UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

Gabrielle Fidelis de Castilho

Similarity Analysis between Sustainability Reports and the Sustainable Development Goals using Machine Learning

São Carlos

2023

Gabrielle Fidelis de Castilho

Similarity Analysis between Sustainability Reports and the Sustainable Development Goals using Machine Learning

Monograph presented to the Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Data Science.

Concentration area: Data Sciences

Advisor: Profa. Dra. Gleici da Silva Castro Perdoná

Original version

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

F451s

Fidelis de Castilho, Gabrielle
Similarity Analysis between Sustainability Reports and the Sustainable Development Goals using Machine Learning / Gabrielle Fidelis de Castilho; orientadora Gleici da Silva Castro Perdoná. -- São Carlos, 2023.
30 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Aprendizado de Máquina. 2. Processamento de Linguagem Natural. 3. Aprendizado Profundo. 4. Relatórios de Sustentabilidade. 5. Objetivos de Desenvolvimento Sustentável. I. da Silva Castro Perdoná, Gleici, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

ACKNOWLEDGEMENTS

I want to thank my advisor, Dr. Gleici da Silva Castro Perdoná, for all the guidance, support, and understanding. And to gratefully acknowledge Fundação de Apoio à Física e Química and the Swedish Institute for providing me with funding to pursue my graduate studies.

I must also express my most profound gratitude to my loved ones, especially Daniel Thomás Ramos Franco de Sá and Dr. Thalita Reis da Silva, for believing, supporting, and encouraging me. This accomplishment would not have been possible without them.

ABSTRACT

CASTILHO, G. F. de **Similarity Analysis between Sustainability Reports and the Sustainable Development Goals using Machine Learning**. 2023. 29p. Monograph (MBA in Data Sciences) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

This work intends to assess the degree of similarity between Brazilian companies' ESG (Environmental, Social and Governance) reports and the 17 Sustainable Development Goals. Four different machine approaches were deployed and tested to achieve this ambition while using human analysis as a benchmark for the evaluation. The final results show that the supervised variant of SimCSE (Simple Contrastive Learning of Sentence Embeddings) is the best method for our study regarding congruence with the human approach. The results of this work provide an initial step towards monitoring best practices that address environmental, social, and economic issues.

Keywords:: Machine Learning. Natural Language Processing. Deep Learning. Sustainability Reporting. Sustainable Development Goals.

RESUMO

CASTILHO, G. F. de **Análise de Similaridade entre Relatórios de Sustentabilidade e Objetivos de Desenvolvimento Sustentável usando Aprendizado de Máquina**. 2023. 29p. Monografia (MBA em Ciências de Dados) - Centro de Ciências Matemáticas Aplicadas à Indústria, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

Este trabalho tem como objetivo avaliar o grau de semelhança entre os relatórios ESG (do inglês Environmental, Social, and Governance) de empresas brasileiras e os 17 Objetivos de Desenvolvimento Sustentável. Quatro diferentes abordagens computadorizadas foram implementadas e testadas para atingir esse objetivo com a análise humana servindo de referência para a avaliação. Os resultados finais mostram que a variante supervisionada do SimCSE (do inglês Simple Contrastive Learning of Sentence Embeddings) é o melhor método para o nosso estudo em termos de congruência com a análise humana. Os resultados deste trabalho fornecem um passo inicial para o monitoramento de boas práticas que abordam questões ambientais, sociais e econômicas.

Palavras-chave: Aprendizado de Máquina. Processamento de Linguagem Natural. Aprendizado Profundo. Relatórios de Sustentabilidade. Objetivos de Desenvolvimento Sustentável.

LIST OF FIGURES

LIST OF TABLES

LIST OF FRAMES

LIST OF ABBREVIATIONS AND ACRONYMS

BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag of Words
EGD	European Green Deal
ESG	Environmental, Social, and Governance
GRI	Global Reporting Initiative
IDF	Inverse Document Frequency
LDA	Latent Dirichlet Allocation
ML	Machine Learning
NLP	Natural Language Processing
OCR	Optical Character Recognition
PDF	Portable Document Format
RoBERTa	Robustly Optimized BERT Approach
SBERT	Bi-Encoder Sentence Transformer
SDG	Sustainable Development Goals
SimCSE	Simple Contrastive Learning of Sentence Embeddings
SNLI	Stanford Natural Language Inference
STS	Semantic Textual Similarity
UN	United Nations
TF	Term Frequency
TF-IDF	Term Frequency - Inverse Document Frequency
TM	Text Mining
USE	Universal Sentence Encoding
USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos

CONTENTS

1 INTRODUCTION

Sustainable development is a phenomenon that impacts our society worldwide. Research has shown that sustainability compliance, often related to Environmental, Social, and Governance (ESG), provides significant long-term performance enhancements when integrated into investment evaluation (WANG; DOU; JIA, 2016).

Over recent years, companies have started to target mitigating their environmental impact and adapting their practices to the dynamic sustainability context (LUCCIONI; E.; DUCHENE, 2020). Sustainability reporting has become a requirement for certain companies in many countries, being recognized as a factor as important as the company's financial disclosures (HALKOS; NOMIKOS, 2021). ESG disclosures inform stakeholders about companies' engagement with sustainable development practices. It not only represents its commitment to the UN 2030 Sustainable Development Goals (SDG) agenda, but it is also a benchmark of the economic robustness of a corporation in the years to come (SERVAES; TAMAYO, 2017).

