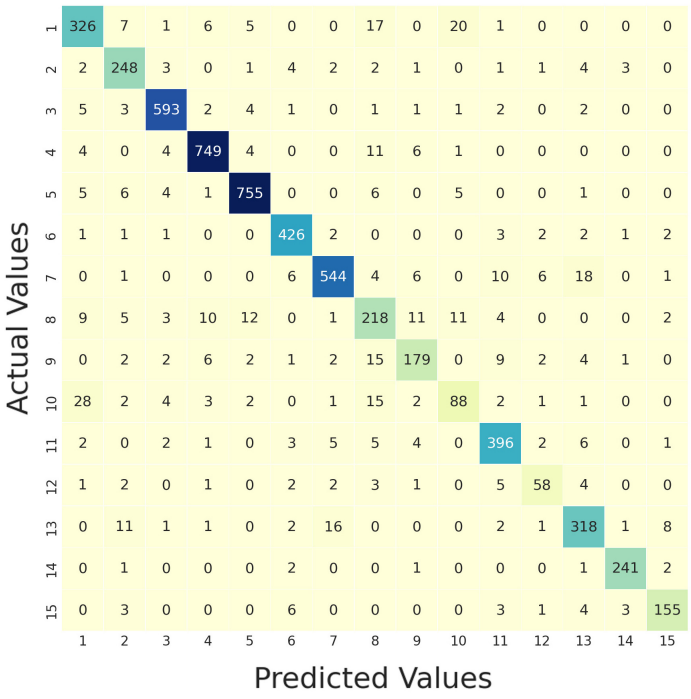

Figure 9. BERT Confusion Matrix
F1 Score: 0.91

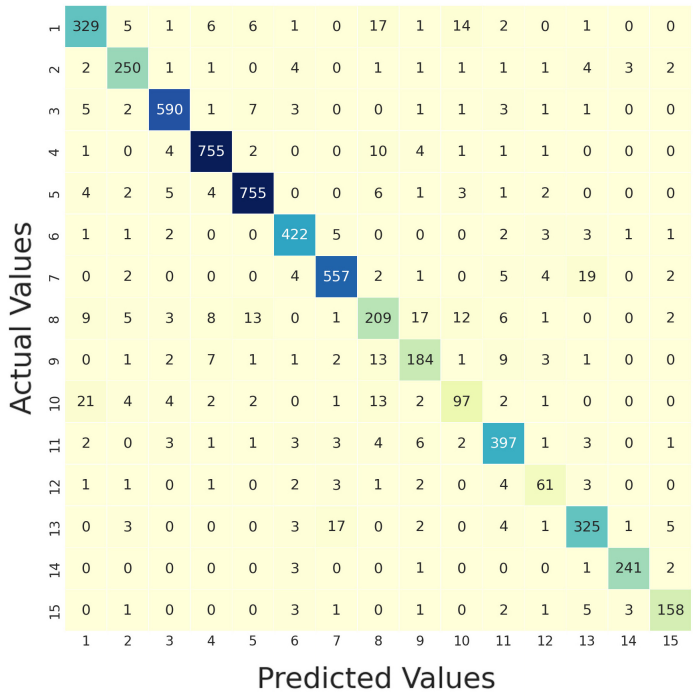

Figure 10. RoBERTa Confusion Matrix
F1 Score: 0.92

artificial intelligence (AI) strategy documents of over 50 countries to examine the linage between SDGs and AI development, particularly in the Global South.

 doi:10.20944/preprints202210.0381.v1

Author Contributions: Conceptualization, Merih Angin and Gökhan Dikmener; methodology, Merih Angin and Pelin Angin; software, Beyza Taşdemir, Cenk Arda Yılmaz and Gökcan Demiralp; validation, Beyza Taşdemir, Cenk Arda Yılmaz and Gökcan Demiralp; formal analysis, Beyza Taşdemir, Cenk Arda Yılmaz and Gökcan Demiralp; investigation, Merih Angin and Pelin Angin; resources, Merih Angin; data curation, Beyza Taşdemir, Cenk Arda Yılmaz, Gökcan Demiralp and Gökhan Dikmener; writing—original draft preparation, Merih Angin, Pelin Angin, Beyza Taşdemir, Cenk Arda Yılmaz, Gökcan Demiralp, Mert Atay; writing—review and editing, Merih Angin, Pelin Angin, Mert Atay and Gökhan Dikmener; visualization, Beyza Taşdemir, Cenk Arda Yılmaz, Gökcan Demiralp; supervision, Merih Angin, Pelin Angin and Gökhan Dikmener; project administration, Merih Angin and Pelin Angin; funding acquisition, Merih Angin.

