International Journal of Basic and Applied Sciences, 14 (2) (2025) 401-409

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/9g302r53
Research paper

Deep Learning-Based Classification of Comments and Reviews for Sustainable Development Goals (SDGs) with Web Application Implementation

Dhanya D [1*], S. Kalaivany [2], M. Suresh Anand [3], Rakhi [4], G. Jaya Raju [5], Anurag Vijay Agrawal [6], N. Sivakumar [7], Ravindra Namdeorao Jogekar[8]

[1] Department of Artificial Intelligence and Data Science, Mar Ephraem College of Engineering and Technology, Kanyakumari, Tamil Nadu 629171, India
[2] Department of Information Technology, Mailam Engineering College, Villupuram, Tamil Nadu 604304, India
[3] Department of Computing Technologies, School of Computing, SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu 603203, India
[4] Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, Paschim Vihar, New Delhi 110063, India
[5] Department of Computer Science and Engineering, Aditya University, Surampalem, Andhra Pradesh 533437, India
[6] Department of Electronics and Communication Engineering, Bhagwant Institute of Technology, Bhagwantpuram, Muzaffarnagar, Uttar Pradesh 251315, India
[7] Department of Artificial Intelligence and Data Science, Panimalar Engineering College, Chennai, Tamil Nadu 600123, India
[8] Department of Computer Science and Engineering, Jhulelal Institute of Technology, Nagpur, Maharashtra 441111, India
**Corresponding author E-mail: dhanvis@gmail.com*

Received: April 18, 2025, Accepted: June 18, 2025, Published: June 26, 2025

Abstract

This work aims to develop an intelligent system that leverages natural language processing (NLP) and deep learning to analyze and interpret textual data in the context Sustainable Development Goals (SDGs). The core objective is to identify semantic relationships between input text and specific SDGs, thereby enabling automated classification and supporting sustainable decision-making. The work focuses on the application of data augmentation techniques to enhance training datasets, refinement of existing classification models through hyperparameter tuning, and the proposal of a novel classification model to improve accuracy and reliability. Additionally, the work includes the development of a user-friendly web application with extended functionalities, allowing users to input text manually or upload text files to determine the most relevant SDG. This integrated approach aims to bridge the gap between unstructured textual data and structured sustainability frameworks, providing an innovative tool for researchers, policymakers, and organizations working toward global sustainability goals.

Keywords: *Learning; Sustainable Development Goals; Web Application; Classification of Comments*

1. Introduction

The Sustainable Development Goals (SDGs) are a set of global objectives that, in general terms, different countries aim to eradicate poverty, protect the planet, and ensure prosperity for all as part of a new agenda (Nilsson et al., 2018; Holzinger et al., 2019). These goals aim to ensure that by 2030, all people enjoy peace and prosperity. The 17 SDGs are integrated, meaning they recognize that action in one area will affect results in others and that development must balance social, economic, and environmental sustainability (Holzinger, Carrington, & Müller, 2020).

In Colombia, the agenda has been implemented, and, to this end, it provides spaces for ongoing dialogue between all sectors of the national, departmental, and municipal governments. Furthermore, initiatives are generated that enjoy greater ownership by other actors in society (Chou et al., 2022).

Due to their global importance, it is vital for all nations to be able to measure the growth or decline of these goals. This is done in order to take the necessary measures and decisions to achieve both theoretical and practical improvements in the different areas covered by the SDGs by 2030, improving the quality of life of the global population.

Faced with this challenge, artificial intelligence (AI), and especially machine learning (ML) and deep learning (DL), appear to be a significant collaborative effort for different countries to achieve their goals. Artificial Intelligence can be defined as the combination of algorithms designed to create machines that exhibit the same capabilities as humans (Holzinger, Malle, Saranti, & Pfeifer, 2021). Within this

vast science, there is the branch of Machine Learning, which, through algorithms, provides computers with the ability to identify patterns in massive amounts of data and make predictions (Corbett-Davies & Goel, 2018). Finally, within this branch is Deep Learning, which is essentially a neural network that attempts to simulate the behavior of the human brain, ingesting and processing unstructured data, such as text and images, and thereby automating feature extraction, eliminating human dependence (Barocas & Selbst, 2016).

There are already several ways in which technology, and especially Artificial Intelligence, are helping around the world to achieve the SDGs. For example, AI enables the development of sustainable agriculture through decision-support systems by suggesting the best variety to grow for small farmers (Leszczynski & Zook, 2020).

