Tinbergen Institute (2023), , 1–29
Supervisor: Meike Morren

THESIS

Comparative Analysis of Natural Language Models Regarding Sustainable Development Goals

Nazlıcan Eroğlu Çetin*

Tinbergen Institute, Amsterdam, the Netherlands
*Corresponding author. Email: n.eroglu@tinbergen.nl

Abstract

Developing NLP models to classify texts into the United Nations Sustainable Development Goals (SDGs) is one of the popular research inquires thanks to the the power of large language models. By leveraging advanced computational techniques, these models enable us to decipher the underlying themes and intentions within texts. They provide valuable insights into how any given text align with the diverse objectives of sustainable development. From commercial services to research institutions, we observe that there are many studies which aim to develop natural language models which are trained on SDG classification task. Even though the aims of these models are the same, previous literature shows that their results are not aligned. In this study, we aim to propose a comparative study between these models while we investigate some of the selected models at a deeper level. Ultimately, we plan to propose an answer for the difference between these models.

Keywords: Sustainable development goals, natural language processing, large language models, BERT, embeddings, keyword analysis

Contents

1. Introduction

United Nations has launched Sustainable Development Goals (SDGs) in 2015 for the first time and since then they have been expanding both the comprehensiveness and depth of these goals. One of the cutting-edge technology that researchers can utilize in the pursuit of the effectiveness of these goals is the recent developments in large language models (LLMs). Models trained on vast textual domains have the capability to accurately identify the most pertinent goal within any given text such as abstracts of academic works and policy documents from governmental organizations. The ultimate benefit of this is that it provides a comprehensive insight into the distribution of resources aligned with these objectives. Thanks to this, we can learn more about the level of attention and support given to each SDG.

There are various research groups around the world which focuses on this objective each with slightly different models and methods. One interesting thing to note at this point is that these models produce divergent results even though they are trained on the same objective. While it is natural to see such disperancies across models given the initial differences occurring in various steps, it is worrying to see very contrasting predictions for relatively "easy" texts. We believe that there is more research needed on this point specifically to guide the community of this research agenda in a more aligned path. Even though there are several studies which compares the fine-tuned versions of these models and assess their performance, we believe that we should investigate further to provide a causal reasoning for these differences. There is no study within our knowledge which aims to compare some of these models at such a deeper level. This research aims to fill this gap in the literature.

Having said that, we acknowledge that this is an complex task as there are multiple moving parts in each model where it is not always possible to provide a fair comparison. Due to this, we first provide a simple baseline keyword and embedding analysis to see the potential prediction problems among SDGs. To achieve this, we have first created a unique dataset where we have gathered keywords from 5 different ontologies and queries. Later, we have deployed Ridge regressions to pin down the top keywords for each SDG and analyzed their embeddings extracted from two different models. This way, we were able to see an intuitive difference between the clustering performances of contextual models and simple frequency based models.

Secondly, we move to the comparison of the selected three models: OSDG.ai, ChatGPT and Aurora. We inspect the overall performance and potential biases in each model by utilizing different datasets with different characteristics. These datasets are again curated specifically for this study with the aim of pinning-down not only the performances in various standard evaluation metrics but also to inspect where each model is struggling.

This study is structured as follows: First, we provide an extensive literature review on the works that focus on the SDG classification using natural language processing. Second, we focus on the existing models that we are going to test and provide an overview on their characteristics. Following this, we explain our methodology thoroughly with the necessary theoretical background. Both the methodology and results sections are divided into two part for the easier readability. Methodology section is followed by results section where we provide both our findings and their interpretations. Lastly, we discuss our findings further and point possible research directions.

2. Literature Review

Text classification has become one of the most popular tasks for language models. It involves the categorization of textual data into predefined classes or categories based on its content. This results in a more efficient organization, retrieval, and analysis of large volumes of text. The importance of text classification stems from its ability to automate and streamline processes that would otherwise be time-consuming and resource-intensive.

Some of the existing models' performance on not only text classification but also on annotation has become an interesting research agenda. For instance, Gilardi, Alizadeh, and Kubli 2023 have

shown that ChatGPT developed by OpenAI outperformed crowd-workers for text-annotation tasks. Using a sample of 2,382 tweets, they demonstrate that ChatGPT surpasses crowd-workers in several annotation tasks, including relevance, stance, topics, and frames detection. The zero-shot accuracy of ChatGPT outperforms crowd-workers in four out of five tasks. Additionally, the intercoder agreement of ChatGPT exceeds that of both crowd-workers and trained annotators for all tasks. This means that we have higher level of consistency and reliability. Another reliable advantage of using LLMs is the cost-effectiveness. In their paper, they showed that The per-annotation cost of ChatGPT is significantly lower, at less than $0.003This makes it approximately twenty times cheaper than utilizing platforms like MTurk for crowd-workers. This cost advantage further enhances the efficiency of text classification tasks and it enables organizations and researchers to process large volumes of text data more economically.

In the context of sustainable development, text classification holds particular significance. With the increasing focus on achieving the United Nations' Sustainable Development Goals, the ability to classify texts related to sustainable development initiatives, policies, and progress becomes crucial for effective decision-making. Text classification enables the automated sorting of documents into specific SDG categories and it facilitates the analysis of large corpora of text to gain insights into the progress made. Ultimately, we aim to have more informed policy interventions.

The UN has recognized the significance of text classification in advancing its own objectives with respect to sustainable development. It faces the challenge of effectively managing the extensive amount of text it generates each year. These texts can be related to resolutions, speeches, reports, and various other documents. LaFleur 2019 have presented a proof-of-concept process for developing a machine learning-based system that addresses this need. They have trained an Latent Dirchlet Allocation (LDA) algorithm with 18 topics where the extra topic acts as a filter to capture common words among all the 17 representative texts. (Blei 2012) Since LDA is an unsupervised method, they have utilized the open source tool Mallet to train LDA by using 17 representative texts. In that sense, their paper presents a semi-supervised approach. They found that SDG 17 has the highest overall representation average as a result of its popularity in recent years which is followed by SDG 10.

Another work by the UN focuses on the exploration of time-series trends for four categories of resolutions, Sustainable Development, Justice and Law, Human Rights, and Peace and Security, using a traditional labeled classifier (Kim, LaFleur, et al. 2020). In their supervised approach, they have utilized the annual Secretary-General reports which gives them well defined examples of language that describes each of the categories they wish to analyze. Across the 62nd to 73rd sessions of the General Assembly, they have found that there has been an overall increase in the number of resolutions across all four categories. Notably, the category of Sustainable Development stands out as the number of resolutions has remained constant at 57 throughout the last four sessions.

Another research project from the UN presents an extensive formal knowledge organization system (KOS) to represent SDGs. KOS comprises an ontology that models the fundamental elements of the Global SDG indicator framework. Joshi et al. 2021 This framework encompasses 17 Goals, 169 Targets, and 231 unique indicators, along with over 450 related statistical data series maintained by the global statistical community to monitor SDG progress. Additionally, KOS includes a dataset that encompasses these elements. The ontology serves several purposes: Some of them are formalizing and establishing unique identifiers for the various components of the SDGs and their indicator framework. Their plot application, LinkedSDG, automatically extracts relevant SDG concepts mentioned in a given document. Users can interact with the application through an interactive wheel chart that displays all related SDGs.

Borchardt et al. 2022 presents the work from the European Commission where they investigate how the SDGs are integrated into policy-making by analyzing the semantic connections between the policy initiatives of the current Commission and the 2030 Agenda. They focus on classic text mining and natural language processing techniques where extract keywords from relevant EU policy

documents and use the keyword frequencies to derive insights (Bird, Loper, and Klein 2009). Their results show that the Commission's commitment to the Green Deal priority reflected in strong prevalence of SDG 13. The impact of the pandemic and the subsequent recovery efforts are also clearly captured in the results of SDG 3.

There is a good amount of models developed by independent research groups. One of the main challenges in classifying SDGs is the overlap among the goals and the contextual differences of the diverse SDGs. As a solution, Hsu, LaFleur, and Orazbek 2022 proposed a combinatorial fusion algorithm (CFA) to enhance the precision of the classification process by combining a topic model classifier and a semantic link classifier. CFA is an ML/AI framework that utilizes the rank-score characteristic function and cognitive diversity to combine multiple scoring systems. They showed that the precisions are equal or better for combined models than the best individual model.

Alternative to these works which utilizes topic modelling approaches, there are also other projects which aim to use pre-trained language models such as BERT or DistilRoBERTa (Devlin et al. 2018; Liu et al. 2019). One of such works is the SDG Prospector project proposed by Jacouton, Marodon, Laulanie, et al. 2022. They are mainly concerned about the classification of sustainability reports of Public Development Banks where they aim to map the extent and nature of these banks' consideration of the SDGs in their strategic and operational narratives. They apply DistilRoBERTa on institutional documentation of 237 Public Development Banks from 2016 to 2020 and their results suggests that SDG narrative predominantly revolves around the "productive" SDGs, such as SDG 8 (Decent Work and Economic Growth) and SDG 9 (Industry, Innovation, and Infrastructure). On the other hand, narratives concerning the protection of biodiversity (SDGs 14 and 15) and the reduction of poverty and inequalities (SDGs 1 and 10) are relatively minor components of these banks' activity reports. Also, there is an increasing emphasis on SDG 13 (Climate Action) across the sample. They observe a positive correlation between the size of banks' balance sheets and their consideration of environmental SDGs.

Another model which utilizes BERT is proposed by Guisiano, Chiky, and De Mello 2022. They have fine-tuned the classic BERT model with their database and launched a web based tool: SDG Meter. They have reported an accuracy of 98% in their results. They note that the BERT model has demonstrated its ability to learn and recognize SDGs by leveraging the specific information within each text. However, it lacks the knowledge to differentiate what is not an SDG.

