RGSA – Revista de Gestão Social e Ambiental
REVISTA DE GESTÃO SOCIAL E AMBIENTAL

RGSA – Revista de Gestão Social e Ambiental
ISSN: 1981-982X
Submission date: 08/23/2024
Acceptance date: 10/25/2024
DOI: https://doi.org/10.24857/rgsa.v18n12-039
Organization: Interinstitutional Scientific Committee
Chief Editor: Ana Carolina Messias de Souza Ferreira da Costa
Assessment: Double Blind Review pelo SEER/OJS

ODSBAHIA-PTBR: A NATURAL LANGUAGE PROCESSING MODEL TO SUPPORT SUSTAINABLE DEVELOPMENT GOALS

Êmeris Silva Santos [1]
Leonardo Evangelista Moraes [2]

ABSTRACT

Objective: The present study aims to propose an approach for the objective classification of texts in Portuguese in relation to the Sustainable Development Goals (SDGs) of Brazil's 2030 Agenda.

Theoretical Framework: The study uses natural language processing (NLP) techniques with deep learning, using pre-trained models such as BERTimbau Base, DeBERTinha and Albertina. In addition, it considers the existing gaps in the literature regarding the classification of texts in Portuguese related to the 17 UN SDGs and also including three new SDGs proposed in the document Guide Agenda 2030: Integrating SDGs, Education and Society, prepared in 2020 in partnership between UnB and UNESP, SDGs 18 (Ethnic-Racial Equality), 19 (Art, Culture and Communication) and 20 (Rights of Indigenous Peoples and Traditional Communities).

Method: La investigación es exploratoria, descriptiva y aplicada, con enfoque cuantitativo y procedimientos experimentales. Los modelos previamente entrenados se ajustaron al conjunto de datos de etiquetas múltiples creado específicamente para la tarea. La Base BERTimbau presentó el mejor rendimiento y se utilizó como base para la creación del modelo ODSBahia-PTBR, evaluado con métricas como precisión (82%), recuerdo (72%) y F1-Score (77%).

Results and Discussion: El ODSBahia-PTBR logró una precisión del 95% al traducir y clasificar el conjunto de datos OSDG. Los resultados ponen de manifiesto la efectividad del modelo en la identificación y categorización de textos alineados con los ODS, siendo especialmente relevante para el seguimiento de las interseccionalidades entre los ODS propuestos.

Research Implications: The SDGbahia-PTBR model has practical implications by offering an innovative tool for different stakeholders to monitor and analyze initiatives aligned with the SDGs, contributing to the evaluation and promotion of the 2030 Agenda.

Originality/Value: This research is a pioneer in including SDGs 18, 19 and 20 in Portuguese-language text classifiers, offering an unprecedented and applicable approach to sustainable monitoring in Brazil and other Portuguese-speaking countries.

Keywords: Agenda 2030, Natural Language Processing, Artificial intelligence, Sustainability.

ODSBAHIA-PTBR: UM MODELO DE PROCESSAMENTO DE LINGUAGEM NATURAL PARA APOIAR OS OBJETIVOS DE DESENVOLVIMENTO SUSTENTÁVEL

RESUMO

Objetivo: O presente estudo tem como objetivo propor uma abordagem para a classificação objetiva de textos em português em relação aos Objetivos de Desenvolvimento Sustentável (ODS) da Agenda 2030 do Brasil.

Referencial Teórico: O estudo utiliza técnicas de processamento de linguagem natural (PLN) com aprendizado profundo, utilizando modelos pré-treinados como BERTimbau Base, DeBERTinha e Albertina. Além disso, considera as lacunas existentes na literatura quanto à classificação de textos em português relacionados aos 17

[1] Universidade Federal do Sul da Bahia, Porto Seguro, Bahia, Brasil. E-mail: emerissantos@gmail.com
Orcid: https://orcid.org/0009-0007-8893-9424
[2] Universidade Federal do Sul da Bahia, Porto Seguro, Bahia, Brasil. E-mail: leomoraes@ufsb.edu.br
Orcid: https://orcid.org/0000-0002-6198-0618

ODS da ONU e incluindo também três novos ODS propostos no documento Guia Agenda 2030: Integrando ODS, Educação e Sociedade, elaborado em 2020 em parceria entre a UnB e a UNESP, os ODS 18 (Igualdade Étnico-Racial), 19 (Arte, Cultura e Comunicação) e 20 (Direitos dos Povos Originários e Comunidades Tradicionais).

Método: A pesquisa é exploratória, descritiva e aplicada, com abordagem quantitativa e procedimentos experimentais. Foram ajustados modelos pré-treinados ao conjunto de dados multirrótulo criado especificamente para a tarefa. O BERTimbau Base apresentou o melhor desempenho e foi utilizado como base para a criação do modelo ODSBahia-PTBR, avaliado com métricas como precisão (82%), recall (72%) e F1-Score (77%).

Resultados e Discussão: O ODSBahia-PTBR alcançou 95% de acurácia ao traduzir e classificar o conjunto de dados OSDG. Os resultados destacam a eficácia do modelo na identificação e categorização de textos alinhados aos ODS, sendo particularmente relevante para o monitoramento das interseccionalidades entre os ODS propostos.

Implicações da Pesquisa: O ODSBahia-PTBR apresenta implicações práticas ao oferecer uma ferramenta inovadora para diferentes partes interessadas monitorarem e analisarem iniciativas alinhadas aos ODS, contribuindo para a avaliação e promoção da Agenda 2030.

Originalidade/Valor: Esta pesquisa é pioneira ao incluir os ODS 18, 19 e 20 em classificadores de texto em língua portuguesa, oferecendo uma abordagem inédita e aplicável ao monitoramento sustentável no Brasil e em outros países lusófonos.

Palavras-chave: Agenda 2030, Processamento de Linguagem Natural, Inteligência Artificial, Sustentabilidade.

ODSBAHIA-PTBR: UN MODELO DE PROCESAMIENTO DE LENGUAJE NATURAL PARA APOYAR LOS OBJETIVOS DE DESARROLLO SOSTENIBLE

RESUMEN

Objetivo: El presente estudio tiene como objetivo proponer un enfoque para la clasificación objetiva de los textos en portugués en relación con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030 de Brasil.

Marco Teórico: El estudio utiliza técnicas de procesamiento del lenguaje natural (PLN) con aprendizaje profundo, utilizando modelos preentrenados como BERTimbau Base, DeBERTinha y Albertina. Además, considera los vacíos existentes en la literatura en cuanto a la clasificación de los textos en portugués relacionados con los 17 ODS de la ONU y también incluye tres nuevos ODS propuestos en el documento Guía Agenda 2030: Integración de los ODS, Educación y Sociedad, elaborado en 2020 en asociación entre la UnB y la UNESP, los ODS 18 (Igualdad étnico-racial), 19 (Art, Cultura y Comunicación) y 20 (Derechos de los Pueblos Indígenas y las Comunidades Tradicionales).

Método: La metodología adoptada para esta investigación comprende [describir de manera concisa el diseño del estudio, incluido el enfoque, los participantes, los instrumentos, los procedimientos, etc.]. La recolección de datos se realizó mediante [explicar los métodos específicos utilizados, como entrevistas, cuestionarios, observaciones, entre otros].

Resultados y Discusión: Los resultados obtenidos revelaron [sintetizar los principales resultados de la investigación]. En la sección de discusión, estos resultados se contextualizan a la luz del marco teórico, destacando las implicaciones y relaciones identificadas. En este apartado también se consideran posibles discrepancias y limitaciones del estudio.