Another example of the use of Artificial Intelligence for the benefit of SDGs is mentioned by Alejandro Arango, leader of digital transformation and advanced analytics at Ecopetrol, who suggests that AI can contribute to the fulfillment of health-related SDGs. Thanks to the interconnectivity between a large part of hospitals, global health conditions can be monitored and mapped in real time (Tvaronavičienė et al., 2020; Limba et al., 2017).

In this work, we will focus on improving the processing of people's texts/comments to classify their content among the 16 different SDGs. To do this, we will draw on a previously completed master's degree work, which highlighted difficulties in classifying SDG categories with limited existing data/samples. Next, the survey of the research works in the above context is provided below.

2. Literature Review

The following studies show how explainable and robust sentiment models can enhance short-text understanding, supporting interdisciplinary applications such as social science, linguistics, and SDG-related communication analysis.

Pradhan et al. (2017) conducted a comprehensive assessment of the SDGs' linkages, revealing the complex web of synergies and trade-offs that exist between different goals. The study emphasises the need to use correlation analysis to identify links and improve policy and intervention alignment. This foundational work supports the notion that understanding semantic linkages in text data can help discover hidden links between SDGs.

Naudé & Vinuesa (2021) look at how the COVID-19 epidemic affects data availability and digital inequality, emphasising how a lack of data prevents informed decisions. Their findings underscore the importance of inclusive AI models that work successfully across different degrees of internet access, driving the development of tools capable of interpreting heterogeneous and imperfect textual material for SDG alignment.

Kroll, Warchold, & Pradhan (2019) investigate whether the endeavours effectively changed SDG trade-offs into synergies. The authors conclude that, even with awareness, the execution of procedures continues to confront problems in consistently achieving synergy. These findings highlight the importance of smart systems capable of recognising both conflicting and complementary linkages among SDGs through automated analysis.

Fonseca, Domingues, & Dima (2020) proposed a paradigm for mapping SDG links using content analysis and systemic thinking. The study presents graphical models that depict the intensity and direction of interactions between goals, emphasising the need to visualise and analysing textual links, which is a key component of the current project's classification approach.

Van Roy et al. (2021) provide a comprehensive assessment of national AI policies in Europe, focusing on ethical considerations, trust, and governance. This provides a policy-oriented environment for bringing AI into sustainability frameworks, which strongly aligns with the current project's goal of providing explicit and accountable NLP tools for SDG categorisation.

Wang et al. (2019) provide an overview of adversarial attacks in machine learning, focussing on flaws in security-sensitive applications. Their contributions are critical to this research, which seeks to develop robust classification models capable of withstanding manipulation or misclassification in actual applications.

Tripathi et al. (2021) emphasise the need for robustness and reliability in data-driven knowledge discovery for manufacturing and production, citing common model deployment issues. The study supports the inclusion of hyperparameter tuning and model optimisation as project objectives to ensure consistent performance across several settings.

Montes & Goertzel (2019) urge for distributed and democratised AI systems to solve centralised biases and limitations. Their vision encourages the development of web-based tools that are open, accessible, and capable of incorporating multiple inputs for interpreting the SDGs — a concept that is included in the goals of this project's web application.

De Laat (2018) examines algorithmic decision-making in terms of transparency and accountability, focusing on the use of machine learning with Big Data. His work focuses on the ethical elements of AI, arguing for interpretability and explainability – issues that can be addressed through clear categorisation findings and user interfaces.

Kearns (2017) presents preliminary findings on equitable algorithms for machine learning, proposing frameworks that balance accuracy and fairness across demographic categories. This viewpoint is critical for the current research, which must ensure that SDG classifications do not reflect or spread social biases.

Montiel-Vázquez et al. (2022) propose an explainable AI method to detect empathy in short texts, using traditional NLP techniques combined with machine learning classifiers like Random Forest and SVM. They incorporate SHAP values to enhance interpretability, making it clear which linguistic features contribute to empathetic labeling. This is especially relevant in fields like counseling, education, and healthcare, where understanding emotional tone is critical.

Gao et al. (2022) address the challenge of analyzing sentiment in short texts by introducing a hybrid CNN–BiGRU model. The CNN captures local patterns, while the BiGRU handles context over sequences, improving the detection of sentiment related to specific aspects. This method is useful for applications such as feedback analysis, where context is minimal, but precision is essential.

Tan et al. (2023) review sentiment analysis techniques, covering both traditional and deep learning approaches. They highlight the need for multilingual and explainable sentiment models, particularly for use in diverse social and cultural settings. Their discussion of domain adaptation and low-resource languages is particularly relevant for scaling sentiment tools in global applications.