Given all these attempts to classify any given texts into one or more SDG goals, one natural research agenda becomes the comparison of these models. Several studies have been conducted to compare the performance of different NLP approaches in various tasks. These papers aim to provide insights into the strengths and weaknesses of different models and algorithms, shedding light on their comparative performance. By evaluating and contrasting the results obtained from these approaches, researchers aim to identify the most effective techniques for specific NLP tasks. Ultimately, the goal is to further improve the overall performance of natural language processing systems.

One important paper in this literature is Morales-Hernández, Jagüey, and Becerra-Alonso 2022. This study presents a comparative analysis of different classification approaches applied to scientific papers related to sustainable development goals. The methods evaluated include the One-Versus-Rest multi-class classifier and three multi-label problem transformation techniques. The performance of the method and algorithm combinations is assessed on both balanced and imbalanced datasets and they use specific evaluation metrics to measure classification effectiveness. The findings indicate that utilizing the label powerset transformation method in conjunction with the support vector machine classification algorithm yields promising results. The combined approach achieves an impressive accuracy rate of up to 87% for imbalanced datasets, 83% for datasets with equal instances per label, and an impressive 91% for multiclass datasets.

Another paper which presents a comparative study regarding the word-embeddings is Shen et al. 2018. Word embeddings, derived from extensive unstructured text data, serve as fundamental

components in Natural Language Processing (NLP). These embeddings offer the ability to group words with similar meanings and they capture semantic similarities. In their research, they undertake a comprehensive comparative analysis between simple word-embedding-based models (SWEM) and word-embedding-based recurrent/ convolutional neural network models. Contrary to expectations, SWEMs demonstrate comparable or even superior performance across most of the evaluated cases. Building upon this insight, they introduce two novel pooling strategies for learned word embeddings. First, a max-pooling operation is proposed to enhance interpretability. Second, a hierarchical pooling operation is introduced, designed to retain spatial information, including n-gram details, within text sequences. This study presents a valuable insight in understanding the word-embeddings better. This is especially beneficial when researchers aim to understand the differences stemming from very similar models such as BERT.

The last paper we would like to present in literature review is Kashnitsky et al. 2022. This article outlines an improved approach to mapping research to the SDGs by incorporating feedback and refining the query set. The methodology includes sub-queries for targets and indicators as well as the application of a machine learning model to enhance recall. The latest version, "Elsevier 2022 SDG mapping," captures twice as many articles on average compared to the previous version while maintaining precision above 80% . Following that, the approach was evaluated by comparing the results with those generated by Digital Science, Aurora European Universities Alliance, and the University of Auckland. Their conclusion is that no single approach stands out as the best performer across all validation datasets, although they observed that Elsevier queries exhibited slightly more stability.

In the next section, we will discuss some of the existing models in a more detailed fashion as they are crucial in understanding our methodology.

3. Existing Models

Another tool that aims to provide the mapping between any given text and SDG is OSDG 2.0 which was introduced in 2022 (Pukelis et al. 2022). The distinguishing aspects of this initiative are twofold: First one is regarding the data collection procedure. When the OSDG team initially embarked on developing ML models for OSDG 2.0, they encountered a shortage of accurately labeled data required for SDG classification model development. They recognized the chance to make a contribution to SDG research by curating and sharing high-quality datasets. To pursue this objective, they initiated the OSDG Community Platform (OSDG CP), an interactive data collection initiative described comprehensively in the data section of this paper. Secondly, they support not only English like most of the other studies, but also the content in Arabic, Danish, Dutch, Finnish, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, Swedish, and Turkish.

In their methodology, the initial phase involves the utilization of ML models that have been trained on data gathered through the OSDG CP. These models are responsible for conducting an initial assessment of texts and suggesting preliminary SDG labels. They play a crucial role in maintaining context and selecting texts that fall within the general thematic areas of the SDGs. In the subsequent stage, OSDG relies on its ontology/keyword map to validate the initial labels. The process involves matching keywords to identify definitive evidence that establishes a connection between a given text and a specific SDG. For a specific SDG label to be assigned, both the ML models and ontology approaches must concur.

The second model we want to cover in this section is the model developed by the Aurora Universities network Vanderfeesten, Jaworek, and Keßler 2022. The significance of this study is that they have created "Aurora SDG Queries" according to a comprehensive feedback survey conducted among 244 senior researchers throughout Europe, as well as a thorough text analysis. To assess the efficacy of these queries, their accuracy was evaluated. The results give an average precision of 70% for all SDGs collectively and an average recall of 14%.

Another important point to note about the Aurora model is that they utilize a pre-trained multilingual BERT model. More interestingly, they have trained 169 individual models for each of the SDG targets instead of training 1 multi-label model. They argue that training a model for each target was preferred because they could fine-tune each one of these models separately. Moreover, due to variations in the quantity of training data for each target, training a single multi-label model would necessitate complex calculations of target weights. As a result, they obtain 169 individual binary classifiers that determine whether a paper belongs to a specific target.

Lastly, we want to briefly mention ChatGPT as it is another state-of-art model which can easily be utilized in the task of classifying texts into SDGs Radford et al. 2018. They present evidence that substantial improvements can be made in specific text classification tasks by deploying a two-step approach: generative pre-training of a language model on a diverse collection of unlabeled text, followed by discriminative fine-tuning for each individual task. Unlike previous works in this area, their approach incorporates task-specific input transformations during fine-tuning. This enables effective transfer while minimizing the need for extensive modifications to the model architecture.

In terms of classification into SDGs, one way to use ChatGPT is to use techniques like transfer learning. We basically take the model which is trained on a large corpus of text and then fine-tune it on our classification task using the labeled data. One thing to note is that fine-tuning is currently available only for davinci, curie, babbage and ada base models.

In table 1, some major characteristics of the existing models have been reported to give an overall summary to the readers. More elaboration on this table will be provided in the discussion part of this study with the additional interpretation that we derive from the findings.

Table 1. Some characteristics of existing models

Models	Info on training dataset	Base model/method	Number of parameters
OSDG.ai	OSDG Community dataset which is annotated by a large number of community members. It has 41,689 text excerpts derived from publicly available documents, including reports, policy documents, and publication abstract.	The first stage uses the ML models to propose initial SDGs and the second stage uses ontology/keyword map to verify the initial labels. ML and ontology approaches must align for final SDG label.	N/A
Aurora	English abstracts in the corpus of 1.4 million research papers using the Aurora SDG Query model v5.	Multilabel BERT	110 million parameters
ChatGPT-3	Trained on WebText2, a library of over 45 terabytes of text data	Generative Pre-trained Transformer 3 architecture	175 billion parameters

4. Data

We have used two different datasets for this report. The first dataset incorporated in this study, the OSDG Community Dataset, comprises paragraph-length text excerpts culled from a wide variety of publicly accessible documents, such as reports, policy papers, and abstracts of various publications. A considerable portion of these documents (exceeding 3,000 in number) are drawn from sources related to the United Nations, including the SDG-Pathfinder and SDG Library, many of which already have SDG labels assigned. Each textual excerpt, typically 90 words in length, includes 3 to 6 sentences. The dataset contains a total of 32,120 such excerpts, with an aggregate of 172,769 labels assigned across them. These texts are all scrutinized by volunteers working with the OSDG

Community Project. More detailed information about the creation of the dataset can be found on the publicly available website of the dataset OSDG, Lab, and PPMI 2021.

The original dataset has 7 columns. It has a doi column which is the digital object identifier of the original document, id column which is the unique text identifier, text excerpt from the document, the SDG label that the text is validated against, negative labels which is the the number of volunteers who rejected the suggested SDG label and similarly positive labels which is the number of volunteers who accepted the suggested SDG label. Lastly, it has an agreement column which is constructed as follows:

$$\text{agreement} = \frac{|\text{labels}_{positive} - \text{labels}_{negative}|}{\text{labels}_{positive} + \text{labels}_{negative}}$$

which gives us the level of "difficulty" of the text for classification purposes.

We would like to report some descriptive statistics. In table 2, you can see the average values of agreement rates, positive labels and negative labels per SDG alongside the counts of each SDG. We see that SDG 16 is the most represented and SDG 12 is the least represented SDGs. However, we believe that there is enough number of texts for each SDG to perform the planned analyses.

From the table, we can also see that on average people are likely to agree with the initial labels. We cannot do a cross-SDG comparison for positive/negative labels from this table because we do not know the total number of annotators for each SDG. For instance, we see much higher positive labels as well as much higher negative labels in SDG 12 compared to other SDGs. This might indicate that more people were analyzing SDG 12. Nevertheless, we can look at the average agreement rates per SDG. We see that SDG 3 has the highest agreement rate and SDG 2 has the lowest agreement rate. This can give us a clue that SDG 3 is easier to be classified whereas SDG 2 is much harder to classify. We also see that the difference between mean positive labels and mean negative labels are not necessarily correlated with the agreement rates.

Table 2. Descriptive statistics of OSDG dataset per SDG

SDGs	Mean Agreement	Mean Positive Labels	Mean Negative Labels	Count
1	0.608138	4.965984	1.898683	2734
2	0.560343	4.532357	2.102971	2457
3	0.776473	5.934176	0.965787	2689
4	0.736174	6.074599	1.488770	3740
5	0.724804	5.372291	1.876210	4338
6	0.673423	3.781883	1.502664	2815
7	0.719593	6.044619	1.467520	3048
8	0.704703	4.502982	0.974818	1509
9	0.643605	7.099762	1.680760	2105
10	0.568350	5.531004	2.491142	2032
11	0.685295	5.256039	1.185332	2277
12	0.576457	9.973827	3.145307	1108
13	0.667196	5.649857	1.664605	2102
14	0.774658	5.411919	1.730061	1141
15	0.610243	6.082594	2.070929	2143
16	0.597043	4.529444	1.569620	5451

Since we also plan to do a keyword analysis, we want to report the most common words in all texts in the form of word cloud. The result can be seen in figure 1. To get this figure, we have performed tokenization with word_tokenize function from the NLTK library. Further, we have

used stopwords.words('english') function from NLTK to get a list of English stopwords, and then these stopwords are removed from the tokenized words. This list contains words such as 'is', 'an', 'the', 'in' etc. For our main analyses, we also give more information about the pre-processing steps in the next subsection.

