A Transfer Learning Approach for SDGs Classification of Sustainability Reports

Ata Nizamoglu[*◇], Lea Dahm[*◇], Talia Sari[*◇], Vera Schmitt[*], Salar Mohtaj[*†], Sebastian Möller[*†]

[*]Technische Universität Berlin, Berlin, Germany
[†]German Research Centre for Artificial Intelligence (DFKI), Labor Berlin, Germany

{nizamoglu | lea.dahm | t.sari}@campus.tu-berlin.de
{vera.schmitt | salar.mohtaj | sebastian.moeller}@tu-berlin.de

[◇] The first three authors have contributed equally

Abstract

In 2015, The United Nations (UN) provided a blueprint for sustainable development in various domains. This Blueprint described 17 Sustainable Development Goals (SDGs), such as "No Poverty", "Zero Hunger", or "Gender Equality". Subsequently, many companies have started publishing yearly sustainability reports, explaining their efforts with respect to the SDGs. However, the manual assessment of these reports is an infeasible task, and the automatic processing of text documents is necessary to aggregate information about the distribution of SDGs throughout various domains. In this research, we have developed and measured the performance of various natural language processing models from classical to transfer learning-based models to identify the targeted SDG in sustainability reports. Hereby, transformer-based models show the best performance for this task, especially BERT-based models, such as RoBERTa. The results show, that the approach of automatically processing text documents to classify SDGs in various documents is feasible and can be used to aggregate information about which SDGs are covered by which companies and industry domains.

1. Introduction

Various types of organizations such as private entities and scientific and governmental institutions publish documents reporting about activities related to Sustainable Development Goals (SDGs). In 2015, all member states of the United Nations agreed upon the 2030 Agenda for Sustainable Development, which serves as a framework for addressing poverty, improving health and education, reducing inequality, promoting economic growth, and addressing the challenge of climate change (United Nations, 2015). The UN member states have defined 17 SDGs along with their corresponding targets that need to be achieved. As a result, the private sector has taken steps to address the SDGs by releasing annual sustainability reports, which detail the actions taken by companies to tackle the challenges related to the respective SDGs.

The information used to create the *Annual SDG* progress report comes from a variety of sources, including policy recommendations, sustainability reports, and progress reports. SDG experts analyze these sources and consolidate the information they contain in order to create the report (Guisiano et al., 2022b). However, identifying which SDGs are addressed in text documents published by both the private and public sector is a time-consuming task. This is particularly challenging as the number of documents addressing the SDGs increases each year, making manual scanning and classification a daunting task (Mhlanga et al., 2018). The use of machine learning models can aid in scanning large volumes of documents for content related to the SDGs, thus supporting the process of identifying and analyzing sustainability-related information on a large scale.

There have been attempts to address the infeasibility of manual assessment of the massive amount of SDG-related text documents. Most of the approaches focus on applying deep learning models to automatically classify SDGs in text documents, such as the *SDG-Meter* applying transfer learning models to map SDGs in progress reports (Guisiano et al., 2022b), and sustainability report classification based on the Open Source SDG (OSDG) Community Dataset (Angin et al., 2022), or SDG-oriented artifact detection in various types of text documents (Hajikhani and Suominen, 2022).

In this work, the focus is on analyzing various text sources (e.g. scientific publications and information from the UN website), which can vary greatly in terms of length, word count, and structure. There are no standards defining how to structure and describe any efforts made to address one or more of the 17 SDGs. Thus, the manual assessment of SDGs of different text sources is very challenging, and automated procedures are required to process the increasing amount of text sources and aggregate the activities and progress made by different industries, companies, and scientific and public institutions.

Therefore, in this paper different natural language processing (NLP)-based models will be applied to assess their performance in classifying the 17 SDGs in different text sources. For this purpose, a new dataset containing $41,351$ sentences from various text sources addressing SDGs has been gathered. The dataset has been used for training and testing the NLP models (mainly transformer-based models), such as BERT, RoBERTa, XLNet, and a stacking model under which we combined the predictions from the four best-performing models. Overall, RoBERTa achieves the best performance of $86,3\%$ F1, which is in line with the findings from (Guisiano et al., 2022b). The main contributions of this work are as follows:

1. Collect and scrape SDG-related text data from vari-

ous text sources ($41,3k$ sentences),

2. Apply different data-preprocessing strategies for balancing the class distribution to improve the overall performance of teh transformer-based language models,

3. Analyzing the performance of state-of-the-art pretrained transformer models on the task of automated SDGs classification as a multi-class classification task.