Implicaciones de la investigación: SDGbahia-PTBR tiene implicaciones prácticas al ofrecer una herramienta innovadora para que diferentes actores monitoreen y analicen iniciativas alineadas con los ODS, contribuyendo a la evaluación y promoción de la Agenda 2030.

Originalidad/Valor: Esta investigación es pionera en la inclusión de los ODS 18, 19 y 20 en los clasificadores de texto en portugués, ofreciendo un enfoque inédito y aplicable al monitoreo sostenible en Brasil y otros países de habla portuguesa.

Palabras clave: Agenda 2030, Procesamiento del Lenguaje Natural, Inteligencia artificial, Sostenibilidad.

1 INTRODUCTION

In 2015, United Nations (UN) member states adopted the 2030 Agenda, a global action plan structured around 17 Sustainable Development Goals (SDGs) and 169 associated targets (UN, 2015). These goals and targets cover the three essential dimensions of sustainable development: economic, social and environmental. Five crucial areas are emphasized—people (SDGs 1–6), prosperity (SDGs 7–12), planet (SDGs 13–15), peace (SDG 16) and partnership (SDG 17) (Morton, 2015). *et al* ., 2017). The 2030 Agenda sets an ambitious deadline, stipulating that such goals must be achieved by the year 2030 (UN, 2015).

However, to implement this sustainability agenda, a global partnership involving governments, the private sector and civil society is necessary, mobilizing all available resources, as these actions are focused on people, the planet and prosperity (UN, 2015). It is essential that different sectors and stakeholders collaborate to achieve the SDGs, since the results of the SDGs are interconnected and influence each other, involving human, technical and natural systems.

The 2030 Agenda has the motto “Leave no one behind” and, in some of its goals, addresses issues of social justice, cultural diversity and combating discrimination. However, according to Cabral and Gehre (2020), the 17 SDGs do not specifically and visually address some population groups that are representative of the Brazilian and Latin American realities.

From this perspective, the authors propose to include in the 2030 Agenda three new objectives "Racial Equality (SDG 18)", "Art, Culture and Communication (SDG 19)", and "Rights of Original Peoples and Traditional Communities (SDG 20)", aiming give visibility to groups that have been historically invisible and neglected in global development agendas, ensuring that important aspects for local development and sustainability are addressed.

While proposals like this seek to fill gaps and expand the scope of the 2030 Agenda, they also underscore a fundamental issue that underlies the implementation of all SDGs: the need for reliable and accessible data.

There are still significant data and knowledge gaps in tracking policies and inputs to monitor progress towards the SDGs and assess the impact of implemented initiatives. These include the lack of consistent and comprehensive data, inequalities in the availability of information across regions and population groups, and the lack of standardization and interoperability in data collection systems. These challenges make it difficult to assess progress towards the SDGs, especially when measuring qualitative aspects.

Furthermore, people with limited knowledge of the SDGs may face difficulties when trying to correlate their local activities to the broader context of the SDGs, as they do not have a comprehensive understanding of the 2030 Agenda. This correlation is often carried out subjectively, based on individual judgments, which can also lead to inconsistent and biased interpretations.

Therefore, this work has the general objective of presenting an approach that allows correlating local activities with the SDGs in an objective manner, exploring the use and application of Natural Language Processing (NLP) techniques, focusing on the following specific objectives:

a) Build and validate a multi-label training corpus in Portuguese, corresponding to the 17 UN SDGs and three new SDGs, to provide a database that helps improve the evaluation of classification models.

b) Evaluate and compare the performance of the pre-trained models BERTimbau, Albertina, and DeBERTinha, in order to identify which one best adapts to the training corpus.

c) Develop ODSBahia -PTBR, a text classification model adjusted for the Portuguese language, improving the model selected in the previous stage to improve its ability to classify texts related to the SDGs.

d) Analyze the performance of ODSBahia-PTBR in relation to other models that classify texts related to the SDGs, evaluating the results in terms of precision , recall *and* F1 - *score* .

The adjusted model, called ODSBahia-PTBR, is expected to be a tool to autonomously and objectively assess how textual descriptions of activities align with the SDGs, including the new proposed goals (SDGs 18, 19 and 20). This work not only expands the existing literature on the topic, but also offers a practical tool that can assist in monitoring and evaluating progress towards the SDGs.

2 THEORETICAL FRAMEWORK

Artificial Intelligence (AI) is increasingly present and transforming society. In recent years, there has been a significant improvement in several applications, including those related to knowledge representation and decision-making, such as Natural Language Processing (NLP), which involves the engineering of computational models and processes to solve practical problems aimed at understanding human languages (Otter *et al.* , 2021).

NLP studies emerged to simplify user interactions and meet the desire to establish communication with the computer through natural language.

According to Khurana *et al.* (2022), NLP can be classified into two main parts. The first is the Natural Language Understanding (or linguistic) task, which employs several linguistic levels to understand and process human language comprehensively, including phonology, morphology, lexicon, syntax, semantics, pragmatics, and discourse. And the second part is the Natural Language Generation task, which consists of producing sentences and paragraphs within a context.

2.1 ATTENTION-BASED MODELS

Before the rise of attention-based models, the literature points to different technologies and approaches in the field of natural language processing that were used to solve tasks related to text and language processing. These previous technologies and approaches include rule-based models, statistical methods, support vector machines (*SVM) and* other machine learning algorithms, and recurrent neural networks *(RNN*) (Hajikhani & Suominen , 2022).

Despite the advances brought by these technologies, the search for more effective and adaptable models led to the development of the encoder-decoder architecture (Cho *et al.* , 2014), which stood out in natural language processing tasks, such as the automatic translation explored by Sutskever *et al* . (2014).

In this architecture, the encoder processes each item in the input sequence, generating an intermediate representation or context vector *that contains* information about the entire input sequence. After processing the entire input sequence, the encoder sends this vector to the decoder, which in turn uses this information to produce the desired output, be it a translation, a classification, an automatically generated answer, among other possibilities, since this architecture can be applied to a variety of NLP tasks.

However, these traditional models such as RNNs and LSTMs (*Long Short- Term Memory*), face difficulties in dealing with long-term dependencies in large sequences, in addition to incurring computation overhead to perform the decoding operation and time-consuming training due to the problem of gradient fading and exploding. To solve such problems, the attention mechanism (Bahdanau *et al* ., 2014), initially used in machine translation models with RNNs (Luong *et al* ., 2015).

Attention mechanisms are computational methods that mimic human attention, allowing algorithms to focus on specific parts of the input data. In NLP, neural attention detects correlations between text components, mapping the structure of the language and its contextual and ordering relationships.

Hu (2020) presents in his review different variants of the attention mechanism proposed to deal with a variety of tasks in NLP, among them self-attention . The *self* - attention mechanism is used integrally in the *Transformers* neural network architecture , presented in 2017 in the work of Vaswani *et al* . (2017). This mechanism is used to deal with sequences of *tokens* (words), allowing the model to capture complex and long-range relationships between the elements of the input sequence.

self-attention mechanism , each *token* in the input sequence is represented by its respective vectors: query vector (Q), key vector (K) and value vector (V), they are derived from the input vector through separate linear transformations. Then, an unnormalized attention matrix is calculated through the dot products between the vectors Q and K, which represents the similarity between all pairs of *tokens* in the sequence.

Softmax function to obtain attention weights that represent the relative importance of each *token* relative to the other *tokens* in the input sequence. These attention weights are multiplied by the value vectors (V), resulting in a context vector, which is a weighted combination of the input vectors. Self -attention is computed in parallel for all *tokens* in the input sequence, allowing the model to obtain a more comprehensive representation of the *token sequence* in NLP tasks.