Building on these insights and the identified limitations in existing models and implementations, there is a clear need to develop more adaptable, transparent, and context-sensitive tools. This necessity directly informs the direction of our proposed intervention.

2.1 Proposed Solution

Given the analysis of the problems and limitations of the current state of the art, it is proposed to redesign the classification algorithm and the website with some modifications to improve the statistics obtained and to create a website with more detailed information and new functionality.

To achieve this, before training the models, it is proposed to perform a text augmentation process in the categories with a total data size of less than 500 each. This is to see how the algorithms and transformers perform with artificial data created from the original data.
In addition, it is proposed to use the same embedding generation model, since this was the one that obtained the best results in the previous study, but to perform a new hyperparameter search and layer selection for the neural network that classifies the texts.
Finally, it was decided to create a new website that would display the two main SDG categories that the algorithm says the entered text refers to, with a bar chart showing the total probability of each SDG topic, and a new feature that would allow multiple texts contained in the same file (PDF or Excel) to be classified.
To translate this design vision into a functional and scalable tool, a structured development process was initiated. This process not only focuses on enhancing the algorithmic accuracy but also on ensuring usability, interoperability, and scalability of the tool across various use cases.

3. Solution Development

This section describes the implementation of the tool, starting with the processing of the supplied data, then explaining and detailing the text augmentation process and selection, model training, and hyperparameter selection.

3.1 Pre-Processing

First, an exhaustive analysis of the data was carried out, including its quality, content, variables, and, in general, data preprocessing. Initially, it can be observed that the data provided contains several columns, which are: text, Spanish_texts, sdg, labels_negative, labels_positive, agreement, and large. Of these variables, it was decided to select only the first three, that is, the texts in both languages (Spanish and English) and their respective categories, since the other columns do not contain representative information for the work's objectives.
Second, since the data was previously extracted and translated, there is a possibility that it may not be completely clean for processing and training the models. Therefore, validation of the respective text languages was performed, as well as a review of special characters. For this, two different libraries were used. The first, called langdetect, provided by pypi, allowed texts to be filtered in the languages in which they were written, allowing them to be verified and, if necessary, deleted. Only texts in English and Spanish were left in their respective columns. The second, called ftfy, allowed all Spanish texts that were not encoded properly to be corrected. That is, there were texts in which, probably due to data export, the vowels with accents were not read correctly. For example, instead of an 'ó', an 'Ã³' appeared. All these semantic problems were solved.

3.2 Text Augmentation

As mentioned previously, one of the biggest difficulties of the existing algorithm was classification into categories with limited data. Therefore, the first challenge was to find a way to generate artificial data for some SDG categories that were sufficiently similar in content to the original texts, without being an identical or near-identical copy. To achieve this, it was decided to perform the text augmentation process in all categories with a total data volume of less than 500, that is, SDGs 9, 10, 12, and 15.
The first option attempted was the use of open-source libraries such as nlpaug (Tripathi et al., 2021). This library has several variations depending on the type of algorithm used and the language. This library has three types of augmentations, each with different algorithms: character augmentation, word augmentation, and sentence augmentation. This work implements the latter two.
There are different algorithms for word augmentation. However, due to the nature of the library, most of them work only with English texts, so depending on the algorithm chosen, the texts used are either Spanish or English. The first algorithm used was word augmentation with synonyms. This obtains words from the original text and replaces them with synonyms from WordNet, a lexical database in English and other languages. For this algorithm, thanks to WordNet, it was possible to select the preferred language.
When using this algorithm, initially with short texts and examples, it was evident that the generated texts had a significant variation from the original. For example, an example text "The quick brown fox jumps over the lazy dog" was replaced by the following augmented text: "The speedy brown fox jumps over the inattentive dog." However, when using the texts provided for the SDGs, it was noticeable that, since they were longer texts, the difference between them was not usually very large. Typically, only the main verb of the sentence was replaced and substituted with a synonym, for example (Table 1):

Table 1: Comparison of augmented texts with NLPAUG

Original	Synonym Augmenter
SDG target 9.c is highly unlikely to be achieved by 2020. In practice, it is virtually impossible to enjoy the internet effectively through a 2G connection. Only 76% of the world's population has access to a 3G signal, and only 43% has access to a 4G connection. Thus, most of the world remains unconnected, most of it in developing countries.	SDG target 9.c is highly unlikely to be achieved by 2020. In practice, it is virtually impossible to enjoy the internet efficiently over a 2G connection. Only 76% of the world's population has access to a 3G signal, and only 43% has access to a 4G connection. Thus, most of the world remains unconnected, mostly in developing countries.