The rest of the paper is organized as follows; Section 2. summarizes the state-of-the-art literature on the classification of SDGs in different domains. An overview of the dataset and the pre-processing steps are presented in Section 3., and the experimental setup and results are discussed in Sections 4. and 5., respectively. Finally, we conclude the paper and the system in Section 6..

2. Related Work

The number of documents reporting SDGs by companies and (international) organizations continues to increase, and new approaches to processing this information have been proposed. Hereby, NLP methods significantly contribute to developing solutions to automatically process large text documents reporting progress made with respect to SDGs. Recent advancements in deep learning for various NLP tasks have led to the development of large language models showing high performance for complex NLP tasks (Angin et al., 2022). Hereby, transformer-based methods show the most promising results in detecting SDGs in text documents, achieving a high-performance (Angin et al., 2022). For the processing of scientific reports, Smith et al. (2021) have assessed *Doc2vec* in combination with clustering (Le and Mikolov, 2014) to analyze similarities of SDGs in scientific research documents (Smith et al., 2021). Transformer-based models have been applied by Guisiano et al. (2021), who developed a tool based on the Bidirectional Encoder Representation from Transformer (BERT) model (Devlin et al., 2018) to facilitate faster processing and classification of the SDGs in text, by focusing on SDG 17 (Guisiano et al., 2022b). Yet, the UN emphasizes that the SDGs are interlinked and applying models to detect one SDG, will facilitate working on the other SDGs as well (United Nations, 2015). For example, improving health and education are fundamental elements to ending poverty and reducing inequality. (Smith et al., 2021) used NLP methods to analyze interdependencies between the goals, aiming to provide insight into overlaps in public conversation. The findings indicated that certain terms played a central role in addressing multiple SDGs. For instance, the term "gender" was found to be significant in discussions pertaining to both goal 5: gender equality, and goal 4: quality education. Another example is the term "development assistance", which was commonly referenced in relation to goal 2: zero hunger, goal 3: good health and well-being, goal 10: reduced inequalities, and goal 17: partnerships for the goals. The positive and negative correlation between indicators of the goals is analyzed in (Pradhan et al., 2017). This research assessed the degree to which the 231 indicators that comprise the 17 goals are complementary to each other, or are trade-offs. The findings of this study revealed that most indicators were considered to be synergistic within and across different SDGs. Nevertheless, certain indicators for particular goals were found to be contradictory to each other. While the automatic detection of SDGs in text documents does not enable a qualitative assessment of the impact of the mentioned SDGs or identify instances of "greenwashing" in the text, it can assist in consolidating information on which SDGs are being addressed by various companies or industry domains. This, in turn, can streamline further processing of the information. Thus, this research focuses on the multi-class classification of SDGs in sustainability reports. Specifically, the interrelated nature of the SDGs is treated as a multi-label task, where the classifier can assign sentences to one or more labels (Guisiano et al., 2022a).

3. Dataset

The subsequent section outlines the data collection procedure and the dataset utilized for training and evaluating NLP models for the task of automatically classifying SDGs.

3.1. Data Collection

To evaluate different NLP model performances with respect to the SDG classification task, a dataset was used containing 2219 sentences with the corresponding SDG labels[1]. The dataset covers two main sources: (1) scientific papers and sustainability reports from different companies, and (2) the SDG descriptions of the United Nations. From each resource, the SDG-related sentences have been extracted and aligned with the corresponding labels. The data is split into train and test sets, where the train set contains 37216 instances and the test set has 4135 instances. Some key statistics from the dataset are presented in Table 1.

Attribute	**#**
Number of instances (i.e., sentences)	41351
Max length of texts (in character)	1931
Min length of texts (in character)	11
The average length of texts (in character)	547

Table 1: Summary of statistics and frequency distribution of the dataset

Figure 1 highlights the difference between the length of texts in the dataset for each source. As depicted in the figure, the sentences from scientific papers tend to be longer compared to sentences from the UN source.