2.1.1 Transformers

Transformer is a recent and widely used approach in sequence-to-sequence learning (*seq2seq), based on* self-attention mechanisms that allow learning long-range dependencies between input and output sequences, in addition to offering high parallelization, accelerating training and inference on GPUs or TPUs (Vaswani *et al* ., 2017).

Its core structure includes stacks of encoders and decoders, each composed of self-attention sublayers multifocus , fully connected *feedforward* neural networks , and cross-attention mechanisms that integrate information between input and output. The *Transformer* *architecture* has been applied to various tasks such as language translation, sentiment analysis, text generation and classification, and is used in models such as T5, which transforms all tasks into text-to-text problems, and GPT, which focuses on text generation with only decoder layers. Models such as BERT, on the other hand, exclusively use encoder layers for tasks such as text classification, highlighting the flexibility of the Transformer in different contexts.

2.1.2 Bert

The *Bidirectional Encoder Representations from Transformers* (BERT), developed by Google (Devlin *et al* ., 2019), consists of a stack of encoding layers based on the *Transformer architecture that incorporates specific objectives and a multi-headed* self-attention approach . It is part of the family of foundational models (Bommasani *et al* ., 2022), and which can be adapted to a wide range of tasks. It is available in two main variants: *BERT-Base* (12 layers) and *BERT-Large* (24 layers). Although *BERT-Large* is able to capture more complex relationships between words due to its larger number of layers, it also requires more computational resources for training and inference.

Compared to autoregressive models like GPT, which generate predictions primarily based on the context preceding the current word, and models that only consider the context to the right of a word (*right-to-left models*), BERT uses a bidirectional approach, it is trained to predict words in a sentence given all its contextual information.

BERT is pre - trained on unsupervised learning tasks using large amounts of data (approximately 3.3 billion words extracted from online books and articles) of unlabeled text, allowing it to learn general language representations. *Pre* -training is the dominant approach to transfer learning, where a model is trained on a surrogate task (often just as a means to an end) and then adapted to the task of interest later through fine-tuning (Bommasani et al., 2002). *et al* ., 2022).

During pre-training, BERT is subjected to two unsupervised tasks, Masked Word Prediction *(Language Model (MLM) and Next Sentence* Prediction *Prediction* (NSP). In the MLM task, a percentage of words are randomly masked (replaced with a [MASK] *token*), and the model attempts to predict these words based on the context of neighboring words. The NSP task provides the model with pairs of real and fake sentences, created by combining parts of different sentences, in order to train it to predict whether the second sentence is, in fact, the next sequence of the first and to correctly differentiate them. These tasks allow the model to understand the relationship between different parts of the text and capture broader context information (Islek & Oguducu , 2020).

In fine-tuning , *the* model is trained on specific tasks with labeled data, adjusting the parameters and weights of the neural network to maximize performance on the task at hand (Rogers et al., 2020).

Kaliyar highlighted *et al* . (2021), at this stage there are three important features to consider. First, it is the preprocessing of long text, since *BERT-base* , for example, the most commonly used model, has a maximum sequence length limit of 512 *tokens* . However, there are variants of the BERT architecture with extended capacity to handle longer texts, such as Longformer (Beltaby *et al* ., 2020), which can process sequences of up to 4096 *tokens* . Therefore, when dealing with long texts, it is necessary to ensure that they fit within this limit. Second, it is important to consider layer selection, since the base BERT model consists of an embedding layer and 12 encoder layers; Third, the problem of overfitting , which occurs when a model fits very well on the training data, but has difficulty generalizing to new data.

Indeed, BERT has been applied in a variety of other NLP tasks (Qiu *et al* ., 2021), such as document classification (Adhikari , 2019), recommender systems (Ray *et al* ., 2021), named entity recognition (Virtanen , 2021), and so on. *et al* ., 2019), sentiment analysis (Zhang *et al* ., 2020), and question answering (Yang & Choi, 2019). The success of this architecture is directly linked to its ability to understand the nuances and subtleties of natural language, capturing the context and relationships between words effectively.

When it was published, BERT revolutionized the field of NLP. Since then, other variants have been developed to optimize the model and adapt it to different purposes. Among them, DistilBERT stands out , a compact version of the BERT model that has been reduced in size, while maintaining 97% of its language understanding capacity and being 60% faster (Sanh *et al* . 2020); RoBERTa (Liu *et al* ., 2019) trained on more than 30B words, obtained better results than BERT in question-and-answer and language understanding tasks; XLNet (Yang *et al* ., 2020), trained on 33B words, performed better than BERT in 20 tasks; GPT-3 (Brown *et al* ., 2020) trained on a corpus with approximately 400B words, represents the third iteration of the GPT series of models that gained popularity, especially through the ChatGPT application, standing out for its conversation and text generation capabilities.

Additionally, specialized models such as BioBERT (Lee *et al* ., 2020), SciBERT (Bentagy *et al* ., 2019), BERTabaporu (Da Costa *et al* ., 2023) and BERTweet (Nguyen *et al* ., 2020) have been adapted for specific tasks or domains. These variants maintain the underlying *Transformer architecture* , providing significant advances in several NLP tasks.

In this study, three language models pre-trained on Brazilian Portuguese corpora were employed, each characterized by its specific architecture and adaptation to handle NLP tasks in this linguistic context. They are presented below:

Albertina PTBR 100M (Rodrigues *et al* ., 2023): This model was trained on a selection of 3.7 billion *tokens* from the OSCAR dataset, which includes documents in over a hundred languages, including Portuguese. It is a derivative of the DeBERTa architecture (He *et al* ., 2020). This model was designed to optimize the understanding and generation of text in Brazilian Portuguese, incorporating language-specific features.

BERTimbau (Souza *et al* .,2020): this model has achieved state of the art in named entity recognition (NER), sentence textual similarity (STS) and textual implication recognition (RTE) in Brazilian Portuguese. In this study, the *BERTimbau Base* aviator trained in BrWaC (*Brazilian Portuguese Web as Corpus*).

DeBERTinha (Campiotti *et al* ., 2023): an adaptation of the DeBERTaV3 XSmall model pre -trained in English for Brazilian Portuguese NLP tasks, with 40 million parameters. In addition to BrWac , DeBERTinha also used the Carolina dataset.

2.1.3 Evaluation Metrics

Evaluating the quality of a model requires the use of quantitative metrics relevant to the problem at hand. In this study, precision, *recall* and *F1-Score were used* as performance evaluation measures.

According to Rezaeenour *et al* . *(2022) the precision* metric in classifier evaluation aims to determine the proportion of samples correctly classified as positive. It is calculated by dividing the number of true positives by the total number of positives identified, according to Equation 1.

$$Precision = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Positivos} \quad (1)$$

The recall metric operates along the same lines as the precision metric, but uses false negative samples. Its formula uses the ratio of the number of true positives to the sum of true positives and false negatives, as per Equation 2. This metric measures the completeness of the classifier, where a low value indicates a high rate of false positives.

$$recall \;=\; \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos\ /\ Falsos\ Positivos} \qquad (2)$$

The *F1-score* is the harmonic mean of precision and *recall* , providing a balanced metric that considers both false positives and false negatives, as per Equation 3. It is useful when there is an imbalance between labels.