Due to the high similarity of the generated texts with the augmented texts, it is not advisable to use this technique to generate new data. Having many similar fragments will cause the pre-trained models to generate similar embeddings, generating identical data for training the neural network. This results in biases in the representations and overfitting of the model. In other words, augmenting such similar data is more harmful than beneficial.
The set of algorithms can be grouped into three actions: inserting, changing, or deleting words. This can be done in different ways. This can be done through TF-IDF similarity insertion (a formula that calculates the weight of a word in a document (Montes & Goertzel, 2019), replacing words with their antonyms, removing random words, or inserting words based on their contextual similarity in embeddings using BERT or DistilBERT, etc. Despite implementing all the algorithms, the results obtained were like those previously presented and therefore were not considered optimal enough to meet the proposed objectives. Therefore, the OpenAI API is proposed.
The ChatGPT API is an application programming interface that allows users to experience an Artificial Intelligence-based chatbot (De Laat, 2018). Based on these API features, it was decided to develop a feature that allows obtaining the original texts from the categories selected for augmentation and, using a prompt (a set of words that trigger content generation through Artificial Intelligence software (Kearns, 2017), generating these new artificial texts with similar content, but not the same.

As mentioned before, the prompt could be defined as an "order" that the AI interprets and generates content (sentences, words, paragraphs) based on this prescription. To generate these artificial texts, the following prompt is passed to the API in Spanish using the davinci-002 engine (Montiel-Vázquez et al., 2022):
"Given the following text in Spanish: 'original_text', generate a new text similar in theme and content. You must not generate the same text."
Where 'original_text' is the original data/text provided for this work. This is done iteratively, in which the original texts are selected in an orderly manner until the sum of the total amount of data (original and augmented) is greater than 500. This ensures that the source of each augmented data is a different text. Thanks to this tool, it was possible to obtain new data whose content and theme are like the original, without being a copy of it. An example is the following (Table 2):

Table 2: Comparison of augmented texts with OpenAI

Original	OpenAI API	ODS
Disparities in industrial productivity between rich and poor nations remain stark. For example, the MVA per capita was only $114 in LDCs, compared to $4,938 in Europe and North America, in 2018. With a relatively small amount of capital investment and a predominantly local resource base, small-scale industries generate a substantial amount of employment and self-employment. However, one of the biggest challenges these industries face is access to loans or lines of credit for day-to-day business activities. Adequate financing is crucial for these industries to grow, enabling them to innovate, improve efficiency, expand into new markets, and create new employment opportunities. Although 31.5% of small-scale industries (manufacturing and services) worldwide benefit from loans or lines of credit, regional differences are notable.	Economic inequalities between different nations are also present in the industrial sphere. The value added of production per capita was $114 in the Least Developed Countries in 2018, while in Europe and North America it was $4,938. Since many of these industries are small-scale, they generate significant employment and self-employment. However, one of the major challenges they face is access to loans and financing. These solutions allow them to innovate, improve efficiency, expand into other markets, and generate employment.	9

As can be seen in Table 2, the texts generated by the OpenAI API attempt to convey the same meaning as their predecessor. However, we can see that the differences between the two texts are very significant, with only data and proper nouns generally retained. However, we can see a limitation of these texts: when using the API, it is necessary to specify the maximum number of tokens in the generated text (max_tokens parameter). Despite setting this value equal to the number of words in the original text, the texts are often truncated, as is the case in the example in Table 1, where the final paragraph fails to paraphrase it. Despite this limitation, this does not prevent the models from being trained or generating any type of over-specialization, so it was decided to work with this API to perform data augmentation. Upon completion of the data augmentation, the following data (Table 3) were obtained:

Table 3: Total number of original and augmented texts per SDG category

ODS	Original	Augmented
1	787	787
2	567	567
3	1425	1425
4	1635	1635
5	1693	1693
6	1071	1071
7	1233	1233
8	683	683
9	439	550
10	369	550
11	940	940
12	247	490
13	726	726
14	579	579
15	406	500
16	1669	1669
Total	14469	15098

The augmentation process aimed to balance underrepresented categories by generating synthetic data until each reached at least 500 samples. However, SDG 12 fell slightly short of this goal due to the limited number of unique original texts available for augmentation. Since the augmentation procedure maintained a 1:1 mapping between original and synthetic samples to ensure diversity, the process was capped when the original dataset was exhausted.