Furthermore, Figure 2 shows the frequency of how often different SDGs are represented in the dataset. Hereby, it is clearly visible that the dataset is not balanced, as the frequencies of the SDGs differ significantly. SDGs 5 and

[1]The dataset and the related code can be found in the following GitHub Repository: `https://github.com/ataniz/SDGs-Classification-of-Sustainability-Reports`

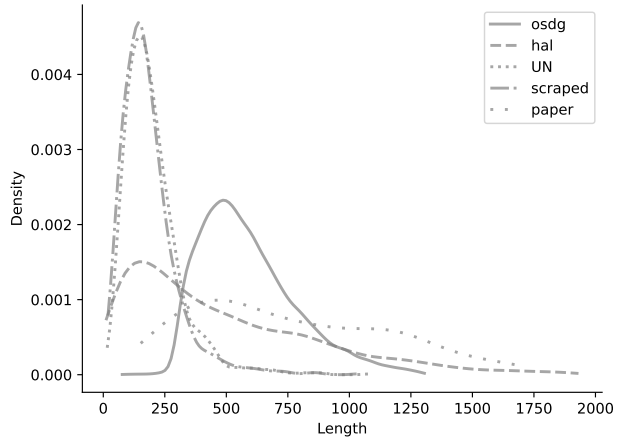

Figure 1: The distribution of sentence lengths from paper and UN sources

4 are over-represented in the dataset, whereas SDGs 16 and 17 are underrepresented. Our dataset's imbalance can be attributed to the fact that it was collected from diverse sources across different countries. Survey studies conducted by (Kleespies, 2022) have indicated that the correlation between SDG counts tends to be highest in the environmental sector of various countries. The importance of each SDG goal for developing and developed countries, and for high- and low-income countries, is considered differently. For example, in some countries, where students receive an affordable and high-quality education, SDG 4 is considered far less important than other SDGs. A more detailed SDG rating graph according to various countries can be observed in (Sachs, 2022).

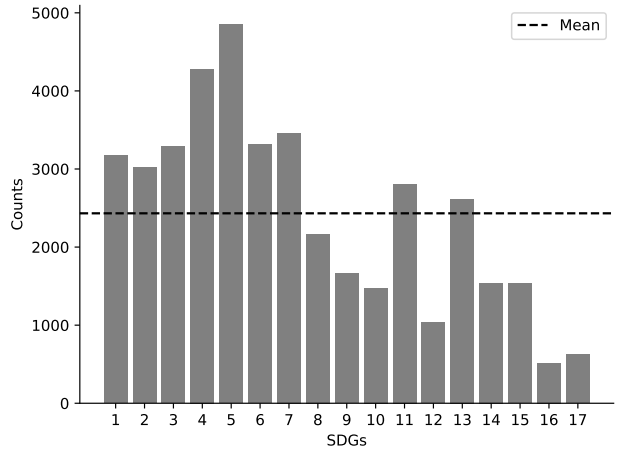

Figure 2: Number of instances per SDG in the dataset

In addition to the mentioned dataset, we scraped data from various sources in order to aid the model's generalization. In total, $41,351$ sentences were added to the training set and used in the training phase. We scraped almost $3,000$ sentences from sustainability reports, and also retrieved the remaining data from (Pukelis et al., 2020) and (Guisiano et al., 2022b)).

For scraping more data concerning sustainability reports two sources have been used: (1) `https://www.sustainability-reports.com/` (Reporting, 2002), and (2) `https://www.online-report.com/` (Nexar and Group, 2003), which aggregate corporate sustainability reports. Hereby, the respective pdfs concerning different sustainability reports have been downloaded first and the text was extracted using the python library *PyPDF2*. For the labeling task, a list of keywords associated with each SDG was compiled to label each sentence separately. A certain number of sentences has been manually verified, to assess how well the automatic labeling task was performed. Since our models are not for multi-label-dataset, we decided to assign sentences with multi-label to just a single label. For example, the sentence "More than 130 million women and girls around the world lack access to education, and women account for two-thirds of the 750 million adults without basic literacy skills" could be labeled as SDG 4 or SDG 5. Since the sentence focuses more on women's lack of access to education, this sentence has been labeled as SDG 5. Furthermore, additional data was used to extend and balance the dataset. For further balancing attempts, synthetic data generation approaches have been used, such as the python library *NLPaug*[2], to replace synonyms and back-translate text, which is also known as *round trip translation*. However, up-sampling approaches based on data augmentation did not improve the overall performance of the transformer-based models. The results in the following section are based on the original imbalanced dataset.

3.2. Pre-processing

The data has been scraped from different sources, which requires the application of text normalization as a pre-processing step to improve the generalization of applied NLP models.