However, satisfying current disclosure standards is voluntary. It can include nonspecific information and embedded figures or tables, making such documents hard to compare regarding layout and critical performance indicators. Also, due to the abundance of data, sustainability analysts must investigate hundreds of pages of reports to find relevant information.

This work intends to quantify the degree of similarity between ESG disclosures and the United Nation's (UN) SDG indicators and targets. Specifically to identify to what degree the report discusses these goals. The degree of similarity is calculated using a semantic measure obtained by combining Information Retrieval and Natural Language Processing (NLP) methods, thus allowing stakeholders to conduct further analysis and considerations. We present the methods and materials in this work.

This study is restricted to the English language, as pre-trained models in Portuguese are much more restricted in variety, and to Brazilian companies due to the geographical interest of the author. However, it should not restrict the impact and influence of this work on future developments in other languages or countries.

In this monograph, we start in Chapter 2 by examining fundamental concepts, and related work, including state-of-the-art. Chapter 3 describes the steps to develop the models and the technical and design details. Chapter 4 investigates the results. In Chapter 5, we cover the conclusions and recommendations for future work.

2 LITERATURE REVIEW

2.1 The United Nations' 2030 Sustainable Development Goals

The UN's 2030 Agenda for Sustainable Development comprises 17 Sustainable Development Goals (Frame 1), each divided into indicators and targets. The Agenda promotes that ending poverty and other deprivations must go in parallel with "strategies that build economic growth and address a range of social needs, including education, health, social protection, and job opportunities" (UNITED NATIONS, 2015) while accounting for environmental protection and climate change. The SDGs call for profound changes in every country, considering the different national stages of development, capacities, and specific difficulties. Thus, "all countries share responsibility for achieving the SDGs, and each has a critical role, requiring coordinated efforts by governments, civil society, research, and business" (UNITED NATIONS, 2015).

The UN Secretary-General provides a yearly report of the implementation progress based on current data from statistical systems at national and regional levels. An independent group of scientists also writes the quadrennial Global Sustainable Development Report to inform the review discussions at the General Assembly every four years (UNITED NATIONS, 2015).

2.2 Sustainability reporting and Natural Language Processing

The first registers of businesses describing their economic, social, and environmental impact date back to the 1970s. Then, environmentally sensitive industries started publishing corporate social responsibility reports in the 1980s. These reports were mostly limited to more prominent companies. However, the increased awareness of businesses' importance in achieving sustainable development motivated a recent adoption among small and medium enterprises, institutions and non-profit organizations (HSU; WEN-HAO; WEI-CHUNG, 2013).

Disclosures under an ESG framework aim to inform investors, stakeholders, customers, and regulators about their sustainability-related actions and make their efforts quantifiable. In fact, according to the Global Reporting Initiative (GRI), reporting is neither "a matter of communication nor a mere data gathering or compliance exercise. It helps organizations to reflect on their sustainable strategies, set goals, measure performance, and manage change" (GLOBAL REPORTING INITIATIVE, 2013).

Recent norms on non-financial disclosures have incentivized studies of sustainability reporting. However, even when companies rigorously follow the existing standards for sustainability reporting, they provide diversified documents containing qualitative and

Frame 1 – The 17 sustainable development goals to transform our world

Goal	**Ambition of the goal**
1: No Poverty	End poverty in all its forms everywhere
2: Zero Hunger	End hunger, achieve food security and improved nutrition and promote sustainable agriculture
3: Good Health and Well-being	Ensure healthy lives and promote well-being for all at all ages
4: Quality Education	Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all
5: Gender Equality	Achieve gender equality and empower all women and girls
6: Clean Water and Sanitation	Ensure availability and sustainable management of water and sanitation for all
7: Affordable and Clean Energy	Ensure access to affordable, reliable, sustainable and modern energy for all
8: Decent Work and Economic Growth	Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all
9: Industry, Innovation and Infrastructure	Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation
10: Reduced Inequality	Reduce inequality within and among countries
11: Sustainable Cities and Communities	Make cities and human settlements inclusive, safe, resilient and sustainable
12: Responsible Consumption and Production	Ensure sustainable consumption and production patterns
13: Climate Action	Take urgent action to combat climate change and its impacts
14: Life Below Water	Conserve and sustainably use the oceans, seas and marine resources for sustainable development
15: Life on Land	Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss
16: Peace and Justice Strong Institutions	Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels
17: Partnerships to achieve the Goal	Strengthen the means of implementation and revitalize the global partnership for sustainable development

Source: (UNITED NATIONS, 2015)

quantitative information, irregular layouts, and embedded tables and figures, making strategies, priorities, and results hardly comparable. This factor, combined with the amount of documentation produced worldwide, makes examining sustainability reports through NLP a potentially meaningful scientific endeavour (ZHOU; WANG; YUEN, 2021).