3.3 Generating Embeddings with the Pre-Trained Model

Now that we have a new dataset with some augmented categories, the next step is to generate embeddings for each of the data/texts. As mentioned in the theoretical framework, an embedding is basically a numerical representation of a text that is easy for a machine to understand. For this purpose, we used the model selected in the thesis on which this work is based. This work compared the pre-trained models with their respective results, measuring metrics and other information to select the best model for this data.
The final decision was to use the pre-trained Google DistilRoBERTa model with the Hugging Face (Transformers) library, using the AutoTokenizer and Automodel methods, and finally manually perform pooling on the last hidden layer of the model, which results in a 768-dimensional vector (Tvaronavičienė et al., 2020). After properly transforming all the texts into their corresponding embeddings, the data was properly separated into training, validation, and testing. This was done through the following distribution: 64% training, 16% validation, and 20% testing.

3.4 Training and Validation

Once the embeddings were obtained and the dataset was separated to obtain the training data, the neural network responsible for text classification was trained. A 5-layer neural network was used (Figure 1): Input, Dropout, LSTM, GlobalMaxPooling1D and Output (Dense Layer), so that the architecture will be used to begin testing and training the model.

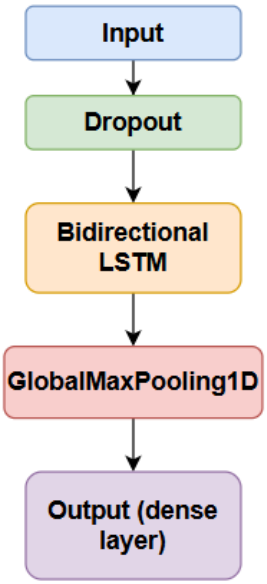

Fig. 1: Initial neural network

First, keep in mind that the input is a layer of 768 neurons, because the embeddings have the same dimension; that is, each dimension of the embedding is a neuron in the input layer. Second, since we are classifying into 16 categories corresponding to each SDG, the output layer has 16 neurons, each of which represents the probability (value between 0 and 1) that the input embedding corresponds to that SDG. For training, the data set 'X' (input features) is the generated embeddings, and the data set 'Y' (output label) is a vector in which all dimensions are 0 except for the one corresponding to the text category, which has a value of 1.

Training begins using a neural network like the one presented in the previous thesis, with some variations in the callback, minimum delta, and patience functions. However, it was decided to add a second LSTM layer so that the dimensionality reduction between layers would not be as large. To understand this, it is important to understand that we start with a layer of 768 neurons and end with a layer of 16 neurons, so the intermediate layers must reduce the dimensionality of the data.

After adding the second LSTM layer to the neural network after the first, a hyperparameter search is performed for the units that each of these layers should have. This search is performed using the Grid Search method, which allows the search for the best combination of results among all possible combinations. The DropOut layer percentage was also added to this search. This layer randomly "turns off" some of the neurons (depending on the given percentage) so that the model does not overspecialize. The research showed that the first LSTM layer should have 512 neurons, the second should have 128, and the DropOut layer should have a percentage of 0.3.

After a second training session, using the values obtained from the hyperparameter search, and performing accuracy validation and model validation during the training epochs, it became evident that, despite having a DropOut layer intended to prevent the model from overspecializing, this error was still occurring (Figures 2-3). This was evident because the values of these metrics, as the training and validation epochs progressed, diverged. Therefore, it was decided to implement a second DropOut layer, in addition to adding a kernel regularizer and a recurrent regularizer in the bidirectional layers (LSTM). In addition, the Adam optimizer was added with a learning rate also found with the Grid Search. After making these adjustments, it was possible to obtain a model that was not overspecialized and generally maintained the accuracy values obtained in the previous work.

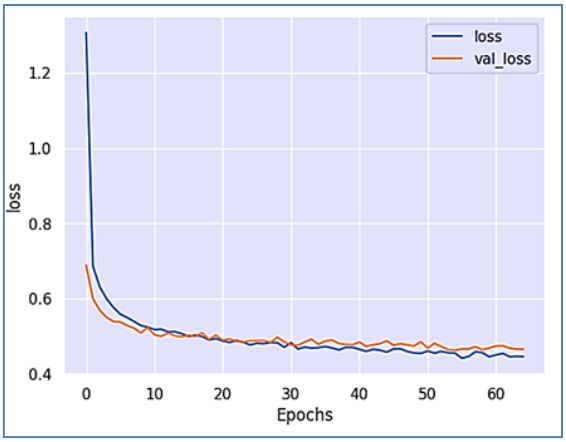

Fig. 2: Values of the accuracy metric for the training and validation data across epochs.