Not only text-normalization has been applied but also further pre-processing steps such as:

1. spelling correction,
2. exclusion of non-English text,
3. duplicate sentence removal,
4. removing email addresses, hashtags, URLs, emojis and emoticons, footnote references, and HTML elements

Moreover, we performed normalization of non-Unicode characters, *noisy* characters, signs, and symbols (e.g., bullet points and hyphenated words quotes). The pre-processing steps were mainly done using Regular Expressions (regex) and the textacy[3] Python library. In addition to these pre-processing steps, we applied stop word removal, lower-casing, and stemming on the input text for classical models in our experiments.

To ensure an accurate evaluation of the model's performance, we split the dataset into different subsets for training, development, and testing. For classical models,

[2] `https://pypi.org/project/nlpaug/`
[3] `https://pypi.org/project/textacy/`

we split the dataset with a 9:1 ratio of training to testing data. For transformer models, we used a split of 8:1:1 ratio of training, development, and testing data, respectively. This split allows us to fine-tune the transformer models on the development set and evaluate their performance on the test set, ensuring that the final results are robust and generalizable.

4. Experiments

In the following we will describe the experimental setup and the implementation of NLP models to classify the SGDs in text data, ranging from traditional NLP models to state-of-the-art transformer-based techniques. As the baseline model, the Naive Bayes and a Support Vector Machine (SVM) model have been implemented, since they have shown promising results in text classification tasks in various domains (Luo, 2021; Xu, 2018). The input texts were converted into vectors based on the **Term Frequency - Inverse Document Frequency** (TF-IDF) weighting schema. In addition to the traditional NLP models, we also fine-tuned a number of pre-trained language models on the training and the scraped datasets. The pre-trained language models have outperformed the classical NLP models in many studies on text classification in different domains such as fake news detection (Mohtaj and Möller, 2022a). Transformer-based models such as BERT (base and large) (Devlin et al., 2018), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), and Ernie 2.0 (Sun et al., 2020) have been implemented. For these models, an additional dense layer has been added with an output size of 17 on top for the classification task.

We used *HuggingFace* (Wolf et al., 2020) and *Flair*(Akbik et al., 2019) Python libraries in *PyTorch* (Paszke et al., 2019) to implement the transformer-based models. Furthermore, a dropout layer has been added to prevent over-fitting, and a linear layer and the softmax activation function have been used to fit the architectures to the multi-class classification task.

Regarding the hyper-parameters, we tested a range of values for different parameters including the *learning rate* (1e-6, 3e-6, 1e-5, 2e-5, 3e-5¡ 4e-5, and 5e-5), dropout probability (0, 0.1, 0.3, and 0.6), and mini-batch size (4, and 8). We used the *AdamW* (Loshchilov and Hutter, 2019) optimizer in all of the experiments and fine-tuned all of the models in 10 epochs.

Hereby, the learning rate of 1e-5 yielded the best F1-score. The dropout probability was set ¡y 0.1 and the mini-batch size to 4 in the final experiments since they showed the best performance compared to the other values.

5. Results

For the evaluation of the applied approaches, the evaluation criteria of accuracy, precision, recall, and F1-score have been used to compare the model performances. The macro F1-score values are presented in Table 2 for all models.

As it is highlighted in Table 2, almost all of the state-of-the-art transformer-based models could outperform Naive Bayes as a traditional model. However, the SVM model shows competitive results compared to the pre-trained models on the task. Among all models that we have implemented for the SDG classification task, *RoBERTa-large* yielded the best results with a meaningful margin compared to the second best model *XLNet-large*.

Classifier	**F1-score**
Naive Bayes	0.572
SVM	0.775
BERT-base	0.770
BERT-large	0.829
Ernie2.0	0.795
XLNet-large	0.836
RoBERTa-large	**0.863**

Table 2: The macro F1 score retrieved by different models

6. Discussion and Conclusion

In this research, the application of different NLP models for the SDG classification task has been explored. The data has been obtained from various text sources, such as scientific publications, the UN description of SDGs, and sustainability reports from different companies. NLP models such as SVM, Naive Bayes, and transformer-based models have been implemented to assess their performance in terms of the macro F1-score. Hereby, the transformer-based models, especially RoBERTa showed promising performance for the multi-class classification task, although other transformer-based models *XLNet*, *BERT*, and *Ernie 2.0* also achieved a sufficient performance for this task. The additional scraped data from sustainability reports $(41, 4k)$ improved the performance of the implemented models, but the up-sampling approaches did not improve the overall performance to balance differing frequencies of SDGs in the dataset.