Funding: This research was funded by H2020 Marie Skłodowska-Curie Actions (H2020-MSCA-IF-2019) grant number 896716. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Data Availability Statement: The OSDG Community Dataset (OSDG-CD) used in this study is publicly available via https://zenodo.org/record/6393942.Yy9bY1JBzqp.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CSR	Corporate social responsibility
DL	Deep Learning
ESG	Environmental, Social and Governance
GRI	Global Reporting Initiative
ML	Machine Learning
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OSDG	Open Source SDG
OSDG-CP	OSDG Community Platform
POS	Part-of-speech
RoBERTa	Robustly Optimized BERT Pre-training Approach
SDG	Sustainable Development Goal
TF-IDF	Term Frequency-Inverse Document Frequency
UN	United Nations
UNDP	United Nations Development Program
UNGA	United Nations General Assembly

References

1. Málovics, G.; Csigéné, N.N.; Kraus, S. The role of corporate social responsibility in strong sustainability. *The Journal of Socio-Economics* **2008**, *37*, 907–918.
2. Lodhia, S.K. The need for effective corporate social responsibility/sustainability regulation. *Contemporary issues in Sustainability accounting, assurance and reporting. Bingley: Emerald Publishing Limited* **2012**, pp. 139–152.
3. Ascioglu, A.; Gonzalez, J.; Zbib, L. Analysis of Sustainability Reports for Top 20 Companies in the S&P 500 Index. *The Journal of Impact and ESG Investing* **2022**, *2*, 82–94.
4. Nations, U. Transforming our world: the 2030 Agenda for Sustainable Development. https://sdgs.un.org/2030agenda.
5. Fonseca, L.M.; Domingues, J.P.; Dima, A.M. Mapping the Sustainable Development Goals Relationships. *Sustainability* **2020**, *12*. https://doi.org/10.3390/su12083359.
6. Bonina, C.; Koskinen, K.; Eaton, B.; Gawer, A. Digital platforms for development: Foundations and research agenda. *Information Systems Journal* **2021**, *31*, 869–902.
7. Deniz, A.; Angin, M.; Angin, P. Understanding IMF Decision-Making with Sentiment Analysis. In Proceedings of the 2022 30th Signal Processing and Communications Applications Conference (SIU), 2022, pp. 1–4. https://doi.org/10.1109/SIU55565.2022.9864926.
8. Sovrano, F.; Palmirani, M.; Vitali, F. Deep Learning Based Multi-Label Text Classification of UNGA Resolutions. *CoRR* **2020**, *abs/2004.03455*, [2004.03455].

 doi:10.20944/preprints202210.0381.v1

9. Kim, N.; LaFleur, M. What does the United Nations "say" about global agenda? An exploration of trends using natural language processing for machine learning. *DESA Working Paper No. 171* **2020**.
10. Yeh, C.; Meng, C.; Wang, S.; Driscoll, A.; Rozi, E.; Liu, P.; Lee, J.; Burke, M.; Lobell, D.B.; Ermon, S. SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning. In Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.
11. Matsui, T.; Suzuki, K.; Ando, K.; Kitai, Y.; Haga, C.; Masuhara, N.; Kawakubo, S. A natural language processing model for supporting sustainable development goals: translating semantics, visualizing nexus, and connecting stakeholders. *Sustainability Science* **2022**, *17*, 969—-985.
12. Nilsson, M.; Chisholm, E.; Griggs, D.; Howden-Chapman, P.; McCollum, D.; Messerli, P.; Neumann, B.; Stevance, A.S.; Visbeck, M.; Stafford-Smith, M. A natural language processing model for supporting sustainable development goals: translating semantics, visualizing nexus, and connecting stakeholders. *Sustainability Science* **2018**, *13*, 1489—-1503.
13. Smith, T.B.; Vacca, R.; Mantegazza, L.; Capua, I. Natural language processing and network analysis provide novel insights on policy and scientific discourse around Sustainable Development Goals. *Scientific reports* **2021**, *11*, 1–10.
14. Toetzke, M.; Banholzer, N.; Feuerriegel, S. Monitoring global development aid with machine learning. *Nature Sustainability* **2022**, pp. 1–9.
15. Pukelis, L.; Bautista-Puig, N.; Skrynik, M.; Stanciauskas, V. OSDG - Open-Source Approach to Classify Text Data by UN Sustainable Development Goals (SDGs). *CoRR* **2020**, *abs/2005.14569*, [2005.14569].
16. Amel-Zadeh, A.; Chen, M.; Mussalli, G.; Weinberg, M. NLP for SDGs: Measuring Corporate Alignment with the Sustainable Development Goals. *The Journal of Impact and ESG Investing* **2022**, *2*, 61–81.
17. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* **2013**.
18. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International conference on machine learning. PMLR, 2014, pp. 1188–1196.
19. Guisiano, J.; Chiky, R. Automatic classification of multilabel texts related to Sustainable Development Goals (SDGs). In Proceedings of the TECHENV EGC2021, 2021.
20. Guisiano, J.E.; Chiky, R.; de Mello, J. SDG-Meter: a deep learning based tool for automatic text classification of the Sustainable Development Goals. In Proceedings of the ACIIDS: 14th Asian Conference on Intelligent Information and Database Systems, 2022.
21. Hajikhani, A.; Suominen, A. The interrelation of sustainable development goals in publications and patents: A machine learning approach. *CEUR Workshop Proceedings* **2021**, *2871*, 183–193.
22. Natural Language Toolkit. https://www.nltk.org/.
23. Miller, G.A. *WordNet: An electronic lexical database*; MIT Press, 1998.
24. Ramos, J.; et al. Using tf-idf to determine word relevance in document queries. In Proceedings of the Proceedings of the first instructional conference on machine learning, 2003, Vol. 242, pp. 29–48.
25. Hugging Face. https://huggingface.co/.
26. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
27. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**.
28. OSDG.; Lab, U.I.S.A.; PPMI. OSDG Community Dataset (OSDG-CD), 2022. https://doi.org/10.5281/zenodo.6393942.