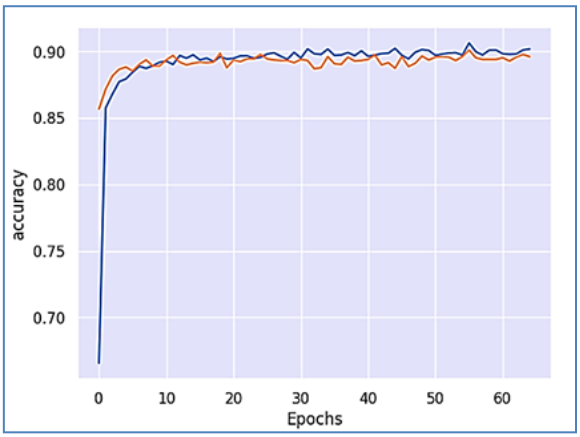

Fig. 3: Values of the accuracy metric for the training and validation data across epochs.

After obtaining this model with these validation metrics, the architecture and final model were saved (Figure 4).

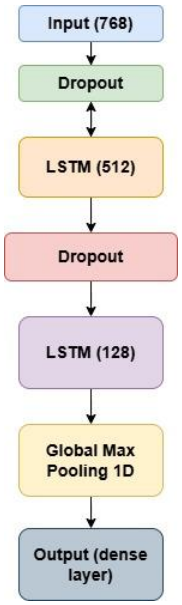

Fig. 4: Architecture of the final neural network.

3.5 Web Tool Construction

Finally, it was decided to build a new web page that would allow users to use the final model to classify texts. For this, the Streamlit tool was used. It has predefined interface components and facilitates implementation.

The first feature allows users to enter text, and the model will classify it. In the interface, next to the text, you can see the two categories most likely to belong to this text, along with their respective images. Additionally, on the right side, a bar chart shows the probability of the text falling into the two categories obtained. Figure 5 classifies the following text:

Clan interests often take precedence over the interests of individual victims, and families choose to reconcile through the customary system rather than seeking redress for the victims. This leads to female rape victims being forced to marry the rapist, following the dictates of village elders and applying customary practices (A/HRC/20/16/Add.3). In Ghana, traditional authorities, such as tribal chiefs in many rural areas, govern issues and disputes related to land and property rights, as well as matters involving "supernatural interference," including allegations of witchcraft. In Afghanistan, Sharia, customary law, the formal and secular legal system, and international law exist in parallel.

Fig. 5: Classification of a sample text on the web interface

As can be seen in the text, it can address various Sustainable Development Goals, given that it touches on several topics. Although this text is labeled SDG 5, the model's primary classification is SDG 16, and the second possible classification is SDG 5. If we pause to read the

text, we can see that it addresses both gender equality and justice and peace, so, understandably, the model has a roughly 50% probability of belonging to these two classifications.
The second feature of the website includes uploading a file (PDF or Excel) and generating classifications from it. This feature is useful when you have many texts to classify and don't have to manually insert each one. A PDF file was imported containing a table whose first column contains the texts to be classified. Below the Browser button, the same table as the PDF will be displayed with an additional column representing the classifications generated by the model. This table can be downloaded as a CSV.
The current implementation has been tested primarily on preprocessed, bilingual (English and Spanish) text aligned with the SDG framework. As such, its performance may degrade when applied to raw, unstructured, or noisy textual inputs, especially those involving other languages, dialects, or domain-specific terminology not represented in the training data.
Having completed the model training and system development phases, it was essential to rigorously evaluate the tool's effectiveness using a portion of data withheld during training. This evaluation aimed to assess how well the model generalizes to unseen data and to identify potential areas for further refinement.

4. Testing and Results

After obtaining the final neural network, the test data (20%) separated before training was used to review the accuracy of the resulting model, yielding the following results (Table 4):

Table 4: Results of the generated model

ODS	Precision	Recall	F1-score
1	0.87	0.78	0.82
2	0.84	0.9	0.87
3	0.94	0.95	0.94
4	0.95	0.98	0.96
5	0.95	0.96	0.96
6	0.93	0.93	0.93
7	0.93	0.94	0.93
8	0.76	0.74	0.75
9	0.82	0.89	0.85
10	0.78	0.67	0.72
11	0.89	0.89	0.89
12	0.91	0.89	0.9
13	0.9	0.85	0.88
14	0.95	0.97	0.96
15	0.91	0.93	0.92
16	0.96	0.98	0.97
accuracy	0.9115	0.912	0.9115
macro avg	0.89	0.89	0.89
weighted avg	0.91	0.91	0.91