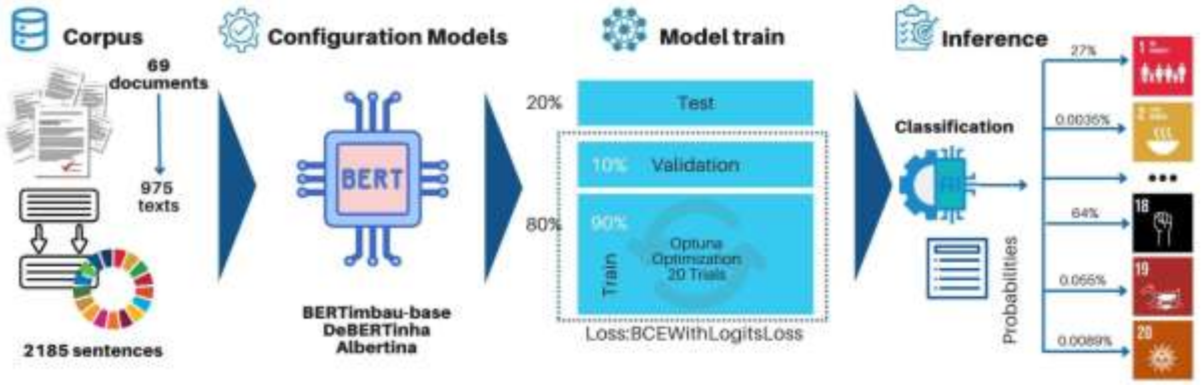
$$F1\ score \;=\; 2\,.\frac{Precision\ .recall}{Precision\ +\ recall} \qquad (3)$$

3 METHODOLOGY

The analytical framework is presented in Figure 2. First, a multi-label corpus in Portuguese was constructed and validated, corresponding to the 17 UN SDGs and the three new proposed SDGs. Then, the pre-trained models BERTimbau, Albertina and DeBERTinha were evaluated and compared to identify the most suitable one for the corpus. The selected model was then adjusted, giving rise to ODSBahia-PTBR, which was evaluated against other text classification models, based on metrics such as precision, *recall* and *F1-score* .

Figure 1

Analytical structure of the ODSBahia-PTBR model methodology.



3.1 CONSTRUCTION AND VALIDATION OF THE TRAINING CORPUS

No corpus was found in Portuguese labeled with the SDGs; therefore, a specific corpus was created for this task from manually selected excerpts of text in Portuguese, mainly from the Luz Reports from 2017 to 2023.

The Luz Report is produced by the Civil Society Working Group for the 2030 Agenda in Brazil (GT Agenda 2030), composed of several Brazilian civil society organizations dedicated to monitoring and evaluating the implementation of the SDGs in the country, with consultation with experts on the various topics covered by the 2030 Agenda. These reports aim to provide an independent and critical analysis of Brazil's progress in relation to the SDGs, covering several dimensions, such as indicators, public policies, advances and challenges. The selected text excerpts contain contextualized information about the SDGs associated with them. In addition, excerpts from news and scientific articles were collected.

A total of 69 documents were analyzed, from which 975 excerpts containing between 3 and 8 sentences were extracted, resulting in 2,185 sentences. These documents were selected for their relevance and comprehensiveness in the context of socio-environmental policies and their relationship with the SDGs. On average, each excerpt contains approximately 70 words. The majority of these excerpts, 614 in total, were extracted from the Luz Report. The selection criterion was based on the clarity with which the context of each excerpt indicated the related SDGs. In cases where the content addressed multiple topics, additional SDGs were assigned to capture the complexity of the issues addressed.

The SDGs were represented in a 20-dimensional vector, where each position indicates the presence (1) or absence (0) of a specific SDG in the analyzed excerpt (Table 1). In this case, a vector [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0] indicates that the text is related to SDGs 2, 13, 15. The assignment of the labels was validated through an analysis of the contexts presented.

Table 1

Examples of records from the training dataset .

Text_PT	One_hot_labels
There are no effective efforts by the three branches of government to reverse the political underrepresentation of women, black people, indigenous people, LGBTQIP+ people, people with disabilities and other social groups in decision-making processes. Women continue to occupy only 16% of the seats in the Chamber of Deputies, while they make up 51.8% of the population; black men and women make up 24.4%35 of federal parliamentarians and 56.2% of the population; and only one female deputy represents the 256 indigenous nations still existing in the country.	[0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1]

The reduction in the structure of the National Institute of Colonization and Agrarian Reform (INCRA) and an economic model that favors agribusiness exporting commodities particularly affect traditional peoples and communities (indigenous peoples, quilombolas, among others), including by making land earmarked for land redistribution available to the market. We have noted the intensification of bills that threaten the democratization of land.	[0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1]
Climate change is one of humanity's greatest challenges. The increase in greenhouse gas (GHG) concentrations, intensified by human action since the Industrial Revolution and with enormous growth throughout the 20th century, impacts the entire world. These impacts are evidenced by the increased frequency of climate events such as droughts, floods and strong winds, changes in hydrological cycles and the resulting changes in agricultural productivity patterns.	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0]

The dataset was randomly divided into training (80%) and testing (20%). From the training set, 10% was separated for validation.

Figure 2

Distribution of the number of texts by SDG.

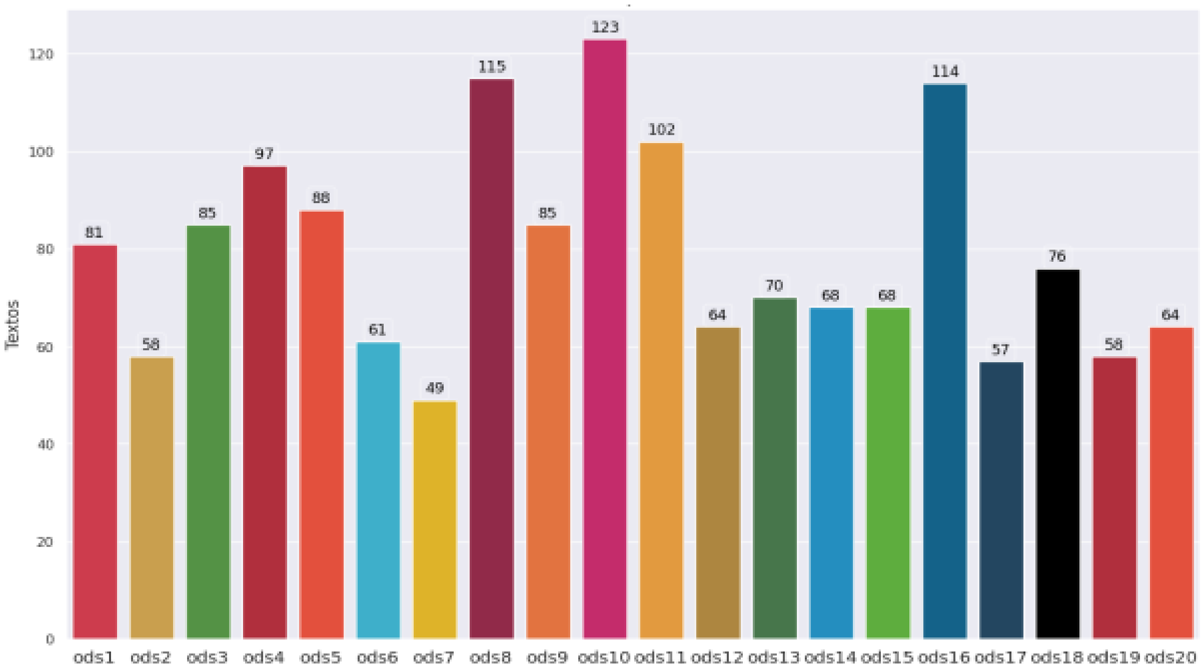


Figure 3 shows the distribution of the texts collected in relation to the 20 SDGs. Since this is a multi-label classification, the same text may be related to more than one SDG. To balance the training and generalization of the model, weighted weights were used in the loss function.