Figure 1. The most common words in OSDG data

The second dataset that we have utilized in our study is the keyword dataset that we have created. The aim of the creation of this dataset was to pin down the most important words per SDG. To this aim we have scanned through the queries and keywords of various models.

Firstly, we have collected the Elsevier queries 2020. These queries mapped publications to the SDGs by increasing the number of search terms used to define each SDG by taking the customer feedback into account. When they complement these queries with a machine learning model, they were able to improve recall by 10% (Jayabalasingham et al. 2019).

Next, we have scanned through queries of Elsevier 2021 SDG mapping (Rivest et al. 2021). It captures twice as many articles compared to the 2020 version while maintaining precision above 80%. This mapping also demonstrates better overlap with SDG queries from other independent projects. This robust dataset is so comprehensive that Times Higher Education (THE) has adopted it for their 2021 Impact Rankings.

Next, we have incorporated SDG Mapper term matrix which is publicly accessible (European Commission 2023). The keywords could be composed of single or multiple terms and they were meticulously identified by a team of specialists. They were derived based on the established definitions of the goals, targets, and indicators set forth in the SDG agenda. More information can be found in Borchardt et al. 2022.

We also added the keyword provided by the University of Auckland's SDG Keywords Dictionary Project Wang, Kang, and Mu 2023. Their objective is to generate an extended list of keywords capable of pinpointing research pertinent to the SDGs.

Lastly, we have utilized the ontology of OSDG model. The details can be found in the existing models part of this report.

The descriptive analysis of this dataset can be found in the table 3. We have 58020 keywords/word groups in total. Most of the values are made of two words. We see that the keywords for SDG 17 is quite low. This is not a problem as OSDG dataset does not have any label for SDG 17. Another thing to note is that, the reason why there is a lot of keywords for SDG 3 in OSDG Ontology is that it refers to a diverse set of illnesses, bacterias and other kinds of biological terminology as SDG 3 related to good health and well-being.

In the next sub-section, we will discuss the pre-processing for the analyses in the later stages of this study.

Table 3. Number of keywords per SDG across different sources

Resource SDGs	Auckland Keywords	Elsevier queries 2020	OSDG Ontology	Rivest et al.	SDG Mapper term matrix
SDG1	32.0	24.0	104.0	421.0	136.0
SDG2	80.0	33.0	796.0	1116.0	127.0
SDG3	88.0	74.0	23614.0	2305.0	225.0
SDG4	84.0	53.0	269.0	734.0	225.0
SDG5	47.0	80.0	94.0	693.0	221.0
SDG6	105.0	63.0	463.0	1760.0	95.0
SDG7	83.0	56.0	695.0	2485.0	90.0
SDG8	72.0	86.0	220.0	791.0	191.0
SDG9	32.0	41.0	1855.0	1366.0	157.0
SDG10	80.0	27.0	130.0	1707.0	163.0
SDG11	90.0	43.0	1222.0	1082.0	154.0
SDG12	87.0	59.0	302.0	690.0	108.0
SDG13	92.0	61.0	421.0	1452.0	102.0
SDG14	58.0	48.0	381.0	2118.0	111.0
SDG15	101.0	59.0	680.0	1427.0	236.0
SDG16	61.0	72.0	467.0	1150.0	269.0
SDG17	0	0	3.0	0	326.0

4.1 Pre-processing

In this section, we will report the pre-processing steps that we have applied. During fine-tuning of BERT and also the extraction of BERT embeddings, we did not use any pre-processing. This is because the libraries we use for BERT has its own tokenizer and it automatically does lower-casing, normalization and stemming. Furthermore, BERT is pretrained to understand the relationships between words in a sentence, including the order of words. Therefore, we can work with the raw data both for the fine-tuning also for the extraction of embeddings.

For the GloVe embeddings, however, the requirements are slightly different. Unlike BERT, the embedding of a keyword stays the same regardless of the context. Therefore, applied tokenization, lemmatization and also lowercasing. Additionally, we have used the stopwords and punkt from the NLTK package in python to remove punctuation and common stopwords in english.

For the analyses on the transfer learning part, we again did not apply any pre-processing. All of the models that we test are contextual models which can work with the raw texts.

5. Methodology

The methodology section is divided in two part for easier readability. The first section is dedicated for the keyword and embedding analyses using OSDG Community dataset. The second part involves descriptive analyses regarding the transfer learning.

5.1 Analyses of OSDG Community Data

The aim of this study is to provide an understanding for the comparison of various NLP models which aim to classify texts into relevant SDGs. However, this is not an easy task as there are several differences across these models such as the size and the type of the data-set that they were trained on, the annotation procedures, number of parameters, training objective, the architectures that the models utilize and so on. Therefore, we believe that we cannot compare these models on fair

grounds by only getting estimations after a simple fine-tuning procedure. This study proposes a simple baseline analysis to understand the differences across the outputs of these models, if any.

To this aim, we start by identifying the most important keywords for each SDG. Later on, we inspect the embeddings of these keywords to say something about SDG classification. We have four reasons for this approach. Firstly, by analyzing the embeddings of these keywords, we can explore the semantic relationships between different SDGs. For instance, if the keywords related to SDG 3 (Good Health and Well-being) and SDG 6 (Clean Water and Sanitation) have high cosine similarity, it could indicate an overlap in the language used to discuss these two goals. This could have implications for other models' classification task, as documents discussing one goal may also discuss the other. Eventually, we can pin down the hardships that the models encounter in a better way.

Secondly, conducting a keyword analysis can provide interpretable insights that purely black-box model comparisons might not offer. This approach can help other researchers to understand why certain models perform as they do by examining which keywords are influential in their predictions.

Thirdly, conducting a keyword analysis allows us to better understand the domain-specific language used in relation to the SDGs. Understanding the most frequently used words and phrases can provide important insights into how these goals are discussed and represented in text.

Lastly, keyword analyses could also serve as a kind of "baseline" model for classification. Since we need to be creative about finding the ways to compare different models, we think starting with very simple models can help us investigate further.

Keyword Analysis:

After the creation of the keyword data-set mentioned in data part, we decided to run 16 different regression models for each SDG to get the most important keywords. For the sake of the example, we will focus on the case of SDG 1. Same approach is followed for each SDG. We have started by creating dummy variable for the SDG 1 labelled texts. We needed correct labels for texts so we assigned value of 1 to the texts where positive labels were larger than negative labels and the agreement is higher than 0.75. With this, we have ensured that these texts were indeed mostly related to SDG 1. Then, we have counted the occurrences of keywords of SDG 1 in all the texts and created columns for each keyword. Since not every word occured in the OSDG dataset, we had many columns which had 0 values only. Since these columns can easily create multicollinearity issue, we decided to keep the keywords that occur within the texts labelled as SDG 1. This way, we were able to shrink the size of the relevant keywords.

However, we still had dimensionality problem: the number of keywords for each SDG was much larger than the number of texts which are correctly labelled for that SDG. These counts are also reported in the result section. To tackle with this issue, we have decided to run Ridge regression. The reason behind this choice is due to its ability handle multicollinearity among predictors. In text analysis, it is very common to encounter multicollinearity due to the nature of language where certain words or phrases often appear together. Ridge regression manages this by shrinking the coefficients of correlated predictors towards each other.

More specifically, in ridge regression, every regression coefficient is equally penalized. This makes it effective in scenarios where there could be multicollinearity, or where there are a large number of predictor variables. Suppose we have n observations with p predictor variables. Ridge regression aims to minimize the following loss function:

$$L^{\text{ridge}}(\boldsymbol{\beta}) = (y - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}. \tag{1}$$

Here, y is the output variable vector of size n, X is an input data matrix of size $n \times p$, $\boldsymbol{\beta}$ is a coefficient vector of size p, and λ is a positive hyperparameter that controls the Ridge penalty. If the outcome of the minimization provides $\lambda > 0$, it indicates that the coefficients $\boldsymbol{\beta}$ have been shrunk towards 0. The analytical solution for Ridge regression is given by:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y. \tag{2}$$

It's important to highlight that standardization is crucial when dealing with Ridge regression, as the variance of X impacts the results. However, since our features are created by counting the number of occurrences within the text, we believe that this standardization is not needed in our case.

Lastly, we fitted Ridge regression model to predict if a text belongs to SDG 1 or not. We have also performed cross-validation to find the best λ for the Ridge regression. After the model is trained, we computed the accuracy of the model using a threshold of 0.5 to convert predicted probabilities to binary predictions. The best λ and accuracies are also reported in the results.

Embedding Analysis:

After performing keyword analysis to identify the most significant terms related to each SDG, we convert these keywords into their corresponding embedding vectors where we have utilized GloVe and BERT models. These embeddings, which capture the semantic meaning and context of each keyword, are then utilized to group keywords for each SDG. This allows for the creation of SDG-specific semantic spaces, where the proximity between words can be interpreted as semantic similarity.

We choose to work with BERT and GloVe. This is because these two models operate on quite different principles and give us different types of information, both of which can be beneficial for understanding different models' performance. BERT leverages the context of words in a sentence in all directions, thus it's sensitive to the order of words and captures the meaning based on the surrounding context whereas GloVe focuses on word co-occurrence statistics and the overall semantic relationships in the entire corpus. It is a global log-bilinear regression model that combines the advantages of two major models in the field of word embedding: which are the global matrix factorization and local context window methods. We believe that the comparison of these can give us valuable insights into the nature of this classification task.