Diverse industries widely use NLP, with applications such as spam detection, chatbot development, and text classification and summarizing. NLP is a group of methods for analyzing, understanding, and translating human language to machines (Eisenstein, 2019). More importantly, they are becoming crucial for the scientific community in studying

techniques to help report sustainability (GREWAL; SERAFEIM, 2020).

For a computer to decipher the natural language, it must convert the written or spoken message into a numerical format. However, there are notable challenges when implementing these techniques since context, homonyms, errors, slang, and sarcasm cannot be inferred and decoded by a computer (EISENSTEIN, 2019). We believe that human judgement on word relations plays an essential role in NLP analysis and will be one of the methods used in this work.

2.3 Related works and state of the Art

Researchers have investigated companies' non-financial disclosures using Text Mining (TM) and NLP for over a decade. Such techniques aid in reviewing aspects otherwise impractical for the human reader, while automatic textual processing helps extract implicit knowledge for subsequent analysis. For this reason, Machine Learning (ML) methods are becoming even more relevant and continue improving processes and decision-making (JO, 2019).

The analysis of corporate environment reports has received particular attention, going from statistical and Bayesian techniques to multi-discriminatory analysis (MODAPOTHALA; ISSAC, 2009). One of the most used methods in the past decade was Latent Dirichlet Allocation (LDA). While unsupervised learning was also adopted to identify patterns and clusters (TREMBLAY; PARRA; CASTELLANOS, 2015).

Aureli et al. used text mining to verify variations of sustainability disclosures before and after an industrial disaster (AURELI *et al.*, 2016). Liu utilized the Term Frequency - Inverse Document Frequency (TF-IDF) to retrieve relevant terms for several machine learning algorithms and analytical models (LIU; CHEN; LI, 2017). Szekely and Brocke performed LDA to record the evolution of sustainability reports topics over 16 years (SZEKELY; BROCKE, 2017). Chae and Park used computational content analysis on ESG-related conversations on Twitter (CHAE; PARK, 2018). And Koundouri et al. proposed a methodology to connect the SDG objectives with the European Green Deal (EGD) policies (KOUNDOURI *et al.*, 2022).

Many studies on applying Machine Learning in sustainability reports are devoted to specific industries to identify characteristics, best practices, and trends related to specific sectors. For instance, Liew et al. identified practices and trends in the chemical industry (LIEW; ADHITYA; SRINIVASAN, 2014). And Zhou et al. used LDA to examine shipping companies' sustainability reports (ZHOU; WANG; YUEN, 2021). Transformer models and different variations of architectures such as BERT have had significant success for domain-specific applications, from question answering to machine translation and natural language inference, in legal-court-case reports, medical databases or family businesses in

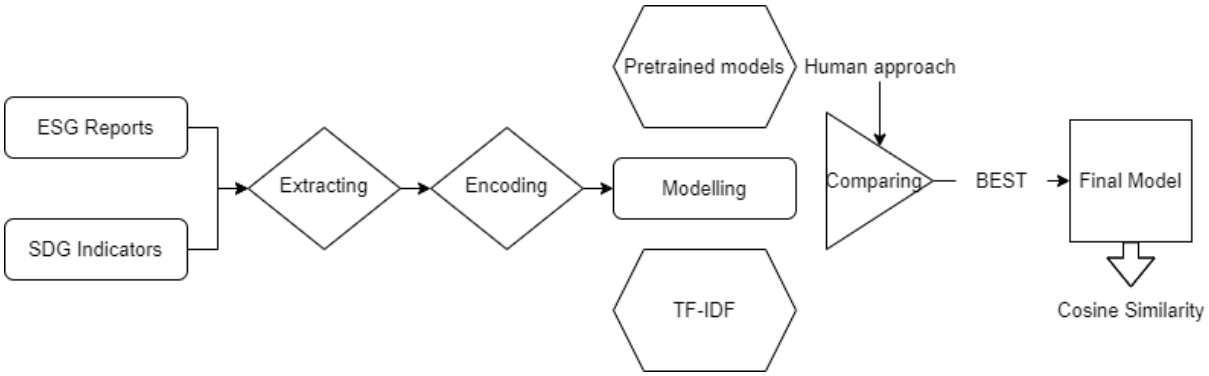
tourism (GUTIERREZ-BUSTAMANTE; ESPINOSA-LEAL, 2022).

This brief overview of the literature demonstrates how exploring similar algorithms could support the investigation of ESG disclosures.

3 MATERIALS AND METHODS

The use of Machine Learning is within the scope of this study to provide a fast classification between the SDGs and ESG reports for present and future works. A methodology allowed for planning the work, assessing the potential difficulties, and focusing on the development. Testing four different algorithms on Google Colab, and using its GPU hardware accelerator, was the procedure we followed when determining a model for this problem. In Figure 1, we propose the procedure that we will cover in detail in the following sections. We applied different libraries to process reports and SDG indicators to extract and clean the text. Afterwards, we applied different ML models. As far as we know, there is no reliable computerized test to endorse the results found. Thus, the "human approach" will be the fifth and final analysis method, which we consider as the gold standard.