3.2 EVALUATION OF PRE-TRAINED MODELS

The Python programming language (== 3.10) was used to handle data structuring and model evaluation. No conventional preprocessing was performed on the raw corpus data, preserving statistical and linguistic features to facilitate feature extraction. This includes maintaining case, retaining punctuation and characters, and preserving stopwords , allowing the algorithm to identify important signals for prediction.

The experiments were performed in the Google Colab Pro+ program using a V100 GPU, running on Ubuntu 22.04.3 LTS with approximately 51GB of available RAM and 16GB of video memory.

The transformer library (== 4.41.2) developed by Hugging Face (Hugging Face, 2016), was used to tune the pre-trained models Albertina PTBR100M (*PORTULAN/albertina-100m-portuguese-ptbr-encoder*), BERTimbau Base (*neuralmind / bert - base - portuguese - cased*) and DeBERTinha (*saguin- nlp / debertinha-ptbr-xsmall*). These models were compared based on the selected performance metrics, and the model that presented the best performance was then selected and adjusted for the development of ODSBahia-PTBR.

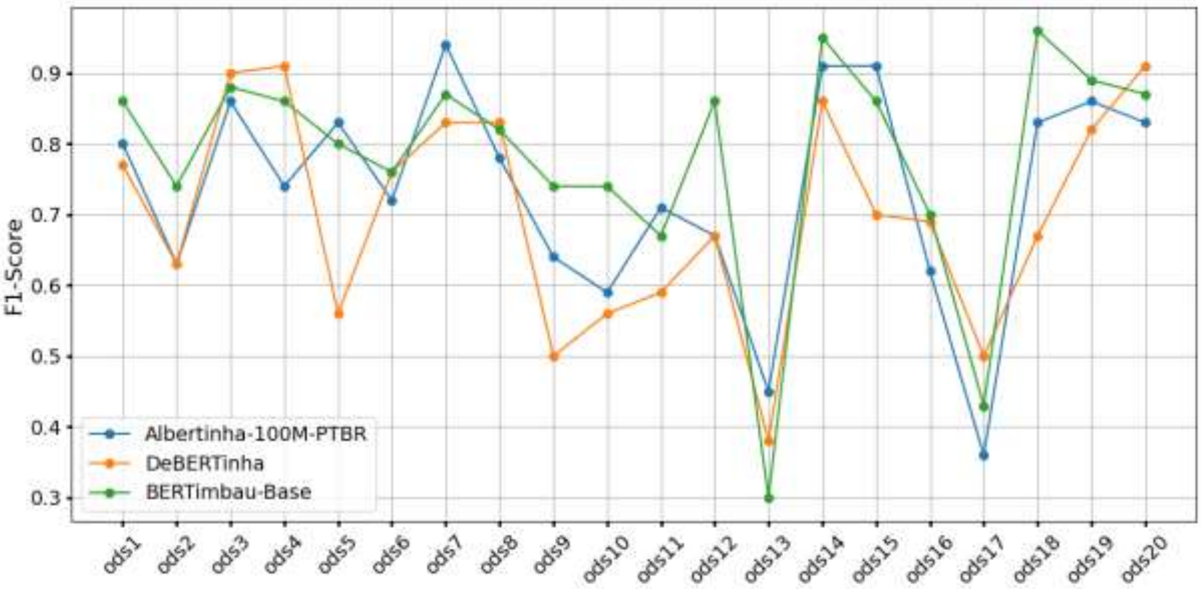
Through the Bayesian optimization library Optuna (== 3.6.1) (Akiba *et al* . 2019), multiple iterations were performed to determine the optimal values of the hyperparameters . After several experiments, the optimized values for the hyperparameters were considered : the number of iterations (*epochs* , ranged from 2^{3} to 2^{5}), the batch size (ranged from 2^{2} to 2^{4}), the learning rate that ranged from approximately $10^{-5}$to 10^{-4}. As a loss function, the Binary Cross Entropy with logit loss (*BCEWithLogitsLoss*) with weight adjustment (*pos_weight) and the* AdamW optimization algorithm (Loshchilov & Hutter , 2019) were used.

4 RESULTS AND DISCUSSIONS

The results presented in this section refer to the performance of the pre-trained models against the test set. The precision, *recall* , and micro *F1-Score metrics* were calculated using *scikit-learn* (== 1.3.2) with a threshold of 0.5. Based on the results obtained (Figure 4), there were small differences in the performance of the models in this multi-label classification task for the 20 SDGs, but they may impact the inferences.

Figure 3

F1-Score by ODS for each pre-trained model.



The models showed very similar performances in terms of *F1-Score* , which suggests that they are all effective in maintaining a balance in predictions. BERTimbau Base and Albertina 100M PTBR performed slightly better compared to DeBERTinha. However, DeBERTinha, despite having a lower *recall (Table 2), still maintains a competitive F1-Score* , indicating that it would also be a viable choice. BERTimbau Base was selected and used as a basis for adjusting the ODSBahia-PTBR model.

Table 2

Results of adjustments to the models .

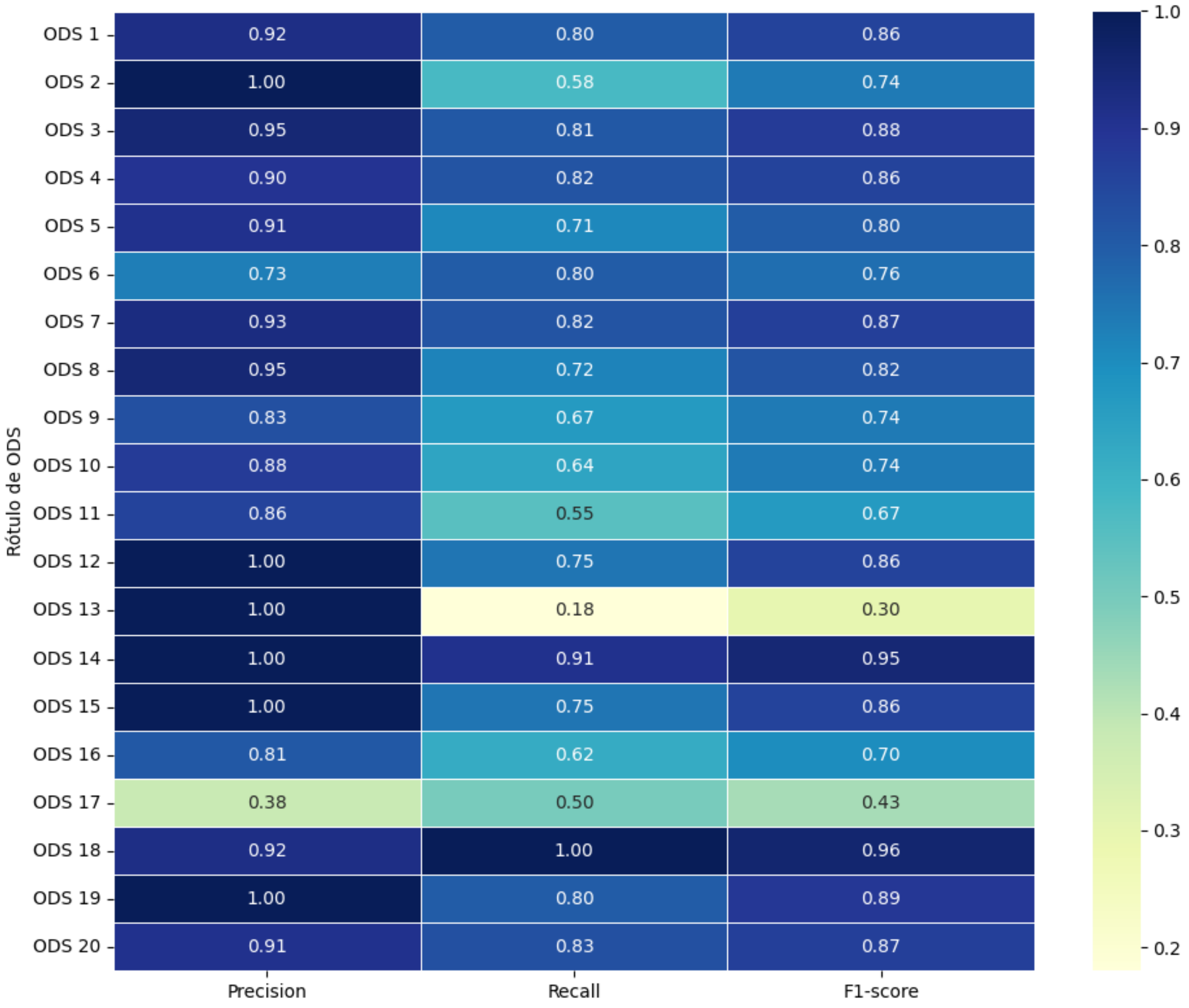
Model	Precision	Recall	*F1-Score*
BERTimbau Base	0.89	0.70	0.79
Albertina 100M PTBR	0.80	0.71	0.75
DeBERTinha	0.84	0.63	0.72