For GloVe embeddings, we have first trained the model with OSDG Community dataset and then extracted embeddings by using the gensim.models package. The GloVe embeddings are adjusted to be 1×300 vectors. We worked with the embeddings of the keywords which occured in at least one of the correct labelled texts for the corresponding SDG. For the BERT model, we have first fine-tuned the model again with the OSDG Community dataset. Then, we get the CLS tokens by using the transformers package. The graph regarding the loss functions for each epoch during fine-tuning can be found in the appendix. The CLS tokens of the BERT model are 1×768 vectors. Since some of the keywords in our keyword dataset consist of more than one word, we have used mean-pooling to get their total embeddings. This also applies to GloVe embeddings. Later on, we have again used mean-pooling to get embeddings for each SDG. We have used these SDG embeddings to calculate cosine similarity. Before that, we have also applied t-SNE to lower the dimension of embeddings for visualization purposes.

To understand the semantic similarity between each SDG, we used cosine similarity measure. Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. In the context of text analysis, each embedding is a high-dimensional vector, and the cosine similarity is a measure of how similar these vectors (and thus, the corresponding keywords) are. By calculating the cosine similarity between embeddings of the keywords, we can quantitatively evaluate how related the keywords of each SDG are to each other. More specifically, cosine similarity is calculated as follows:

$$\text{Similarity}(A, B) = \cos(\theta) = \frac{A.B}{\|A\|.\|B\|}, \tag{3}$$

where A and B are vectors, θ is the angle between them, $A.B$ is the dot product between them and $\|A\|$ and $\|B\|$ are the L2 norm of the vector. L2 norm of n-dimensional vector A is calculated as

follows:

$$\|A\| = \sqrt{A_1^2 + A_2^2 + A_3^2 ... A_n^2}. \quad (4)$$

The range of cosine similarity is from -1 to +1. A larger cosine value, which corresponds to a smaller angle between the vectors, indicates a higher degree of similarity between the vectors.

5.2 Transfer Learning

As the second chapter in our study, we moved to the comparison of already existing models. We plan to work with OSDG.ai, ChatGPT and Aurora. More detailed description of each model can be found in the existing models part of this study. We named this section as transfer learning because all the models are trained on a large and diverse dataset with the same aim: SDG classification. This pre-trained models has already learned generic features and representations that are useful for various related tasks. Therefore, we believe that they are good candidates to test the performance of transfer learning in this domain as we test them with OSDG Community dataset.

For the estimations of OSDG.ai, we have used the web-based open-source tool. Unlike other models, OSDG.ai is able to give more than one SDG predictions. We have taken this into account when we calculate the evaluation metrics. We hypothesize that OSDG.ai will have the best performance as we are using the data that it had been trained on. Ability of giving multiple-labels is another advantage for this model.

Secondly, we have used ChatGPT's Ada model to get the predictions. We have used our own API key and set the temperature to 0 before we get the predictions to minimize the "creativity" or variance of the answers. We have also used the same quesiton prompt for all the texts to not to introduce any imbalance in our outputs.

Lastly, we have also used Aurora to get predictions. Aurora model also give the probability of each SDG for any text that we give. Similar to the ChatGPT predictions, it gives only one label for each text excerpt.

To be able to interpret the results in the light of several standard evaluation metrics, we needed a correctly labelled dataset for all the SDGs we have in OSDG dataset. To this aim, we have reduced the OSDG dataset in the following way: We have picked 5 text excerpts which had very low agreements, 5 texts excerpts which had high agreements and more positive labels than the negative labels and also 5 excerpts that have high agreements and more negative labels than the positive labels. As we reduced the dataset in this way for each SDG, we ended up with a dataset which has 15 example texts for each SDG. Since we use them only for testing the models, we believe that this number should be reasonable.

The reason why we did the selection in the described way is that we wanted to create a set of "hard" and "easy" texts for classification purposes. We hypothesize that the low agreement texts will result in very disturbed distributions in the estimations of the models. Ideally, models should give SDG distributions close to the uniform distribution as the number of SDGs are the same in the test data. We plan to visually analyze the models in each category. Following that we will compare the performance of the models based on accuracy, recall, precision and F1 score.

In the next section, we will report the analyses described in this section.

6. Results

6.1 Results of Analyses of OSDG Community Data

Keyword Analysis:

First, we will report our findings regarding the keyword analysis. As explained before, the main reason why we applied Ridge regression is the high volume of keywords relative to the true labelled texts in our reduced data. In table 4, the details regarding the ridge regression can be seen including the number of true labelled texts and number of keywords. After applying the filtering technique we have described in the methodology section, we followed k-fold cross validation to get the optimal value of penalty parameter of λ for each model. We have searched the optimal value in an array of numbers that are spaced evenly on a logarithmic scale. The array we choose contains 50 numbers that are evenly spaced on logarithmic scale between 10^{-4} and 10^{4}. As we get some value of λ for some SDGs, we decided to experiment with other arrays, such as an array starts with 10^{1} and 10^{3} to get more granular results. Since the lambda values do not change significantly, we decide to report our initial findings regarding λs. Another thing to note about λ values is that, in SDG 15 and SDG 6 see a relatively higher penalty rate. This indicates that the coefficients (keyword occurances) shrinked toward zero, effectively reducing the model's complexity. In the table, we can also clearly see that the accuracy of Ridge regression remains high for every model.

Table 4. Descriptive statistics for Ridge models per SDG

	Number of true labelled texts	Number of keywords	Best Lambda	Accuracy of Ridge
SDG1	1101	717	51.795	0.97793
SDG2	749	2152	51.795	0.98249
SDG3	1757	26306	11.514	0.96942
SDG4	2234	1365	51.795	0.96234
SDG5	2322	1135	35.565	0.96330
SDG6	1257	2486	159.986	0.97254
SDG7	1749	3409	51.795	0.96498
SDG8	809	1360	75.431	0.98177
SDG9	1016	3451	23.300	0.97757
SDG10	711	2107	25.596	0.98357
SDG11	1190	2591	30.889	0.97230
SDG12	428	1246	10.986	0.99029
SDG13	1037	2128	28.118	0.97793
SDG14	701	2716	65.513	0.98573
SDG15	829	2503	126.486	0.98333
SDG16	1602	2019	79.061	0.96426

In figures 2 and 3, we report the main output of our Ridge regressions: most important keywords. We have used the absolute value of each coefficient to derive these figures as keywords with negative coefficients also play a role to define the characteristics of that given SDG. In the figures, top 15 keywords/phrases are reported. Related to the λ values, we indeed see that coefficients of SDG 6 and SDG 15 has the lowest magnitudes due the shrinkage.

We believe that keyword analyses give us rather successful results. For instance, when we inspect SDG 6 which is clean water and sanitation, we see that the top 3 most important words are water, water quality and aquifer. Similarly, for SDG 7 which is renewable energy, we see electrification, solar and natural gas as top 3 keywords. Another successful example is about SDG 5 which is gender equality: we see woman, violence against women and female as top 3 keywords.

Another thing to note is that, some SDGs have very high coefficients for the top keyword compared to the other keywords such as SDG 10, SDG 15 and SDG 3. One implication of this might be that these SDGs are more "focused" relative to other SDGs as only few coefficient can do a fairly good job at explaining/defining them. This could be another reason why we see a performance difference across the predictions of SDGs for the same model.

Embedding Analysis:

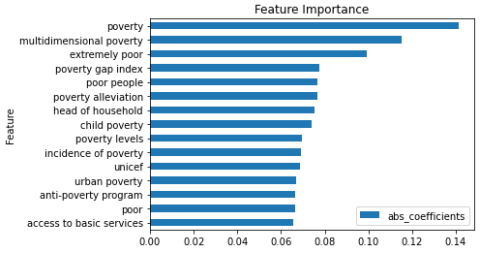
Next, we have utilized BERT and GloVe models to extract the embeddings of the keywords. For interpretation purposes, we decide to visualize them in 2-dimensional space by using t-SNE for dimensionality reduction. The 2D projections of GloVe and BERT embeddings can be seen in figure 4. One way to interpret these figures is to realize the clusters, if any. In the GloVe embeddings, we see small same color clusters in the edges of the set of points concentrated in the middle. In BERT embeddings, we see that color clusters are not very defined in any part of the 2D space. However, this table is hard to read due to the fact that we have 16 different colors. To solve the readability issue, we also reported the visualizations of each SDG embeddings seperately. The results can be see in figure 5. We argue that there is no significant difference in the clustering performance of both embeddings. However, one thing we can notice is the difference across SDGs within one model. For instance, we can argue that SDG 5 is more clustered than SDG 1 in the embeddings of both models. Another relatively more successful clustering can be seen in SDG 7 and SDG 16 as well. Nevertheless, visual inspections do not provide a strong evidence for any interpretation across SDGs.

For this reason, we wanted to create another measure to see the performance of clustering of embeddings across SDGs. To this aim, we have created cosine similarities across each pair of SDGs for both models. The results are reported in figure 6. The values of cosine similarity in any cell of the matrix indicates the degree of similarity between the overlapping SDG row and column. Looking at the overall pictures between these two matrices, we can confidently say that GloVe embeddings differentiates keywords more sharply compared to the BERT embeddings. The cosine similarities between the mean-pooled embeddings of keywords are much lower in GloVe embeddings. We believe that these tables can help us comment on the differences of estimations of different SDG classification models. The overall difference between these two matrices can be explained as follows: BERT's embeddings are context-sensitive, so they might capture more nuanced relationships between words, leading to higher similarities. It could also mean that BERT is recognizing more shared semantic relationships between keywords that might not be as apparent in GloVe.