Figure 1 – Proposed scheme



Source: Author

3.1 The Data

Our case deals with SDG indicators and sustainability reports sourced by the companies websites in Portable Document Format (PDF) format. Here is a brief description of the data:

- Sustainability reports: We downloaded reports made publicly available by Brazilian companies for 2021. We enforced no limitations regarding any aspects of the companies.
- SDG Declaration: The declaration consists of targets and metrics for the Sustainable Development Goals, which were grouped in one file and divided by pages, one for each goal.

3.2 Pre-processing

3.2.1 Text Extraction

As we accessed most of the information through PDF files, advanced methods were used to extract and correctly manipulate the data used in this study. Using Python programming language, our first attempt was to utilize the Optical Character Recognition (OCR) package. We transformed pages into images for the tool to derive the raw data. The resulting texts needed more quality for analysis, requiring comprehensive cleaning before being used within any numerical model. For this reason, we decided to obtain the texts directly from the PDF format with the PDFplumber library. The conversion of each PDF page was iterative, providing both individual pages, and a compiled document. This distinction allows us to detect the specific page where each SDG goal received more emphasis in the analyzed document if such information is needed.

3.2.2 Data Cleaning

An appropriate portrayal of the extracted text is fundamental for any analysis. Therefore, classical text cleaning techniques were applied to improve the quality of the text captured. It removes all neutral or negative aspects that could impact the model's performance. In particular, we decided to lowercase all words and eliminate non-alphanumeric characters, including white spaces and punctuation marks. The Bag of Words approach required extra cleaning to reduce words to a common English root (stemming) and remove stop-words using the Spacy package.

3.3 Method 1: Term Frequency - Inverse Document Frequency (TF-IDF)

Bag of Words (BoW) approach is a collection of classical methods that maps unique words in the entire text corpus. It extracts text features and converts them into a unique vector index, disregarding grammar, word order and syntax. In TF-IDF, the values of the vector for each document are the product of two numbers: the Term Frequency (TF) and the Inverse Document Frequency (IDF). The TF-IDF of each word in each text is the product of the individual TF and IDF scores. The conclusion is that frequent words in one document that are scarce in the entire corpus are crucial for that document and have a high score (LEO, 2022).

- TF: It measures how important that word is to the document. The number of times that word occurs in that document.
- IDF: It measures how rare the word is. The log of the inverse of the fraction of documents in which the word occurs.

The TfidfVectorizer module in the Scikit-learn library normalizes the values so that longer documents do not overpower the calculation. It also prevents 0 divisions. We then compare the cosine similarity between these embedding vectors (LEO, 2022). An example can be observed in Figure 2.

Figure 2 – TF-IDF Vectorizer example

Sentence 1: this is a sentence, a short sentence.
Sentence 2: this is another short sentence.

Word	Sentence 1 TF	Sentence 1 IDF	Sentence 1 TF*IDF	Sentence 2 TF	Sentence 2 IDF	Sentence 2 TF*IDF
this	1	log(2/2)	0	1	log(2/2)	0
is	1	log(2/2)	0	1	log(2/2)	0
a	2	log(2/1)	-0.6	0	-	0
sentence	2	log(2/2)	0	1	log(2/2)	0
short	1	log(2/2)	0	1	log(2/2)	0
another	0	-	0	1	log(2/1)	-0.6

Source: (LEO, 2022)

Though BoW approaches are intuitive, their performance varies greatly, as texts without common words could still be similar in meaning. We can overcome this issue by measuring Semantic Textual Similarity (STS) (LEO, 2022). These approaches are generally more accurate than the non-contextual approaches, which we will cover in the following sections.

3.4 Method 2: Universal Sentence Encoding (USE)

Due to the varied layouts and subjectivity of reports, we decided to deploy a model that has semantic capability. For this reason, deep learning becomes fundamental to finding suitable models.