BERTimbau Base showed a precision of 0.89, which means that the majority of predictions made by the model are correct. This value is important to reduce errors in the classification of texts that are not aligned with the SDGs. The *recall* of 0.70 suggests that the model can correctly identify 70% of the relevant SDGs, although it failed to identify the SDGs in 30% of cases. The F1-Score of 0.79 indicates a good balance between precision and *recall* , suggesting that the model can identify texts that truly belong to the SDG in question while maintaining accuracy (ensuring that texts identified as belonging to the SDG actually do).

By isolating the predictions of the multi-label model for each SDG individually, as in a binary classification problem, the following BERTimbau Base results were obtained:

Figure 4

Binary metrics for each SDG.



Good performance is observed for some labels, such as SDG 3, SDG 5, SDG 14, SDG 18, and SDG 20, where precision and *recall metrics* are high, resulting in a high *F1-Score* . On the other hand, there is room for improvement for some SDGs, such as SDGs 6, 11, 13, and 17, where *recall* is lower. In addition, some labels showed an imbalance between precision and *recall*, such as SDG 2 and SDG 16. SDGs 13 and 17 showed the lowest results, possibly due to the complexity of the topics and the lack of sufficient training examples that adequately cover the nuances of these objectives.

Recently, several studies have explored NLP techniques in various applications related to the SDGs. Fux *et al* . (2022) developed a tool for classifying legal processes and their correlation with the SDGs. To do this, they collected PDF documents from initial petitions and rulings from STF cases, and extracted texts using the optical character recognition (OCR) technique. The study, which used text pre-processing techniques and creation of *embeddings* , employed recurrent neural networks along with other classic approaches, such as binary vectors (*one -hot encoding*) and word frequency representations (*Bag- of - Words - BoW* and *TF-IDF*).

Chen *et al* . (2023) conducted a study with NLP focused on environmental, social and corporate governance (*Environmental, Social and Governance* (ESG) to identify companies aligned with the SDGs based on the text of their sustainability disclosures. Analyzing corporate social responsibility reports of companies in the Russell 1000 index (a US stock market index) between 2010 and 2019, they applied logistic classifiers, support vector machines, and a fully connected neural network to predict SDG alignment. Specifically, they used dictionary-based word embeddings as input features, based on the *Word2Vec* and *Doc2Vec models* , to classify companies' alignment with the SDGs over time.

Kharlashkin *et al* . (2023) present a study on the alignment of university courses with the SDGs. To do so, they used the PaLM 2 language model to generate training data from descriptions of these courses and used the generated data to train some smaller language models aimed at predicting the SDGs. Among the pre-trained models used are BERT, mBERT , RoBERTa , XLM- RoBERTa and BART, which were fine-tuned for multi-class classification tasks . The authors sought to contribute to a better understanding of the SDGs in the university environment, allowing the identification of how the 2030 Agenda is aligned with each course.

Some authors have focused on the task of multi-label classification using English documents related to the SDGs. Guisiano , Chiky & Mello (2021) conducted research in which they developed a multi-label classification tool for texts, employing the BERT model. To do so, they collected already labeled texts related to one or more SDGs from the IISD SDG Knowledge Hub website, ensuring that each text contained a minimum of 512 words. The collection was performed using the *Webscraping technique* . The authors used the *Fast-Bert11 library* , which is built on top of *PyTorch* and supports the BERT and XLNet models , to perform text classification in both multi-class and multi-label scenarios . This approach allowed an analysis with 98% accuracy under a limited amount of training data.

Callaghan *et al* . (2021) used DistilBERT with nested cross-validation to categorize and extract specific information from 102,160 studies on climate impacts, focusing on SDG 13. Using multi-label classifiers to predict impact categories and climate drivers in the included papers, the authors used the results to create a database and map trends in anthropogenic climate change from 1951 to 2018.

The study conducted by Matsui *et al* . (2022) also in the field of multi-label classification of SDGs using the BERT model, involved adjusting a model pre-trained by the Japanese Wikipedia, the base variant of Tohoku -BERT (*cl- tohoku / bert -base- japanese - whole-word-masking*) with the aim of identifying the relationship with SDGs in input sentences, and vectorizing the semantics of the sentences.

Analyzing the results together with the studies mentioned above, the metrics obtained by the adjusted ODSBahia-PTBR model indicate similar performance, as shown in Table 3.

Table 3

Text classifier metrics in relation to the SDGs .

Proposals	**Precision**	**Recall**	***F1-Score***
Matsui *et al.* (2022)	0.93	0.92	0.93
Kharlashkin *et al.* (2024)	0.769	0.803	0.786
This model (ODSBahia-PTBR)	**0.89**	**0.70**	**0.79**
Chen *et al.* (2021)	0.819	0.744	0.745
Callaghan *et al.* (2021) (ODS 13)	-	-	0.71
RAFA 2030 (Fux *et al.* , 2022) (ODS 3,8,10,16)	-	0.881	0.636

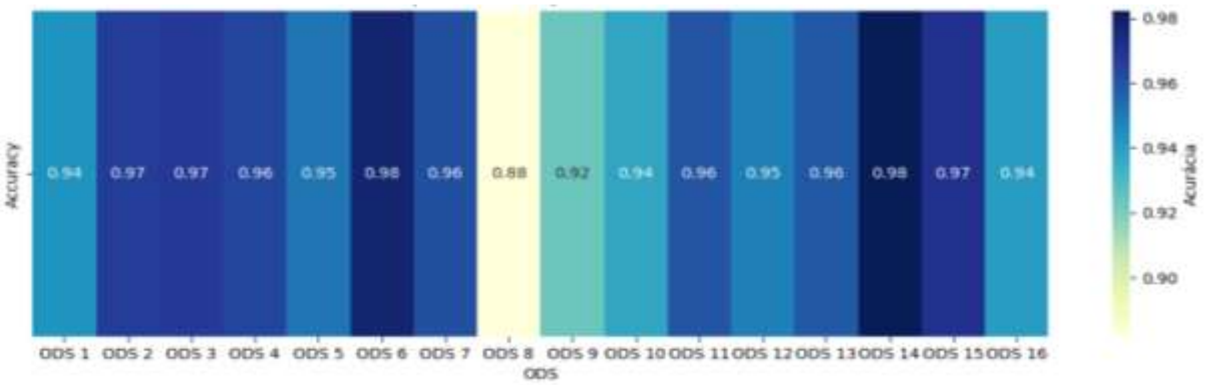
In order to evaluate performance on a larger dataset, the public database OSDG Community Dataset (OSDG-CD) was translated and submitted to the ODSBahia-PTBR classifier. The translation from English to Portuguese was performed using an Application Programming Interface (API) . *Programming Interface*), which made it possible to send translation requests to the Google Translate online platform .