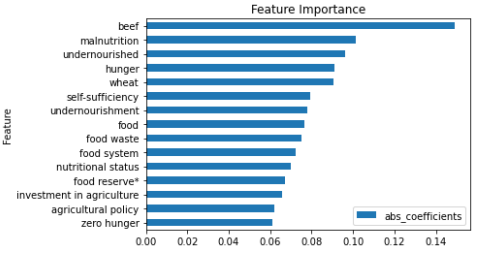
GloVe, on the other hand, is a non-contextual model. It assigns the same vector to a word regardless of its context. It captures global co-occurrence statistics from a corpus. As a result, the relationships represented in the GloVe cosine similarity matrix might be more general and less sensitive to specific contextual nuances. Given these results, we believe that this can be a valuable approach in SDG classification task. If capturing the semantic relations within the text leads to lower performance in terms of prediction, researchers can consider more simple approaches.

A more detailed look in the matrices can also give us the links between SDGs. One of the most problematic parts of the SDG classification is the contextual similarities between some of the SDGs. Even with the human evaluation, it is not always easy to classify texts into one or more SDGs. Due to this reason, it is important to pin down the similarities across SDGs. In BERT embeddings, we see a very high cosine similarities for each pair of SDGs. Therefore, we move the the embeddings of GloVe. We see that cosine similarities of SDG 5 with respect to SDG 6 and SDG 7 are quite low, indicating a sharper differentiation. Another low cosine similarity can be seen between the embeddings of SDG 16 and SDG 7. We believe that some of these results were also reflected in the 2D projections of these SDGs.

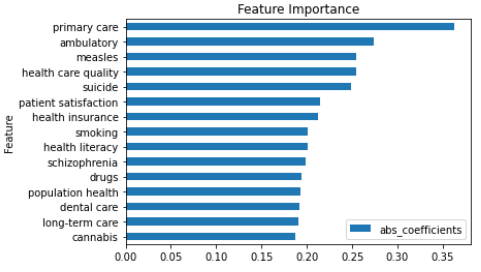
Further, we wanted to analyze the average of GloVe embeddings' cosine similarities for each SDG. We found that the top 3 SDGs which have the highest average similarities are SDG 11, SDG 2 and SDG 9 from highest to lowest. We hypothesize that the models we will test will have the lowest



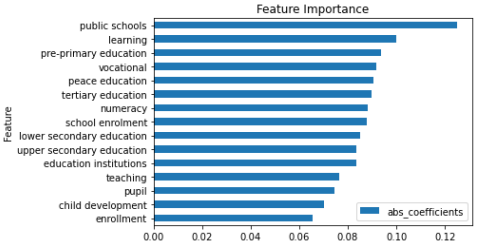
(a) Top 15 phrases for SDG 1



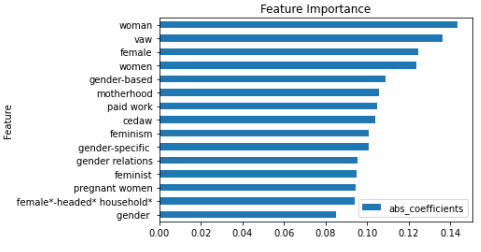
(b) Top 15 phrases for SDG 2



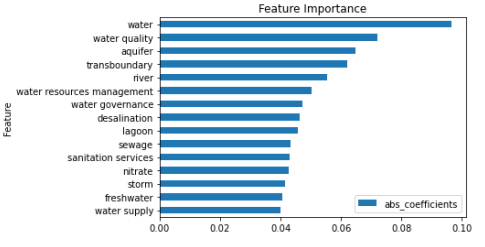
(c) Top 15 phrases for SDG 3



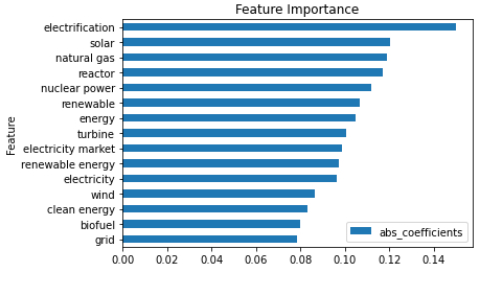
(d) Top 15 phrases for SDG 4



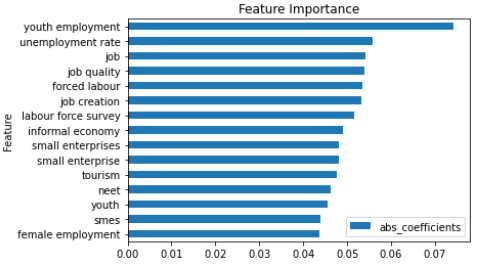
(e) Top 15 phrases for SDG 5



(f) Top 15 phrases for SDG 6

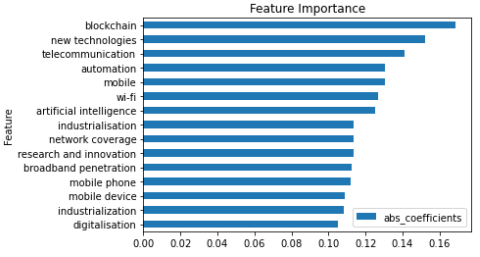


(g) Top 15 phrases for SDG 7

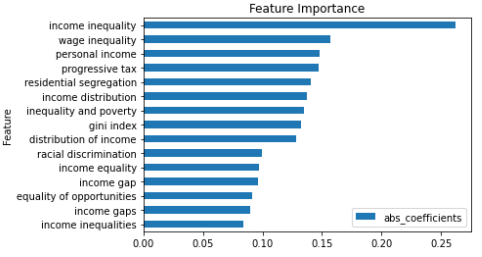


(h) Top 15 phrases for SDG 8

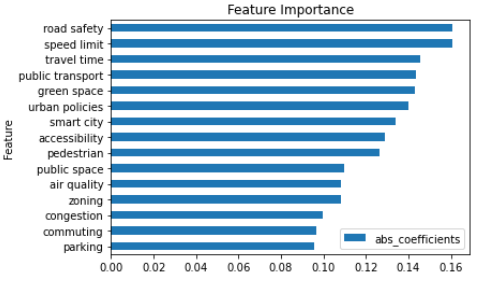
Figure 2. Results of ridge regression for SDG1-8