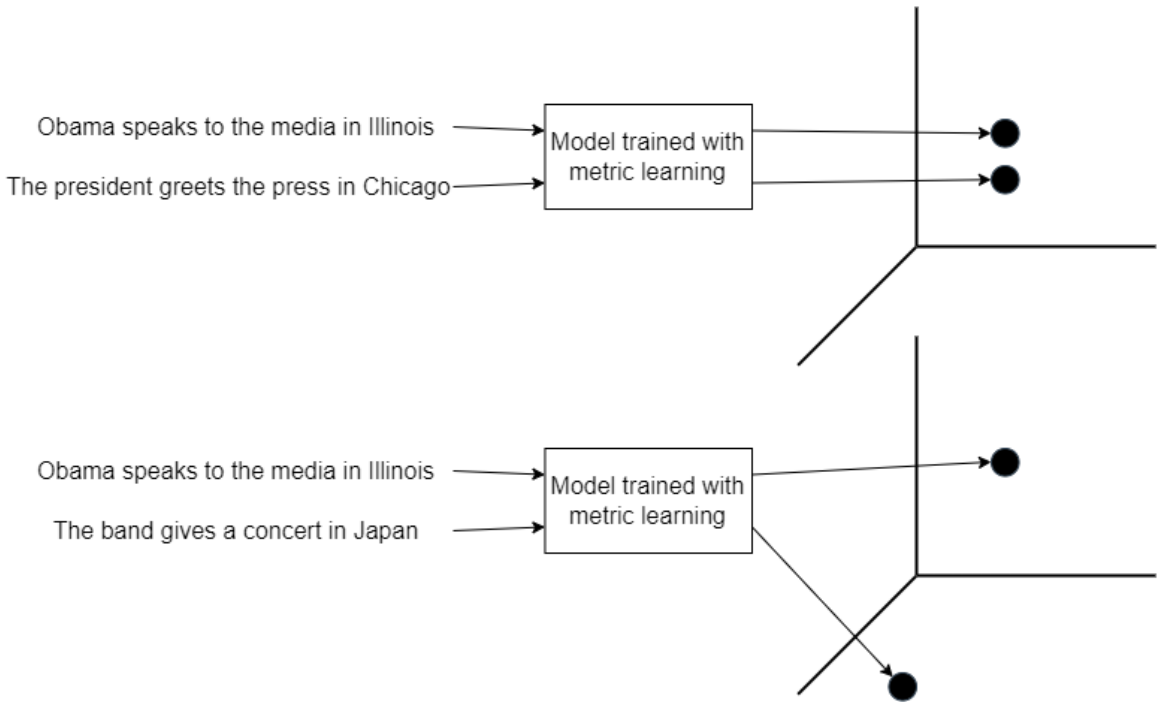
In 2017, Google Research introduced The Transformer, a deep learning model that uses self-attention mechanisms and weights the importance of each part of the text inputted. This innovation accelerated the development of many models, including the Universal Sentence Encoder, a Transformer-based model used for transfer learning (LEO, 2022). Transfer learning means learning how to solve one problem and apply it to a related, but different one (TENSORFLOW, 2022).

The Universal Sentence Encoder makes sentence embedding as simple as for individual words, which we can then use to compute sentence meaning similarity using less supervised training data. The pre-trained model first computes the contextual word embedding for each word in the sentence. Then, it computes the sentence embedding by summing up all of the word vectors and normalizes the sentence lengths by dividing by the square root of its length (LEO, 2022). Google Research open-sourced this pre-trained model on TensorFlow Hub. With the embedding for each sentence, we can calculate the cosine similarity using Scikit-learn.

3.5 Bidirectional Encoder Representations from Transformers (BERT)

Subsequently, using metric learning, researchers developed even better-performing methods based on the Transformer. At its very fundamental level, metric learning uses a neural network to convert texts to embeddings, then design these embeddings so that semantically similar texts cluster near to each other while dissimilar texts are further apart (LEO, 2022) (Figure 3).

Figure 3 – Example of natural language text in vector space



Fonte: (LEO, 2022)

In 2018, Google Research presented the BERT model, revolutionizing NLP, as BERT, a bidirectional transformer, derives its power from its self-supervised pre-training task called Masked Language Modeling. It randomly hides words and trains the model to predict what was removed by providing the previous and following words (LEO,

2022). Training over an extensive collection of texts allows BERT to learn the semantic relationships between the different words in a language. This study used pre-trained models based on the Stanford Natural Language Inference (SNLI).

3.5.1 Stanford Natural Language Inference dataset (SNLI)

The SNLI dataset considers the semantic similarity of 570 thousand independent English sentence pairs. Its purpose is to be a standard for evaluating text and a resource for developing NLP models. Frame 2 shows examples from the development of the dataset. Each pair contains the labels of five human annotators and the consensus regarding entailment, contradiction, and neutral (THE STANFORD NATURAL LANGUAGE INFERENCE CORPUS, 2015).

Frame 2 – Examples of five annotators on SNLI corpus

Text	**Judgment**	**Hypothesis**
A man inspects the uniform of a figure in some East Asian country.	C-C-C-C-C contradiction	The man is sleeping.
An older and younger man smiling.	N-N-E-N-N neutral	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	C-C-C-C-C contradiction	A man is driving down a lonely road.
A soccer game with multiple males playing.	E–E–E–E–E entailment	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	N-N-E-C-N neutral	A happy woman in a fairy costume holds an umbrella.

Source: (THE STANFORD NATURAL LANGUAGE INFERENCE CORPUS, 2015)

3.6 Method 3: Bi-Encoder Sentence Transformer (SBERT)

Over time, several notable improved BERT variants were developed, such as RoBERTa, DistilBERT, and ALBERT. Sentence Transformers are the current state-of-the-art sentence embeddings. It is based on BERT and its variants and is pre-trained utilizing a type of metric learning called contrastive learning. In contrastive learning, the loss function compares the similarity of two embeddings by labelling them as 0 (similar) and 1 (dissimilar). A Bi-Encoder Sentence Transformer model processes each text independently, denoting an embedding vector as the output. Then, we can compare the two documents using cosine similarity (LEO, 2022).

We used the SentenceTransformer library to apply open-source pre-trained models on the SNLI dataset. These models function as follows:

Firstly, use the labelled SNLI dataset as training data and compute the contextual word embeddings of a text using any pre-trained BERT model as an encoder. Secondly,

compute the element-wise average of all the token embeddings to obtain a single fixed-dimension sentence embedding for the entire text. Thirdly, train the model using a siamese network architecture with contrastive loss. The objective is to move embeddings for similar texts closer together and embeddings from dissimilar texts further away. Finally, after training the model, we can compare two texts by calculating the cosine similarity.