OSDG-CD is a public dataset with thousands of paragraph text snippets derived from over three thousand publicly available documents. All of these texts are labeled with the related SDG and have been validated by over 1,400 volunteers from over 140 countries. Each text is composed of 3 to 6 sentences and has on average about 90 words. The CSV dataset file used (version 2024.01 made available on January 1, 2024) has 42,630 text snippets and a total of 306,595 SDG labels, excluding SDG 17 (OSDG, 2024). Each sample provides a consistency score calculated from the annotation results of the different volunteers. Only the samples whose " *labels_positive* " > " *labels_negative* " were used, retaining 35,218 samples.

In addition to correctly labeling the ODSBahia-PTBR, it was able to identify other correlations present in the OSDG-CD texts. Analyzing each label individually, based on a binary classification, ODSBahia-PTBR obtained an accuracy of 0.9523 in classifying the labels for SDGs 1 to 16 (Figure 6).

Figure 5

Accuracy by ODS on the OSDG-CD basis.



Recent studies discussed in this section demonstrate the importance and applicability of using NLP techniques to support the SDGs. This intersection between AI and SDGs opens up a range of opportunities to monitor and address sustainable development challenges more autonomously.

Studies such as those by Fux *et al* . (2022), Chen *et al* . (2023) and Kharlashkin *et al* . (2023) exemplify the practical application of NLP in different contexts, from the classification of legal proceedings to the evaluation of corporate sustainability reports and the analysis of university courses. These works also highlight the importance of considering linguistic and cultural nuances when adapting models for different languages.

The construction of a corpus in Portuguese for the adjusted ODSBahia-PTBR model fills a gap in the literature, enabling a more contextualized analysis of the SDGs. The approach presented here can be expanded and adapted for different purposes, such as the evaluation of public policies, the monitoring of business initiatives, and the promotion of sustainable practices in local communities.

Furthermore, the proposal of this work to cover the classification of texts also in relation to the three new SDGs focused on Racial Equality (SDG 18), Art, Culture and Communication (SDG 19) and Rights of Indigenous Peoples and Additional Tr Communities (SDG 20), allows the ODSBahia-PTBR model to expand the capacity to identify initiatives that impact these groups.

These three new SDGs were introduced in the document “Agenda 2030 Guide: integrating SDGs, Education and Society”, prepared in 2020 in partnership between UnB and

UNESP, highlighting themes that, although not explicitly represented in the 17 SDGs, are of great relevance for Brazil and the world (Cabral & Gehre , 2020). This proposal not only complements the existing SDGs, but also recognizes the interconnection between the various current social demands. The inclusion of these new goals reflects a critical exercise of reimagining and resignifying the characterizations of the 2030 Agenda, aligning with its main objective of ensuring that all individuals are included in sustainable development efforts.

Although the fight against racism is transversally included in other SDGs, such as the eradication of poverty, the guarantee of quality education, and the reduction of inequalities, SDG 18 highlights this issue. It seeks to promote racial equality by confronting all types of racism, highlighting its importance for building sustainable and inclusive development. SDG 19 highlights the social impact of art, promoting the mobilization of social groups, new perceptions of the world, and the construction of knowledge about social dynamics. SDG 20, in turn, aims to preserve and value culture and traditional knowledge, guaranteeing the rights of indigenous peoples, quilombolas, riverside communities, and other social groups.

Therefore, this model, specifically adjusted to the Brazilian context, was developed to understand and process diverse linguistic and cultural nuances, expanding its applicability not only to the 17 UN SDGs, but also to these three additional proposed ones.

5 CONCLUSION

This study presented an approach for classifying texts in Portuguese related to the Sustainable Development Goals (SDGs), through the development and evaluation of the ODSBahia-PTBR model.

Initially, a multi-label corpus in Portuguese was constructed and validated, corresponding to the 17 UN SDGs and three new SDGs proposed in the '2030 Agenda Guide: integrating SDGs, Education and Society'. This corpus was then used for fine-tuning the pre-trained models BERTimbau Base, Albertina and DeBERTinha, which were evaluated and compared based on precision, *recall* and *F1-score metrics* .

Based on the results obtained, there were small variations between BERTimbau Base, DeBERTinha and Albertina 100M PTBR in the selected metrics. Although all these pre-trained models proved to be effective, BERTimbau Base was used as a baseline because it achieved better performance, and was then adjusted, giving rise to ODSBahia-PTBR. This model is published on Huggingface (*odsbahia / odsbahia-ptbr*) and is openly available for use.

We analyzed ODSBahia-PTBR in relation to other classification models described in the literature that revealed satisfactory results in terms of precision, *recall* and *F1-Score* . The model demonstrated to be capable of dealing with the particularities of the Portuguese language, identifying the relations of texts with the ODS with an accuracy of 89%.

This work not only proposed an objective approach for the classification of Portuguese texts in relation to the SDGs, but also demonstrated the importance of choosing and adapting pre-trained models appropriately for different application contexts. Future research can focus on optimizations to improve performance on specific labels, expand the training corpus and explore new data sources for validation, in addition to fine-tuning other pre-trained models.

REFERENCES

Adhikari, A. *et al.* (2019). DocBERT: BERT for Document Classification. DOI: https://doi.org/10.48550/arXiv.1904.08398.

Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. https://doi.org/10.48550/arXiv.1409.0473.

Beltagy, I., Lo, K. & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. DOI: https://doi.org/10.48550/arXiv.1903.10676.

Beltagy, I., Peters, M. E. & Cohan, A. (2020). Longformer: The Long-Document Transformer. DOI: https://doi.org/10.48550/arXiv.2004.05150.

Bommasani, R. *et al.* (2021). On the opportunities and risks of foundation models. DOI: https://doi.org/10.48550/arXiv.2108.07258.

Brown, T. B. *et al.* (2020). Language Models are Few-Shot Learners. DOI: https://doi.org/10.48550/arXiv.2005.14165.

Cabral, R. & Gehre, T. Guia agenda 2030: Integrando ODS, Educação e Sociedade. 1 Ed. São Paulo, 2020. Disponível em: <https://repositorio.unesp.br/server/api/core/bitstreams/60bba95b-fe49-40dd-b01b-7adc68e961a0/content>. Acesso em: 20 jan. 2024.

Callaghan, M. *et al.* (2022). Machine learning-based evidence and attribution mapping of 100,000 climate impact studies. DOI: https://doi.org/10.21203/rs.3.rs-783398/v2

Campiotti, I. *et al.* (2023). DeBERTinha: A multistep approach to adapt DebertaV3 XSmall for Brazilian Portuguese natural language processing task. DOI: https://doi.org/10.48550/arXiv.2309.16844

Cho, K. *et al.* (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Anais...Stroudsburg, PA, USA: Association for Computational Linguistics. DOI: https://doi.org/10.48550/arXiv.1406.1078.

Da Costa, P. *et al.* (2023). BERTabaporu: Assessing a genre-specific language model for Portuguese NLP. Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings. Anais...INCOMA Ltd., Shoumen, BULGARIA.