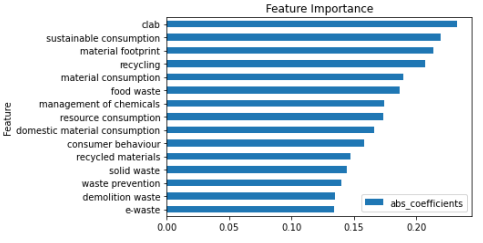
(a) Top 15 phrases for SDG 9



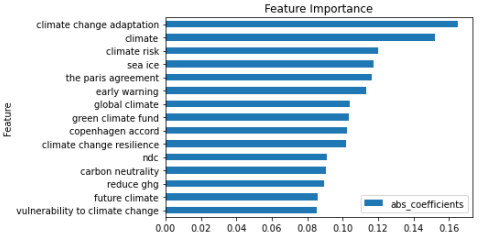
(b) Top 15 phrases for SDG 10



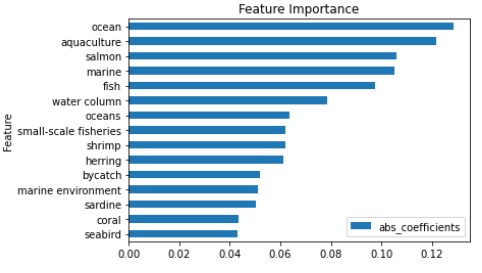
(c) Top 15 phrases for SDG 11



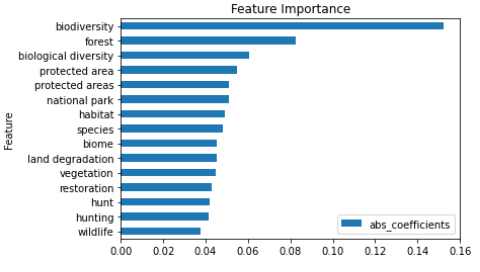
(d) Top 15 phrases for SDG 12



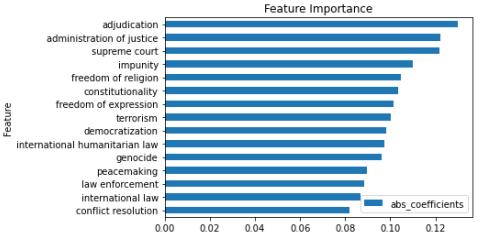
(e) Top 15 phrases for SDG 13



(f) Top 15 phrases for SDG 14

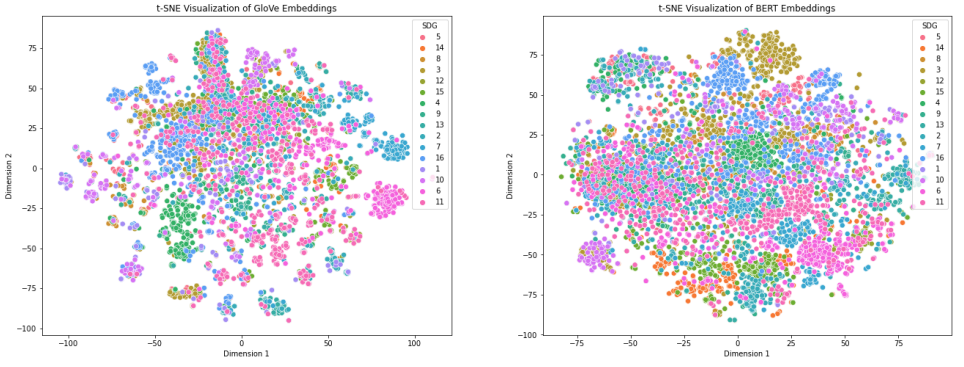


(g) Top 15 phrases for SDG 15



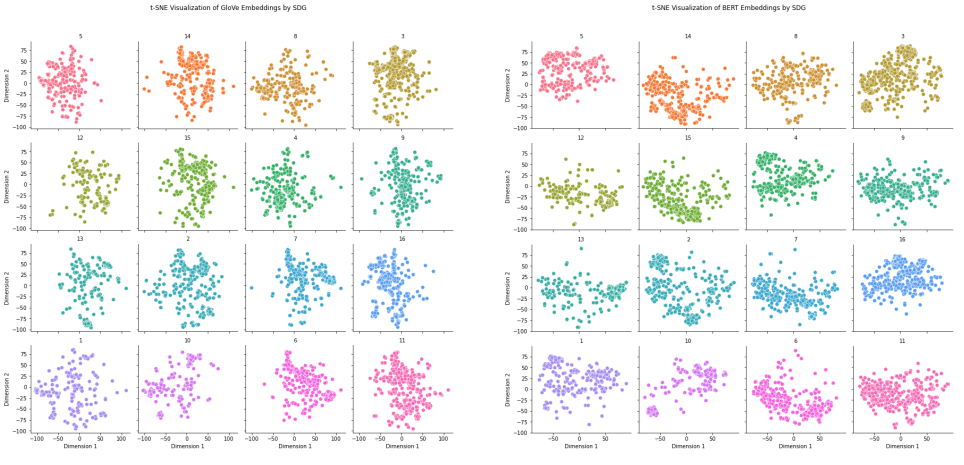
(h) Top 15 phrases for SDG 16

Figure 3. Results of ridge regression for SDG9-16



(a) 2D Projection of GloVe Embeddings **(b)** 2D Projection of BERT Embeddings

Figure 4. 2D Projection of Word Embeddings

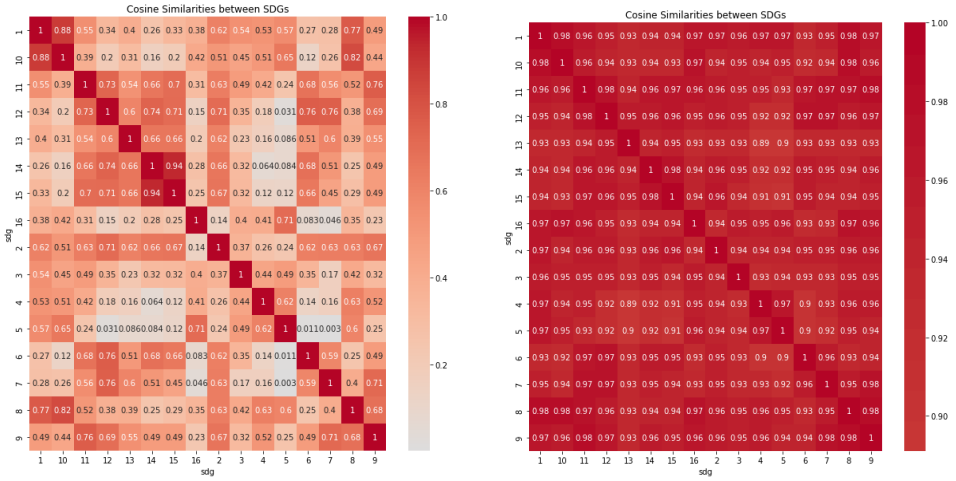


(a) Faceting per SDG for t-SNE visualization of GloVe embeddings **(b)** Faceting per SDG for t-SNE visualization of BERT embeddings

Figure 5. 2D Projection of Word Embeddings per SDG

success rate in these SDGs. Similarly, top 3 SDGs which have the lowest average similarities are SDG 16, SDG 5 and SDG 4 from lowest to highest. We expect higher performance for the predictions of these SDGs.

Next, we will report findings regarding the transfer learning.



(a) Cosine Similarity Matrix for GloVe Embeddings

(b) Cosine Similarity Matrix for BERT Embeddings

Figure 6. Cosine similarity matrix using BERT and GloVe embeddings for SDG groups

6.2 Results of Transfer Learning

We start reporting our analyses with the estimations of the whole reduced dataset that we have created specifically for this section. The whole dataset includes both "hard" and "easy" text excerpts. The result can be seen on figure 7. One thing to note is that, OSDG model can give multiple labels to one single texts whereas the other two cannot. The results should be interpreted by considering these differences.

In figure 7, we see that distribution of the ChatGPT-3 estimations has the closest form to the uniform distribution which is the correct distribution. Ideally, we want to see 15 estimation for each SDG. The SDG that ChatGPT predicts the most is the SDG 4 with 20 predictions whereas the SDG 12 is the least predicted label with 10 prediction. SDG 15, SDG 16, SDG 7 and SDG 10 have predictions closest to 15 which is the correct number of texts for each SDG. Of course, we need to check the accuracy to get a more granular idea.

When we investigate Aurora estimations, we see a distribution far from being uniform. Most of the texts are labelled as SDG 16 and SDG 8. The lowest prediction belongs to SDG 12. By this graph only, we can see that Aurora estimations are likely to do worse than the ChatGPT predictions.

Lastly, we check OSDG estimations. We hypothesized that we would have the best performance here as we are testing the model with the dataset that it is trained on. However, the initial glance does not support this hypothesis. We see that OSDG model is likely to label texts as SDG 1 at a very high rate. The lowest prediction belongs to the 7. Again, we should note that OSDG model can give us multiple labels. This might result in increasing the accuracy but it can also introduce noise to the predictions. In this case, we argue it is the latter case as model tends to label any given text as label 1. Since we could not find any information regarding the selection criteria for labels, we cannot comment further on the performance of the model just by analyzing this graph. It could be that, the model has a very low threshold for passing probability for each SDG. In that case, we could see wrong SDGs like we do in this case. It could be the case that the highest probability that the model gives among SDG labels might actually result in a better performance. Since we do not know the inner dynamics of the model, this requires a further research.

Next, we move the predictions of correctly labelled dataset. As we have described in the methodology part, we have 5 text excerpts for each SDG which has a very high agreement rate from the annotators/community members. This means that these texts are "easy" as humans can easily label them without having much conflicting views. The predictions are reported in figure 8. We again start with ChatGPT predictions. Among the three models, it has the distribution which is closest to the uniform at the count of 5. It has 5 SDGs which has exactly 5 predictions and 9 SDGs which has either 4 or 6.

Aurora predictions seems to do much better in this subset compared to the full dataset. It has 3 SDGs which has 5 predictions. Furthermore, 4 SDGs has either 4 or 6 predictions. Similar to the distribution with the whole dataset, we see that Aurora has a tendecy to label as SDG 1 , SDG 16 and SDG 8. We do not see such a pattern in ChatGPT findings. This might signal the possible difficulty levels across different SDGs.

Lastly, we check the OSDG model. Previously, we saw peaks at SDG 1, SDG 9 and SDG 16. In figure 8 sub-figure c, we again see peak at SDG 9. SDG 1 and SDG 16 are also high. Given these two analyses alongside the results of Aurora, we suspect that these models are struggling with labels of SDG 16, SDG 1 and SDG 9. ChatGPT, on the other hand, does not seem to have such issue. Overall, we also see an improvement in the predictions of OSDG model as it has 4 SDGs which has 5 predictions and 5 SDGs which 4 predictions.

We also analyze the "hard" text examples which has very low agreement rate. This implies that human annotators were also not agreeing with the initial labels of the texts. We expect all the models to perform worse. We hope to see the "biases" of the models more clearly in this section. The results can be seen in figure 9. Indeed, the sub-figures are very different from each other in figure. We see

that Aurora is very keen on predicting texts as SDG 16 or SDG 9 whereas OSDG tends to predict as SDG 1 the most. For some of the SDGs, we even have 0 predictions such as SDG 11 and SDG 12 in Aurora and SDG 6 in OSDG. Compared to these two models, again ChatGPT behaves quite different as it has a more balanced figure across SDGs. Since we also do not have the correct labels for these examples, we do not know success rate for these group. The aim of this specific dataset was to see how each model behaves with the examples that are also hard for humans. Interestingly, we see very different outcomes from each model.

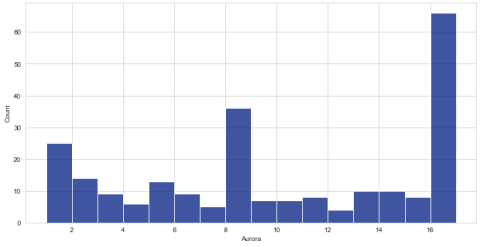
We have also did the similar visualization with the dataset where we have very high negative agreement rates, which means that the initial label offered by the ML part of OSDG is rejected by human annotators at a very high rate. Since the interpretation of this specific subset is rather complex, we added them in appendix for curious readers.

At the last part of the transfer learning, we report the results of some standard evaluation metrics in NLP literature: Accuracy, precision, recall and F1 score. Since we need correct labels to calculate these measures, we have worked with the high positive agreement subset of the data that we have also used in figure 8. The results can be seen on table 5. For the accuracy, we see that ChatGPT has the best value by far with 0.8608 and it is followed by Aurora and OSDG. Since accuracy gives us the number of correct predictions over the total predictions, it is a very straightforward metric. Second, we check the precision across models. We again see that ChatGPT is followed by OSDG and then Aurora. We should again note that OSDG has an advantage at this stage as it gives multiple labels. Precision gives us correctly predicted positive instances out of all instances that the model predicted as positive. That way, we can see that ChatGPT gives us correct prediction %0.87 of a time if it gets correctly. This means there are some false positives in the predictions.