Although Bi-Encoders impress when combined with vector search databases for billions of documents, sentence transformers are fully supervised and demand massive, high-quality, labelled data collection. Thus, adopting sentence transformers to new domains is time-consuming and expensive. Fortunately, some cutting-edge research in semi-supervised and self-supervised learning offer encouraging results.

3.7 Method 4: Simple Contrastive Learning of Sentence Embeddings (SimCSE)

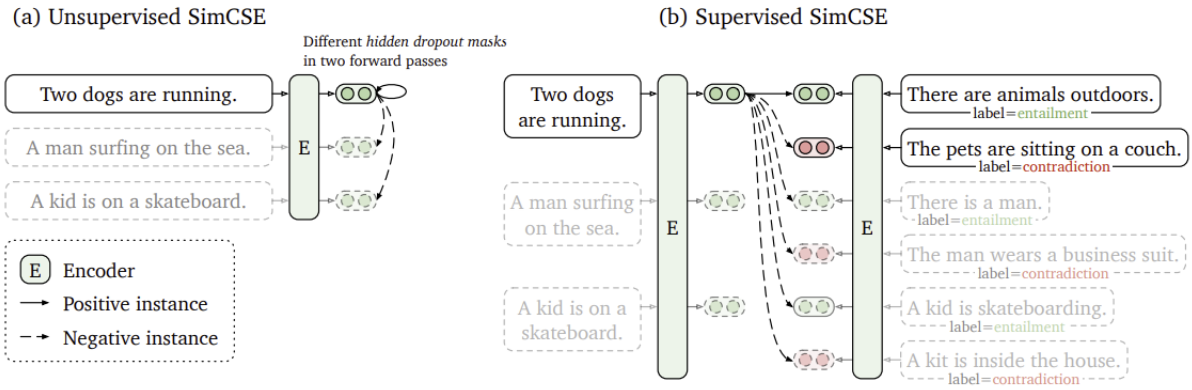
Simple Contrastive Learning of Sentence Embeddings (SimCSE) is a Bi-Encoder Sentence Transformer model trained using the SimCSE approach. We can directly reuse all the code from the Bi-Encoder Sentence Transformer model but change the pre-trained model to the SimCSE models. It is a self-supervised algorithm for learning sentence embeddings that we can either train as a supervised model if labelled data is available or in a completely unsupervised fashion(LEO, 2022). The core idea is as follows:

Firstly, compute the embeddings of a text document using any pre-trained BERT model as an encoder and take the embeddings of the classification token. Secondly, create two noisy versions of the same text embedding by applying two different dropout masks on the original embedding. The model expects them to have a cosine distance of 0. Thirdly, the model expects all the other texts in the batch to have a cosine distance of 1 to the target text embeddings from the previous step. The loss function then updates the parameters of the encoder model. Finally, supervised SimCSE has one additional step where we use a Natural Language Inference labelled dataset to obtain "positive" pairs from texts that are labelled "entailment" and "negative" pairs from texts that are labelled "contradiction". Figure 4 explains the entire process conceptually.The figure and the following explanatory texts are excerpts from the original paper by Gao, Yao, and Chen (GAO; YAO; CHEN, 2021).

(a) Unsupervised SimCSE predicts the input sentence from in-batch negatives, applying different hidden dropout masks.

(b) Supervised SimCSE leverages the SNLI datasets and takes the entailment pairs as positives and contradiction pairs and other in-batch instances as negatives.

Figure 4 – Unsupervised and supervised SimCSE



Source: (GAO; YAO; CHEN, 2021)

3.8 Method 5: Human Analysis

In this work, comprehending the semantic similarity between two documents via machine learning is only a comparative part of a more in-depth independent human analysis. The human method consists of reading the analyzed documents in full, then creating a ranking from 1 to 17, starting from more similar and moving towards dissimilar, indicating the SDGs covered in the document. In our case, the analysis was performed by only one human.

4 RESULTS

Our goal was to assess the degree of similarity of corporations' ESG disclosures with the UN's 2030 Sustainable Development Goals. As an application example, we compared the SDGs and an authentic company report, in this case, Klabin S.A. We selected this company because Klabin is a paper producer, exporter, and the largest paper recycler in Brazil, with solid communication regarding sustainability. Klabin is also part of the World portfolio of the Dow Jones Sustainability Index, with the advantage of having a report with a regular layout, which facilitates text extraction and cleaning (KLABIN S.A., 2023).

In our analysis, we first pre-processed the data and deployed the four models. Then, we created a table with the results in parallel to better visualize the dimension of our comparison. The process takes 5 to 10 minutes per report, which saves considerable time compared to a manual data analysis, which can take several hours. In Table 1, we present cosine valuations for all four methods. The goal was to have an impression of initial values.