In Table 4, you can see that most categories have precision and recall values above 0.9. This means that the model is quite accurate in classifying these categories. Furthermore, the overall accuracy was 91.15%.
You can also see that the model's overall recall is 89.21%. That is, of the original SDG classifications, the model correctly classifies this percentage. Finally, you can also see that the overall accuracy was 89.52%. That is, of the classifications provided by the model, this percentage is correct. Thanks to this, we can say that the model has an accurate rate of approximately 90%.
However, one of the goals of this work is to see how text augmentation can change the metrics for categories whose total data amounts are very small compared to the others. For this purpose, I will use the metrics obtained in Wilches's work (Tvaronavičienė et al., 2020) to compare the results obtained (Table 5):

Table 5: Comparison of results obtained without data augmentation

ODS	Precision Without Aug	Precision	Recall Without Aug	Recall
1	0.88	0.87	0.80	0.78
2	0.87	0.84	0.85	0.9
3	0.92	0.94	0.95	0.95
4	0.95	0.95	0.95	0.98
5	0.93	0.95	0.96	0.96
6	0.90	0.93	0.94	0.93
7	0.90	0.93	0.92	0.94
8	0.79	0.76	0.65	0.74
9	0.71	0.82	0.84	0.89
10	0.59	0.78	0.70	0.67
11	0.89	0.89	0.88	0.89
12	0.78	0.91	0.71	0.89
13	0.88	0.9	0.81	0.85
14	0.96	0.95	0.96	0.97
15	0.88	0.91	0.87	0.93
16	0.95	0.96	0.96	0.98
accuracy	No Data	0.9115	0.89	0.912

In this last Table 5, we can see that most categories improved compared to the previous one. Furthermore, the underlined SDGs were the ones that underwent data augmentation. Specifically, in these SDGs, we can see a significant improvement in precision and recall, especially in SDG 10, which increased from 0.59 to 0.78. That is, an increase of almost 20%. Despite this, we can also see that some categories saw a decrease in performance. However, in none of the cases was there a deterioration greater than 3 percentage points; that is, it is not very significant compared to obtaining better metrics in all the others, especially in those that were augmented.

The developed SDG classification tool holds promise for interdisciplinary applications, particularly in sociology and linguistics. In sociology, it can support large-scale analysis of policy texts, civil society reports, and organizational communications to trace thematic emphases, shifts in discourse, or the framing of development priorities across time and context. In linguistics, its bilingual architecture enables studies of semantic consistency, translation dynamics, and institutional language use across English and Spanish texts. These applications can aid corpus-based research in global governance, discourse analysis, and cross-cultural semantics. However, real-world deployment must account for limitations, including reduced accuracy on underrepresented dialects, culturally specific phrasing, and discursive practices such as metaphors or code-switching, which may not be well captured by current models. Broader linguistic inclusion and validation with diverse, multilingual datasets are essential for enhancing its relevance across social science disciplines.

The developed tool shows great potential for supporting SDG monitoring in linguistically and regionally diverse contexts. Through language detection, multilingual text processing, and domain-specific augmentation strategies, its content can be divided into Spanish and English, and language diversity and resource constraints are often used as globalization when reporting the SDGs. This function is particularly important for interdisciplinary research in sociology and linguistics, where it is important to analyze how sustainable development are represented in cultural and textual landscapes. Nevertheless, future adaptations should address underrepresented languages and regional dialects to increase robustness in real-world application scenarios.

Despite the effectiveness of this tool, there are several challenges that need to be addressed. Using the OpenAI API introduces token constraints that reduce generated texts to compress knowledge, sometimes reducing semantic completeness. Although the model is supported in Spanish and English, extending multilingual processing to underrepresented languages remains a major obstacle, particularly for more comprehensive SDG monitoring across linguistically diverse regions. In addition, user adoption may be hampered by technical requirements, such as understanding model outputs or preprocessing file formats, which may limit access for non-technical stakeholders. Addressing these limitations will be necessary to increase the utility of the tool in practical and interdisciplinary contexts.

Future research could focus on non-English texts such as French, Arabic, or local languages to assess its generalizability across linguistically diverse settings. Another direction is to integrate the model with global SDG tracking platforms (e.g., the UN SDG Data Hub or local policy dashboards) that enable automated analysis of qualitative reports and documents. These extensions will improve sociolinguistic understanding of sustainability narratives and improve operational monitoring of SDG progress in underrepresented regions.

5. Conclusions

There is evidence of a complete improvement in the different metrics (precision and recall) in the four categories in which data augmentation was performed (9, 10, 12, and 15). Outperforming previous models that did not implement text augmentation strategies to balance classes.