Devlin, J. *et al.* (2028). BERT: Pre-training of deep bidirectional Transformers for language understanding. DOI: https://doi.org10.48550/arXiv.1810.04805

Fan, A. *et al.* (2020). Beyond English-Centric multilingual machine translation. DOI: https://doi.org/10.48550/arXiv.2010.11125

Fux, L. *et al.* (2022). Classificação de processos judiciais segundo Objetivos de Desenvolvimento Sustentável da Agenda ONU 2030. Revista da CGU, v. 14, n. 26.

Guisiano, J. E., Chiky, R. & Mello, J. de. SDG-Meter : a deep learning based tool for automatic text classification of the Sustainable Development Goals. hal.science. Disponível em: <https://hal.science/hal-03738404>. Acesso em: 12 abr. 2024.

Hajikhani, A. & Suominen, A. (2022). Mapping the sustainable development goals (SDGs) in science, technology and innovation: application of machine learning in SDG-oriented artefact detection. Scientometrics, v. 127, n. 11, p. 6661–6693. DOI: https://doi.org/10.1007/s11192-022-04358-x

He, P. *et al.* (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. DOI: https://doi.org/10.48550/arXiv.2006.03654.

Hossin; S. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, v. 5, n. 2, p. 01–11.

Hu, D. (2020). An introductory survey on attention mechanisms in NLP problems. Em: Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020. p. 432–448. DOI: https://doi.org/10.1007/978-3-030-29513-4_31.

Hugging Face. Hugging Face – On a mission to solve NLP, one commit at a time. huggingface.co. Disponível em: <https://huggingface.co/>. Acesso em: 11 dez. 2023.

Islek, I. & Oguducu, S. G. (2020). A hybrid recommendation system based on bidirectional encoder representations. Em: ECML PKDD 2020 Workshops. Cham: Springer International Publishing, 2020. p. 225–236. DOI: https://doi.org/10.1007/978-3-030-65965-3_14.

Kaliyar, R. K., Goswami, A. & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia tools and applications, v. 80, n. 8, p. 11765–11788. DOI: https://doi.org/10.1007/s11042-020-10183-2.

Kharlashkin, L. *et al.* (2024). Predicting sustainable development goals using course descriptions -- from LLMs to conventional foundation models. DOI: https://doi.org/10.48550/arXiv.2402.16420.

Khurana, D. *et al.* (2023). Natural language processing: state of the art, current trends and challenges. Multimedia tools and applications, v. 82, n. 3, p. 3713–3744. DOI: https://doi.org/10.1007/s11042-022-13428-4.

Lee, J. *et al.* (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (Oxford, England), v. 36, n. 4, p. 1234–1240. DOI: https://doi.org/10.1093/bioinformatics/btz68.

Liu, Y. *et al.* (2019). RoBERTa: A robustly optimized BERT pretraining approach. DOI: https://doi.org/10.48550/arXiv.1907.11692 .

Luong, M.-T., Pham, H. & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. DOI: https://doi.org10.48550/arXiv.1508.04025.

Manning, C., Raghavan, P. & Schutze, H. (2008). Introduction to Information Retrieval. [s.l.]: Cambridge University Press. ISBN 978-0-521-86571-5.

Matsui, T. *et al.* A natural language processing model for supporting sustainable development goals: translating semantics, visualizing nexus, and connecting stakeholders. **Sustainability science**, v. 17, n. 3, p. 969–985, 2022. DOI: https://doi.org/10.1007/s11625-022-01093-3.

MORTON, S. *et al.*. Sustainable Development Goals (SDGs), and their implementation. **British Medical Bulletin**, v. 124, n. 1, p. 81–90, 2017. Disponível em: <https://doi.org/10.1093/bmb/ldx031>. Acesso em: 19 maio 2023.

Moustafa, Mona. (2022). On SDG 18: War legacy, Resilience, and Healing in Uncertain Times! United Nations Development Programme, Lao PDR, 8 dez. 2022. Disponível em: <https://www.undp.org/laopdr/blog/sdg-18-war-legacy-resilience-and-healing-uncertain-times>. Acesso em: 19 jun. 2024.

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. DOI: https://doi.org/10.48550/arXiv.2005.10200.

ONU BR – NAÇÕES UNIDAS NO BRASIL – ONU BR. A Agenda 2030. 2015. Disponível em: <https://nacoesunidas.org/pos2015/agenda2030/>. Acesso em: 21/01/2024.

OSDG, UNDP IICPSD SDG AI Labe PPMI. (2024). OSDG Community Dataset (OSDG-CD). Zenodo, jan. 01. https://doi.org/10.5281/zenodo.10579179.

Otter, D. W., Medina, J. R. & Kalita, J. K. (2021) A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems, v. 32, n. 2, p. 604–624. DOI: https://doi.org/10.1109/TNNLS.2020.2979670.

Qiu, X. *et al.* (202). Pre-trained models for natural language processing: A survey. Science China Technological Sciences, v. 63, n. 10, p. 1872–1897. DOI: https://doi.org/10.1007/s11431-020-1647-3.

Raffel, C. et al. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. DOI: https://doi.org/10.48550/arXiv.1910.10683.

Ray, B., Garain, A. & Sarkar, R. (2021). An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. Applied soft computing, v. 98, n. 106935, p. 106935. DOI: https://doi.org/10.1016/j.asoc.2020.106935.

Rezaeenour, J. *et al.* (2023). Systematic review of content analysis algorithms based on deep neural networks. Multimedia tools and applications, v. 82, n. 12, p. 17879–17903.

Rordrigues, J. *et al.* (2024). Fostering the ecosystem of open neural encoders for Portuguese with Albertina PT* family. DOI: https://doi.org/10.48550/arXiv.2403.01897.

Rogers, A., Kovaleva, O. & Rumshisky, A. (2020). A Primer in BERTology: What we know about how BERT works. DOI: https://doi.org/10.48550/arXiv.2002.12327.

Sanh, V. *et al.* (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. DOI: https://doi.org/10.48550/arXiv.1910.01108.

Smith, T. B. *et al.* (2021). Natural language processing and network analysis provide novel insights on policy and scientific discourse around Sustainable Development Goals. Scientific reports, v. 11, n. 1. DOI: https://doi.org/10.1038/s41598-021-01801-6.

Souza, F., Nogueira, R. & Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em: Intelligent Systems. Cham: Springer International Publishing, p. 403–417.

Sutskever, I., Vinyals, O. & Le, Q. V. (2014). Sequence to sequence learning with Neural Networks. DOI: https://doi.org/10.48550/arXiv.1409.3215.

Vaswani, A. *et al.* (2017) Attention is all you need. In NIPS. DOI: https://doi.org/10.48550/arXiv.1706.03762.

Vinuesa, R. *et al.* (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature communications, v. 11, n. 1. DOI: https://doi.org/10.1038/s41467-019-14108-y.

Virtanen, A. *et al.* (2019). Multilingual is not enough: BERT for Finnish. DOI: https://doi.org/10.48550/arXiv.1912.07076.

Yang, Z. & Choi, J. D. (2019). FriendsQA: Open-domain question answering on TV show transcripts. Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue. Anais [...] Stroudsburg, PA, USA: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/W19-5923.

Yang, Z. *et al.* (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. DOI: https://doi.org/10.48550/arXiv.1906.08237.

Zhang, L. *et al.* (2020). Sentiment analysis methods for HPV vaccines related tweets based on transfer learning. Healthcare (Basel, Switzerland), v. 8, n. 3, p. 307. DOI: https://doi.org/10.3390/healthcare8030307.