Table 1 – Cosine Similarity

SDG	TF-IDF	USE	SBERT	Sup SimCSE	Unsup SimCSE
1	8.8	31.9	70.0	75.3	42.5
2	11.7	44.1	76.4	79.2	52.3
3	8.4	37.6	71.3	72.8	41.1
4	9.0	31.9	72.1	70.1	26.9
5	7.4	31.1	68.6	65.4	33.9
6	11.2	38.0	74.7	77.5	53.9
7	8.7	33.2	74.4	75.9	47.7
8	10.5	38.0	78.3	76.3	38.3
9	11.6	31.8	77.2	77.7	44.9
10	8.3	25.2	70.4	73.4	40.8
11	12.9	40.0	74.6	76.2	42.6
12	18.0	41.4	79.0	83.4	59.8
13	9.5	37.2	72.8	74.6	45.6
14	10.4	43.0	75.1	77.0	58.5
15	13.7	49.3	75.3	82.5	59.9
16	7.0	31.3	70.8	69.5	37.0
17	11.4	43.4	75.7	77.5	50.3

Source: Author

On the other hand, Figure 5 ranks each goal from 1 to 17, according to its similarity in the human method. Each SDG is then compared to the ranking of the other four methods, providing a total field that indicates the best and worst relations overall. We can observe immediately that the highest relationship among the goals is with SDGs 12 and 15 by almost all methods, including human analysis.

Figure 5 – Machine and human rankings with a total difference

SDG	Human	TF-IDF	USE	SBERT	Sup SimCSE	Unsup SimCSE
12	1	1	5	1	1	2
15	2	2	1	6	2	1
6	3	7	8	8	5	4
9	4	5	14	3	4	9
17	5	6	3	5	6	6
14	6	9	4	7	7	3
2	7	4	2	4	3	5
8	8	8	7	2	8	14
13	9	10	10	11	12	8
7	10	13	11	10	10	7
11	11	3	6	9	9	10
3	12	14	9	13	14	12
10	13	15	17	15	13	13
5	14	16	16	17	17	16
16	15	17	15	14	16	15
4	16	11	13	12	15	17
1	17	12	12	16	11	11
Total difference to Human		42	54	36	26	34

Source: Author

Regarding the results of each approach, TF-IDF performs well, as it captures the expected correlation with goals 12 and 15. However, a few divergences were observed, such as the high connection with goal 11. The most probable reason is the lack of vocabulary interpretation capabilities. We must remember that the results are achieved only by the frequency of words appearing in each text.

The USE model, as well as the other deep learning models, correlate highly with SDG 2. It informs us that there might be a connection that we should look further. After all, machine learning can rapidly find connections not easily perceived by humans. For example, SDG 2 (Zero Hunger) targets agriculture practices that increase productivity and production. Therefore it is related to maintaining ecosystems that improve adaptation to extreme weather, climate change and other disasters and increasingly strengthen soil conditions. The report indeed covers these topics.

The SBERT model is quite peculiar because it found a substantial similarity with SDG 8. Other approaches could not find a similarly strong correlation. The report covers decent work for all, including young people and persons with disabilities, but not at such a high level. It may occur because the vocabulary is similar to other, more closely related, goals.

SimCSE algorithms perform much better in its supervised variant, as its results

proved to be very consistent with our benchmark, the human approach. We can use its results for a fast and robust evaluation of the link between sustainability reports and the SDGs. It is important to remark that although the unsupervised SimCSE model had a performance degradation compared to the supervised one, it would be the go-to method in domains where sufficient labelled data is unavailable or expensive to collect.

5 CONCLUSION

In this research, we proposed a pipeline of information retrieval and Natural Language Processing able to identify the level to which ESG disclosures are in line with the UN's 2030 sustainability agenda. We proposed five different approaches to support the analysis. The first method was the Bag of Words technique. The second method was a pre-trained Universal Sentence Encoder. The third one involves a pre-trained Bi-Encoder Sentence Transformer model. The fourth one is an evolution of SBERT models using a SimCSE approach. The fifth approach was human judgement, our gold standard.

All machine approaches provided a fast classification for sustainability reports regarding the UN's Sustainable Development Goals using cosine similarity of vectors. The proposed approach is an initial step towards monitoring best practices that address environmental, social, and economic sustainability. As organizations often produce poorly structured reports that are complex to read and analyze, a machine learning model can be helpful in this type of operation. Nonetheless, albeit time-consuming, we assumed the human approach to be a standard and concrete criterion for the validity of the other four methods.

The results showed that the first and second models are compelling but need more robustness for our problem. However, we can still utilize them to examine the similarity between ESG disclosures and the SDGs more superficially. The third algorithm performs well in capturing the best connections between the SDGs and the documents examined but not as well as the fourth method in identifying the overall relation. On the other hand, the similarity percentage among all machine approaches and SDGs 1, 4, 5 and 16 were low, confirming the results found by the human method.

The text pre-processing presented here could be improved. Moreover, because the reports provided by companies are generally unstructured, text extraction and data cleaning are time-consuming and can be improved continuously. As text analysis is a vast dominion, further studies could improve access to layouts in different pre-processing stages and perform more advanced actions to improve accuracy.

The results achieved, though limited to the English language and PDF format, show that the technique is viable. Future research could focus on overcoming these constraints to different languages, countries and formats. We also believe that domain-specific word embedding can improve the generalization capacity of these methods. Additionally, extending the human analysis to more than one experts could reduce the bias of the human method.