The model achieved an overall accuracy of 91.15%. This demonstrates the high effectiveness of the combination of the different technologies used (text augmentation, pre-trained models, and neural networks).

It is concluded that, for multi-word texts, open-source libraries are not the best options for data augmentation, given that the generated texts do not vary by more than 4 words from the original text. Therefore, for these cases, it is better to use more sophisticated tools and APIs such as OpenAI.

Given the limitation seen in the OpenAI tool in being able to use the entire original text to generate new texts, without cutting the last sentence, it is recommended to work on data augmentation with other more advanced models from the same tool, such as Davinci-003 or GPT 3.5, which generate more robust texts at a higher cost, possibly without the maximum token limitation.

Because it was necessary to include different methods and technologies to avoid over-specialization in the final neural network, it is also recommended to perform a more in-depth analysis of its architecture to obtain better results with data unknown to the model (not used in training).

References

[1] Nilsson, M., Chisholm, E., Griggs, D., Howden-Chapman, P., McCollum, D., Messerli, P., Neumann, B., Stevance, A.-S., Visbeck, M., & Stafford-Smith, M. (2018). Mapping interactions between sustainable development goals: Lessons learned and ways forward. *Sustainability Science*, 13, 1489–1503.

[2] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9, e1312.

[3] Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: The System Causability Scale (SCS). *Künstliche Intelligenz*, 34, 193–198.

[4] Chou, Y. L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81, 59–83.

[5] Holzinger, A., Malle, B., Saranti, A., & Pfeifer, B. (2021). Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Information Fusion*, 71, 28–37.

[6] Corbett-Davies, S., & Goel, S. (2018). Measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint*, arXiv:1808.00023.

[7] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104, 671.

[8] Leszczynski, A., & Zook, M. (2020). Viral data. *Big Data & Society*, 7, 2053951720971009.

[9] Tvaronavičienė, M., Plėta, T., Della Casa, S., & Latvys, J. (2020). Cyber security management of critical energy infrastructure in national cybersecurity strategies: Cases of USA, UK, France, Estonia and Lithuania. *Insights into Regional Development*, 2, 802–813.

[10] Limba, T., Plėta, T., Agafonov, K., & Damkus, M. (2017). Cyber security management model for critical infrastructure. *Entrepreneurship and Sustainability Issues*, 4, 559–573.

[11] Pradhan, P., Costa, L., Rybski, D., Lucht, W., & Kropp, J. P. (2017). A systematic study of sustainable development goal (SDG) interactions. *Earth's Future*, 5, 1169–1179.

[12] Naudé, W., & Vinuesa, R. (2021). Data deprivations, data gaps and digital divides: Lessons from the COVID-19 pandemic. *Big Data & Society*, 8, 20539517211025545.

[13] Kroll, C., Warchold, A., & Pradhan, P. (2019). Sustainable Development Goals (SDGs): Are we successful in turning trade-offs into synergies? *Palgrave Communications*, 5, 140.

[14] Fonseca, L. M., Domingues, J. P., & Dima, A. M. (2020). Mapping the sustainable development goals relationships. *Sustainability*, 12, 3359.

[15] Van Roy, V., Rossetti, F., Perset, K., & Galindo-Romero, L. (2021). *AI Watch—National Strategies on Artificial Intelligence: A European Perspective*. EUR 30745 EN; Publications Office of the European Union: Luxembourg.

[16] Wang, X., Li, J., Kuang, X., Tan, Y. A., & Li, J. (2019). The security of machine learning in an adversarial setting: A survey. *Journal of Parallel and Distributed Computing*, 130, 12–23.
[17] Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M., & Emmert-Streib, F. (2021). Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing. *Frontiers in Artificial Intelligence*, 4, 22.
[18] Montes, G. A., & Goertzel, B. (2019). Distributed, decentralized, and democratized artificial intelligence. *Technological Forecasting and Social Change*, 141, 354–358.
[19] De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & Technology*, 31, 525–541.
[20] Kearns, M. (2017). Fair algorithms for machine learning. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, Cambridge, MA, USA, 26–30 June.
[21] Montiel-Vázquez, E. C., Ramírez Uresti, J. A., & Loyola-González, O. (2022). An explainable artificial intelligence approach for detecting empathy in textual communication. *Applied Sciences*, 12, 9407.
[22] Gao, Z., Li, Z., Luo, J., & Li, X. (2022). Short text aspect-based sentiment analysis based on CNN + BiGRU. *Applied Sciences*, 12, 2707.
[23] Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13, 4550.