Thus, further research is necessary to apply data preprocessing approaches to balance the dataset to achieve higher classification performance. Moreover, the embedding weights could be obtained from the *n* last layer of the pre-trained models, similar to the related research on text classification problems proposed by (Mohtaj and Möller, 2022b). Furthermore, it is necessary to evaluate the models in additional languages, using data from sustainability reports of companies and international organizations that are also published in other languages.

Overall, the results show that the application of NLP models for the automatic SDG classification task is feasible and can be implemented for the automatic processing of text documents. These approaches can be used to sort text documents according to their relevance concerning different SDGs and aggregate information about the relevancy of different SDGs in various industries and organizations.

Acknowledgment

We would like to express our very great appreciation to Charlott Jakob for providing the dataset and also to Lucy Grey-Gardner for her contribution implementing the NLP models.

7. References

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf, 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019*.

Angin, Merih, Beyza Taşdemir, Cenk Arda Yılmaz, Gökcan Demiralp, Mert Atay, Pelin Angin, and Gökhan Dikmener, 2022. A roberta approach for automated processing of sustainability reports. *Sustainability*, 14(23):16139.

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*.

Guisiano, Jade Eva, Raja Chiky, and Jonathas de Mello, 2022a. SDG-Meter : a deep learning based tool for automatic text classification of the Sustainable Development Goals. In *ACIIDS :14th Asian Conference on Intelligent Information and Database Systems*.

Guisiano, Jade Eva, Raja Chiky, and Jonathas De Mello, 2022b. Sdg-meter: A deep learning based tool for automatic text classification of the sustainable development goals. In *ACIIDS: 14th Asian Conference on Intelligent Information and Database Systems*.

Hajikhani, Arash and Arho Suominen, 2022. Mapping the sustainable development goals (sdgs) in science, technology and innovation: application of machine learning in sdg-oriented artefact detection. *Scientometrics*:1–33.

Kleespies, Dierkes P.W., M.W., 2022. The importance of the sustainable development goals to students of environmental and sustainability studies—a global survey in 41 countries. *Humanit Soc Sci Commun 9, 218 (2022)*.

Le, Quoc and Tomas Mikolov, 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. PMLR.

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*.

Loshchilov, Ilya and Frank Hutter, 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Luo, Xiaoyu, 2021. Efficient english text classification using selected machine learning techniques. *Alexandria Engineering Journal*, 60(3):3401–3409.

Mhlanga, Ruth, Uwe Gneiting, and Namit Agarwal, 2018. Walking the talk: Assessing companies' progress from sdg rhetoric to action.

Mohtaj, Salar and Sebastian Möller, 2022a. The impact of pre-processing on the performance of automated fake news detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer.

Mohtaj, Salar and Sebastian Möller, 2022b. On the importance of word embedding in automated harmful information detection. In *International Conference on Text, Speech, and Dialogue*. Springer.

Nexar and Message Group, 2003. Online reports database. `https://www.online-report.com/report-type/sustainability-report/`. Accessed: 2022-05-30.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pradhan, Prajal, Luís Costa, Diego Rybski, Wolfgang Lucht, and Jürgen P. Kropp, 2017. A systematic study of sustainable development goal (sdg) interactions. *Earth's Future*, 5(11):1169–1179.

Pukelis, Lukas, Núria Bautista-Puig, Mykola Skrynik, and Vilius Stanciauskas, 2020. OSDG - open-source approach to classify text data by UN sustainable development goals (sdgsguisiano). *CoRR*, abs/2005.14569.

Reporting, International Corporate Environmental, 2002. The portal for sustainability reporting. `https://www.sustainability-reports.com/`. Accessed: 2022-05-30.

Sachs, Kroll C. Lafortune G. Fuller G. Woelm F., J., 2022. *Sustainable Development Report 2022*. Cambridge: Cambridge University Press.

Smith, Thomas Bryan, Raffaele Vacca, Luca Mantegazza, and Ilaria Capua, 2021. Natural language processing and network analysis provide novel insights on policy and scientific discourse around sustainable development goals. *Scientific reports*, 11(1):1–10.

Sun, Yu, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang, 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.

United Nations, 2015. The 17 goals — sustainable development. `https://sdgs.un.org/goals`. Accessed: 2022-05-30.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*.

Xu, Shuo, 2018. Bayesian naïve bayes classifiers to text classification. *Journal of Information Science*, 44(1):48–59.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le, 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.