Finally, although the results of this work provide some advice on how a company's

accomplishments can be more easily ventured, due to the points outlined above, it should be taken as only a partial guide to its actions' relevance.

REFERENCES

AURELI, S. *et al.* Sustainability disclosure after a crisis: A text mining approach. **International Journal of Social Ecology and Sustainable Development**, v. 7, n. 1, p. 35–49, 2016.

CHAE, B.; PARK, E. Corporate social responsibility (csr): A survey of topics and trends using twitter data and topic modeling. sustainability. **Sustainability**, 2018.

EISENSTEIN, J. **Introduction to Natural Language Processing**. Cambridge: United States: MIT Press, 2019.

GAO, T.; YAO, X.; CHEN, D. Simcse: Simple contrastive learning of sentence embeddings. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. [*S.l.: s.n.*], 2021. p. 6894–6910.

GLOBAL REPORTING INITIATIVE. **An Introduction to G4**: The next generation of sustainability reporting. Geneva, 2013.

GREWAL, J.; SERAFEIM, G. Research on corporate sustainability. review and directions for future research. **Foundations and Trends® in Accounting**, v. 14, n. 2, 2020.

GUTIERREZ-BUSTAMANTE, M.; ESPINOSA-LEAL, L. Natural language processing methods for scoring sustainability reports—a study of nordic listed companies. **Sustainability**, v. 14, 2022.

HALKOS, G.; NOMIKOS, S. Corporate social responsibility: Trends in global reporting initiative standards. **Economic Analysis and Policy**, v. 69, p. 106–117, 2021. Available at: https://doi.org/10.1016/j.eap.2020.11.008.

HSU, C. W.; WEN-HAO, L.; WEI-CHUNG, C. Materiality analysis model in sustainability reporting: a case study at lite-on technology corporation. **Journal of Cleaner Production**, v. 57, p. 142–151, 2013.

JO, T. **Text Mining: Concepts, Implementation, and Big Data Challenge**. Cham: Switzerland: Springer International Publishing, 2019.

KLABIN S.A. **Website**. 2023. Available at: https://klabin.com.br/en/sustentabilidade. Access at: 02 jan. 2023.

KOUNDOURI, P. *et al.* A methodology for linking the energy-related policies of the european green deal to the 17 sdgs using machine learning. **Working Paper Series**, 2022.

LEO, M. S. **Website**. 2022. Available at: https://towardsdatascience.com/semantic-textual-similarity-83b3ca4a840e. Access at: 03 jan. 2023.

LIEW, W.; ADHITYA, A.; SRINIVASAN, R. Sustainability trends in the process industries: A text mining-based analysis. **Computers in Industry**, v. 65, p. 393–400, 2014.

INTERNATIONAL CONGRESS ON ADVANCED APPLIED INFORMATICS, 6., 2017, Japan. **Text-Mining Application on CSR Report Analytics: A Study of Petrochemical Industry**. 76-81 p.

LUCCIONI, A.; E., B.; DUCHENE, N. Analyzing sustainability reports using natural language processing. **Tackling Climate Change with Machine Learning workshop at NeurIPS 2020**, 2020.

MODAPOTHALA, J. R.; ISSAC, B. Study of economic, environmental and social factors in sustainability reports using text mining and bayesian analysis. *In*: 2009 IEEE SYMPOSIUM ON INDUSTRIAL ELECTRONICS APPLICATIONS. [*S.l.: s.n.*], 2009. v. 1, p. 209–214.

SERVAES, H.; TAMAYO, A. The role of social capital in corporations: A review. **Oxford Review of Economic Policy**, v. 33, n. 2, p. 201–220, 2017.

SZEKELY, N.; BROCKE, J. vom. What can we learn from corporate sustainability reporting?: Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. **PLOS ONE**, v. 12, n. 4, p. 1–27, 2017.

TENSORFLOW. **Website**. 2022. Available at: https://www.tensorflow.org/guide/keras/transfer_learning. Access at: 03 jan. 2023.

THE STANFORD NATURAL LANGUAGE INFERENCE CORPUS. **Website**. 2015. Available at: https://nlp.stanford.edu/projects/snli/. Access at: 03 jan. 2023.

TREMBLAY, M.; PARRA, C.; CASTELLANOS, A. Analyzing corporate social responsibility reports using unsupervised and supervised text data mining. *In*: INTERNATIONAL CONFERENCE ON DESIGN SCIENCE RESEARCH IN INFORMATION SYSTEMS, Ireland. Dublin, 2015. p. 439–446.

UNITED NATIONS. Department of Economic and Social Affairs. **Website**. Geneva, 2015. Available at: https://sdgs.un.org/goals. Access at: 02 feb. 2023.

WANG, Q.; DOU, J.; JIA, S. A meta-analytic review of corporate social responsibility and corporate financial performance: The moderating effect of contextual factors. **Business Society**, v. 55, n. 8, p. 1083–1121, 2016.

ZHOU, Y.; WANG, X.; YUEN, K. F. Sustainability disclosure for container shipping: A textmining approach. **Transport Policy**, v. 110, p. 465–477, 2021.