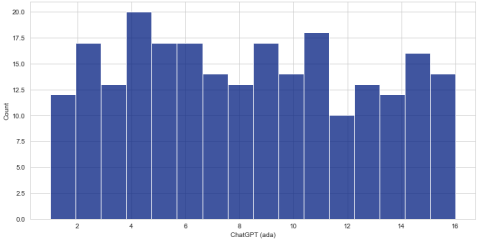
The same ranking among three models also applies to recall and F1 score. F1 score is specifically important if we had a imbalanced dataset, which is not the case here. All in all, we see similar performances for OSDG and Aurora whereas the performance is very high in ChatGPT.

Table 5. Comparison of models based on evaluation metrics

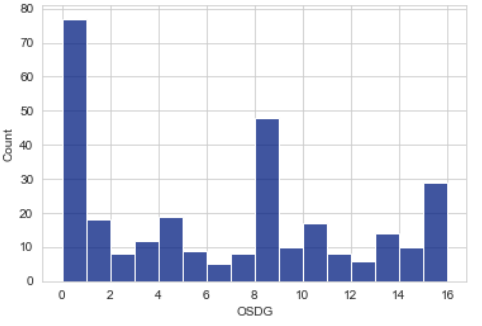
Models	Accuracy	Precision	Recall	F1 Score
Aurora	0.6203	0.6064	0.6471	0.6261
ChatGPT-3	0.8608	0.8686	0.8625	0.8655
OSDG	0.6709	0.6552	0.6719	0.6634



(a) Aurora Estimations

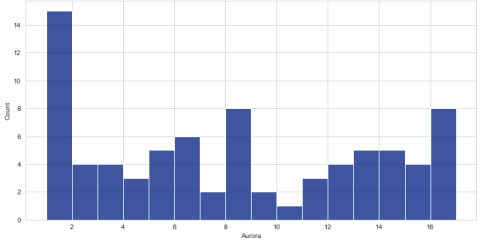


(b) ChatGPT Estimations

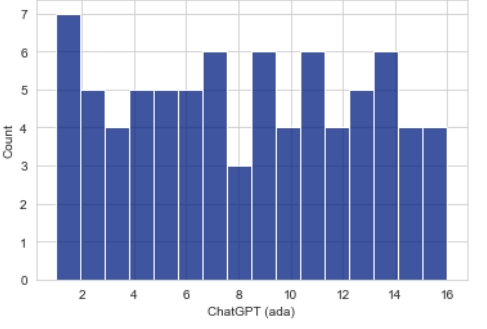


(c) OSDG Estimations

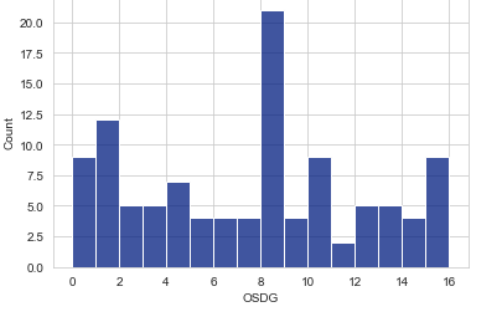
Figure 7. SDG Estimations with the whole reduced dataset



(a) Aurora Estimations

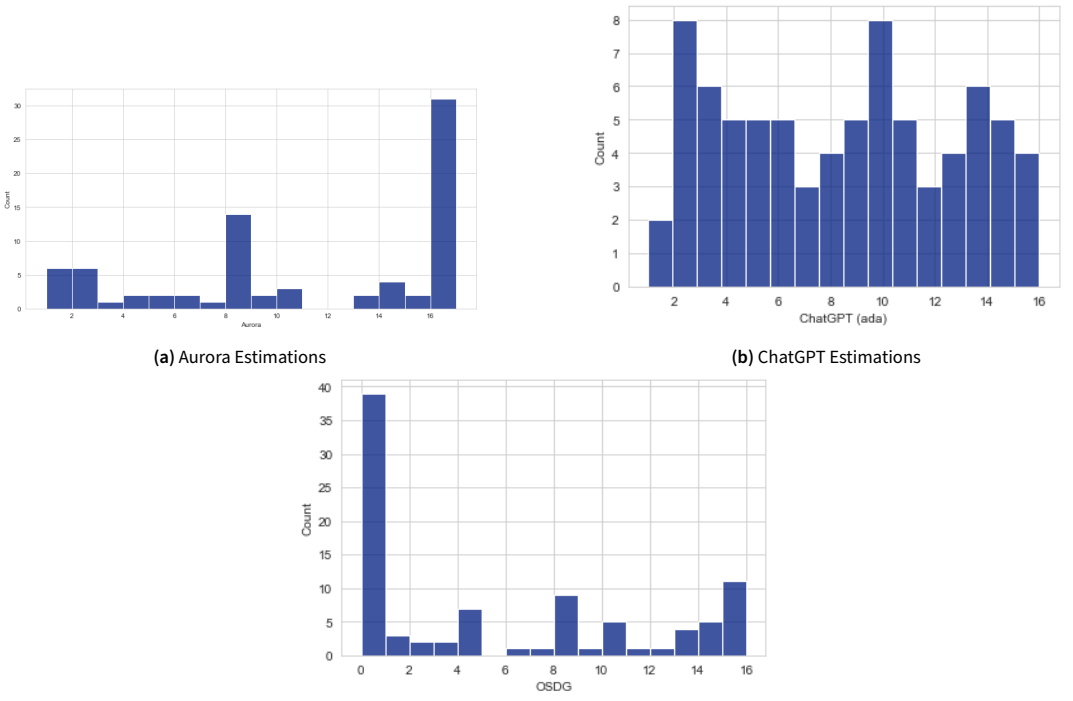


(b) ChatGPT Estimations



(c) OSDG Estimations

Figure 8. SDG Estimations with the correctly labelled dataset



(a) Aurora Estimations

(b) ChatGPT Estimations

(c) OSDG Estimations

Figure 9. SDG Estimations with the low-agreement dataset

7. Discussion

The findings we have found are categorized into two section: keyword/embedding analysis and transfer learning. In the first part, we have pinned down the most related keywords for each SDG after creation of our comprehensive and a unique dataset and analyses of ridge regressions. Later on, we have compared the keyword embeddings of BERT and GloVe models. We have choose to follow this methodology as the first step as it can provide interpretable insights to black-box model comparisons could not offer. In the results, we see that GloVe does a fairly better job in clustering each SDG's keywords into a distinct group. In the cosine similarity matrices, the performance difference becomes very clear. We have discussed the reasons behind this difference thoroughly in the results part. In this part, we would like to provide a link between these findings and what we have seen in the transfer learning.

In cosine similarity matrices, we saw that SDG 5, SDG 16 and SDG 4 had the lowest average similarity in all SDGs. Consequently, we have expected to have a better performance when predicting these SDGs. When we check the frequencies of these SDGs for predictions made with correctly labelled dataset in figure 8, we don't see a significantly better performance for these SDGs. On the other hand, the SDGs that have high average cosine similarity were SDG 2, SDG 9 and SDG 11. When we check their performance in various models, we see that OSDG has a peak on SDG 9, indicating a higher number of predictions for this SDG than the correct number. We can see this on in all three different data specifications which are reported in 7, 8 and 9.

Beside this, the most surprising result is the performance of OSDG data. Even though we have used a sample from its training dataset to do these test, it did not resulted in a high accuracy, recall or precision. Moreover, it has an ability to give multiple labels for predictions. Even though this feature makes it harder for us to compare it with other models, we expected it to improve the recall as it

can predict more classes it misses fewer that should be predicted. Consequently, we can expect a lower precision because it can potentially make more incorrect predictions than other models. For accuracy, the interpretation is not clear. In the light of these, we still see a big difference between OSDG and ChatGPT. We believe that this is one of the most important result of this paper.

The interpretation of the results of the transfer learning section is quite complex. We believe that we should start by looking at the overall differences between these models which is described in table 1. One of the explanations for these results is the differences on the training datasets that each of these models utilize. OSDG model uses the human-annotated dataset which is created with abstracts, policy documents, news reports etc. Compared to this, we see that Aurora is trained only on English abstracts. ChatGPT-3, on the other hand, is trained on a huge size of internet data. We believe that this difference in the level of comprehensiveness had an effect on the findings. We found that as the training dataset becomes more specific, the performance in evaluation metrics decrease.

Another explanation can potentially stem from the difference on the base architectures of these models as we can see in table 1. The level of complexity in these base models results in differences in the ability of capturing the intricate patterns. This can explain the difference between ChatGPT-3 and Aurora.

Lastly, the number of parameters potentially plays a role in these findings. Higher number of parameters enables models to capture a more granular semantic relations in the input text and hence result in a more accurate predictions. Especially in a complex task like SDG classification, this difference can be detrimental.

Given all these initial differences across these models, one can easily ask whether it is even possible to compare these models on a fair basis or not. In an ideal world, it would be better to change only one feature across the models and do the tests again to see the causality clearly. However, in the comparison of NLP models, this is not very feasible. Therefore, we wanted to provide a simple embedding and a keyword analysis in the first part to be able to create a baseline for the upcoming research.

In the light of our results in both parts, we can argue there are two intuitive ways to solve the SDG classification problem. First one is applying simple frequency-based simple models such as GloVe like we demonstrated. The biggest advantage of these type of models is that they provide a very clear and open solution which is easy to interpret due to their transparency. Due to their simplicity, it is also easier to make required amendments in the model with respect to what is needed.

The second groups of models that our research would recommend is the models with high complexity such as ChatGPT. Contrast to the first option, they are usually black-box models where it is not easy to change the inner dynamics. However, they can provide very granular insights to this specific problem as they are able to capture intricate patterns.

The models that fall in-between these two options, like BERT models, do not offer the same level of transparency as basic models nor the same depth of insights as complex models. Hence, they may not be the first-best option given all the other choices. However, researchers in this field should also be guided by the objectives of their own task. For instance, for a project where which solely focuses on financial documents, middle ground models can still provide the needed results.

8. Conclusion

In this study, we have aimed to propose a comparative study between various language models which are trained to solve the task of SDG classification. While our results cannot be generalized to a broad range of researcher agendas, we believe that it provides a good insight to other studies which aims to learn more about the reasons behind the high variance in the outputs of the similar models. We also acknowledge that this is a very complex task where it is not always easy to provide a fair results with respect to every model. However, we believe that our findings can give the initial clue on what the potential important things to look at when making an assessment across similar type of language models.

Our main findings suggests that simple models provide a sharper clustering when it comes to keyword analysis. Moreover, we also found that ChatGPT provides an overall better performances in both in the visual inspections but also in the standard evaluation metrics. We argue that this is due to its ability to assess the given texts at a more granular level while capturing intricate patterns. On the other hand, for models that utilize BERT's variations, we found evidence for biases towards specific SDGs.

There are several points in this field that we think requires further research. The first suggestion is related to the further interpretability of complex models like ChatGPT. In the task of SDG classification, developing methods that provide insights into how these complex models arrive at their decisions can foster greater trust in their predictions. Another way to deal with this task is constructing ensemble models where we can combine the power of both simple and complex models like we see in the example of Hsu, LaFleur, and Orazbek 2022. We believe that exploring this option can improve overall accuracy and mitigate the limitations of individual models. Lastly, human-in-the-loop strategies which combines machine predictions with human expertise can benefit the field and provide extra robustness to existing models.

References

Bird, S, E Loper, and E Klein. 2009. Natural language toolkit (nltk) book. *O'Reilly Media Inc.*

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* 55 (4): 77–84.

Borchardt, Steve, Vignola Giulia Barbero, Daniela Buscaglia, Michele Maroni, and Luisa Marelli. 2022. Mapping eu policies with the 2030 agenda and sdgs.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

European Commission, Joint Research Centre. 2023. Sdg mapper, https://knowsdgs.jrc.ec.europa.eu/sdgmapper.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056.*

Guisiano, Jade Eva, Raja Chiky, and Jonathas De Mello. 2022. Sdg-meter: a deep learning based tool for automatic text classification of the sustainable development goals. In *Intelligent information and database systems: 14th asian conference, aciids 2022, ho chi minh city, vietnam, november 28–30, 2022, proceedings, part i,* 259–271. Springer.

Hsu, D Frank, Marcelo T LaFleur, and Ilyas Orazbek. 2022. Improving sdg classification precision using combinatorial fusion. *Sensors* 22 (3): 1067.

Jacouton, Jean-Baptiste, Régis Marodon, Adeline Laulanie, et al. 2022. *The proof is in the pudding. revealing the sdgs with artificial intelligence.* Technical report.

Jayabalasingham, Bamini, Roy Boverhof, Kevin Agnew, and Lisette Klein. 2019. Identifying research supporting the united nations sustainable development goals. *Mendeley Data* 1 (1).

Joshi, Amit, Luis Gonzalez Morales, Szymon Klarman, Armando Stellato, Aaron Helton, Sean Lovell, and Artur Haczek. 2021. A knowledge organization system for the united nations sustainable development goals. In *The semantic web: 18th international conference, eswc 2021, virtual event, june 6–10, 2021, proceedings 18,* 548–564. Springer.

Kashnitsky, Yury, Guillaume Roberge, Jingwen Mu, Kevin Kang, Weiwei Wang, Maurice Vanderfeesten, Maxim Rivest, Lennart Keßler, Robert Jaworek, Maéva Vignes, et al. 2022. Identifying research supporting the united nations sustainable development goals. *arXiv preprint arXiv:2209.07285.*

Kim, Namsuk, Marcelo LaFleur, et al. 2020. *What does the united nations "say" about global agenda?: an exploration of trends using natural language processing for machine learning.* United Nations, Department of Economic / Social Affairs New York, NY, USA.

LaFleur, Marcelo. 2019. Art is long, life is short: an sdg classification system for desa publications.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

Morales-Hernández, Roberto Carlos, Joaquın Gutiérrez Jagüey, and David Becerra-Alonso. 2022. A comparison of multi-label text classification models in research articles labeled with sustainable development goals. *IEEE Access* 10:123534–123548.

OSDG, UNDP IICPSD SDG AI Lab, and PPMI. 2021. Osdg community dataset (osdg-cd). This CSV file uses UTF-8 character encoding. For easy access on MS Excel, open the file using Data → From Text/CSV. Please split CSV data into different columns by using a TAB delimiter. (October). https://doi.org/10.5281/zenodo.5550238. https://doi.org/10.5281/zenodo.5550238.

Pukelis, Lukas, Nuria Bautista-Puig, Gustė Statulevičiūtė, Vilius Stančiauskas, Gokhan Dikmener, and Dina Akylbekova. 2022. Osdg 2.0: a multilingual tool for classifying text data by un sustainable development goals (sdgs). *arXiv preprint arXiv:2211.11252.*

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Rivest, Maxime, Yury Kashnitsky, Alexandre Bédard-Vallée, David Campbell, Paul Khayat, Isabelle Labrosse, Henrique Pinheiro, Simon Provençal, Guillaume Roberge, and Chris James. 2021. Improving the scopus and aurora queries to identify research that supports the united nations sustainable development goals (sdgs) 2021. *Mendeley Data* 2.

Shen, Dinghan, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline needs more love: on simple word-embedding-based models and associated pooling mechanisms. *arXiv preprint arXiv:1805.09843.*

Vanderfeesten, Maurice, Robert Jaworek, and Lennart Keßler. 2022. AI for mapping multi-lingual academic papers to the United Nations' Sustainable Development Goals (SDGs) (March). https://doi.org/10.5281/zenodo.6487606. https://doi.org/10.5281/zenodo.6487606.

Wang, Weiwei, Weihao Kang, and Jingwen Mu. 2023. Mapping research to the sustainable development goals (sdgs).

Appendix 1. Fine-tuning of Base-BERT-Uncased Model

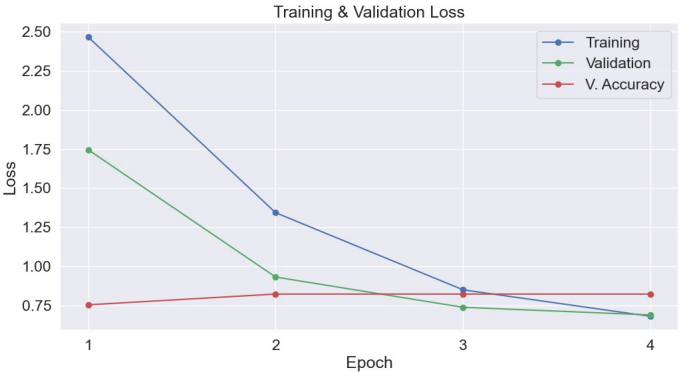
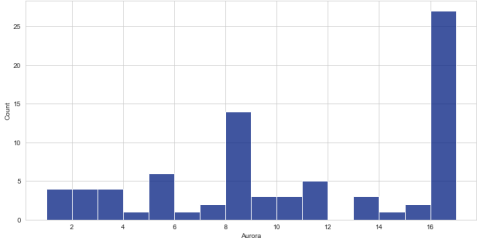


Figure 10. Fine-tuning process of base-BERT-uncased

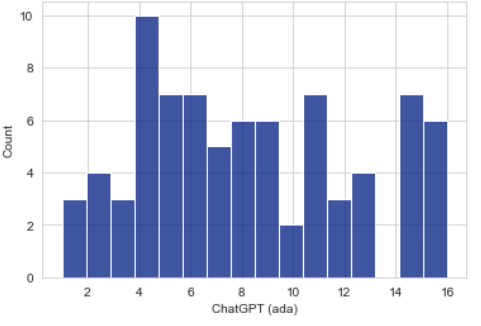
In figure 10, the training and validation loss as well as validation accuracy have been reported for each number of epochs. We see a sharp decline in the losses both in training but also in the validation on the second and third epochs. The rate of decrease becomes much slower between third and fourth epoch. Due to this reason, we believe that we do not need more epochs for fine-tuning. An important thing to note about this figue is that the training loss always remanins on top of validation loss which is an indicator for the absence of over-fitting.

Appendix 2. Predictions for dataset with high negative agreements

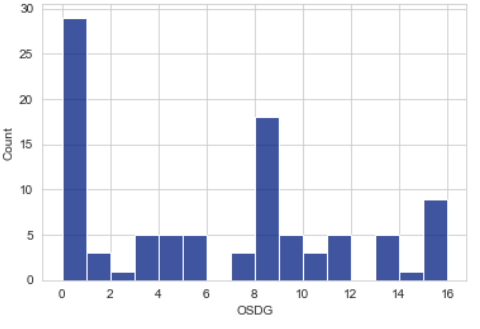
In figure 11, we report the predictions of all the three models we have for the dataset made of text excerpts where we have high negative agreement rates. This means that, most of the human-annotators disagree with the initial labels that the OSDG's machine learning algorithm provides. Since we do not know the correct labels for these texts, we cannot provide an intuitive explanation for these results. However, it should be noted that the previous biases of Aurora and OSDG models can be seen clearly: We see peaks in SDG 16 and SDG 8 in Aurora and peaks in SDG 1 and SDG 9 in OSDG. These patterns were also existent in the previous datasets. In ChatGPT, we do not detect a similar bias.



(a) Aurora Estimations



(b) ChatGPT Estimations



(c) OSDG Estimations

Figure 11. SDG Estimations with the correctly labelled dataset