Identifying document relevance to Sustainable Development Goals using NLG

M.G. Erechtchoukova and N. Safwat

School of Information Technology, York University, Toronto, Canada
Email: marina@yorku.ca

Abstract: Artificial Intelligence (AI) and, specifically, Natural Language Processing (NLP) techniques are considered as catalyzers of sustainable development of human society by providing information technology support for attainment of targets of Sustainable Development Goals (SDGs). The current study aims at investigating applicability of language generative models for identifying representation of SDGs in scientific publications indexed by Scopus database. The study is an initial step in developing an NLP-based framework for evaluation of attainment of SDGs based on documents written in human language. Given that SDGs are articulated in natural language in sentences of different length, comparison of their descriptions with summaries of text documents is expected to identify and quantify relevance of documents to each SDG and its targets in a more comprehensive way compared to the traditional keyword search. The study is based on abstractive summarization and follows the methodological framework presented in Figure 1.

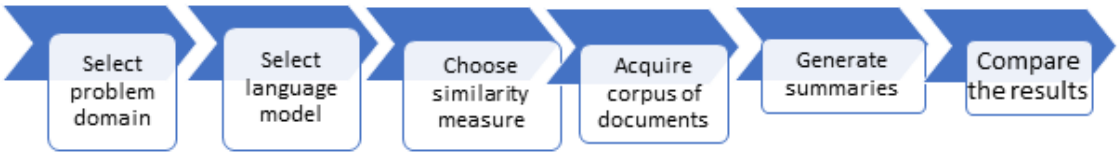

Figure 1. Framework for document analysis using text summarization

Neural language models developed based on Transformer architecture are available as open source software. Several model implementations were selected from the library of HuggingFace platform. The models were evaluated based on the required computational resources and their performance on a small corpus of documents. Two models, namely, BART and T5, were selected for the study due to their relatively low computational cost. These models are pre-trained on large corpora of general purpose text documents. Given multifaceted nature of the Sustainability concept that covers the most important issues of human society, the assumption was made that the models are suitable for text-generation NLP tasks in the selected problem domain. To evaluate text similarity standard NLP similarity measures are not sufficient for accurate extraction of semantics of the texts. Word2Vec similarity was chosen to evaluate relevance of the selected documents to the target texts.

To ensure credibility of the documents, 988 peer-reviewed scientific publications which appeared over the period from January 2022 to May 2023 were extracted from Scopus Elsevier database. Each SDG short description was downloaded from the United Nations web site. Analysis of authors' keywords confirmed that topics dominating in the corpus are relevant to the sustainability domain and cover all three pillars of the Sustainability concept: environment, economy, and society. Abstracts of extracted papers formed the corpus of investigated peer-reviewed publications. Each abstract was fed into two selected models for summarization. For each obtained summary, its similarity to each of the 17 SDGs was calculated based on Word2Vec score. The obtained similarity scores were further analysed using standard quantitative methods of aggregation. The results of the quantitative analysis led to conclusions on SDGs representation in recent scientific publications. The study uncovered that all 17 SDGs found their reflection in the recent scientific publications. Goal 17 is the most representative SDG, while Goal 16 was not in the focus of the recent studies.

The study resulted in an approach to automatic processing of a corpus of text documents aiming to identify relevance of documents to SDGs using quantitative analysis of semantic similarity scores. The approach can be applied to different problem domains and corpora of documents extracted from various sources provided that corresponding target texts are available.

Keywords: *Abstractive text summarization, semantic similarity, sustainable development goals*

1. INTRODUCTION

The sustainable development of human society has been declared as the top priority by the United Nations (UN). UN 2030 Agenda had been articulated in 17 interrelated Sustainable Development Goals (SDGs) covering various directions of societal transition towards a better future. However, successful implementation of these goals requires monitoring and evaluation based on large sets of quantitative and qualitative targets and their indicators. The indicators of attainment of the identified targets are normally published in various documents, e.g., corporate reports or governmental publications, which are unstructured textual files or are scattered across databases with inconsistent and sometimes unknown data organization. This makes the analysis of the available data extremely challenging.

Vinuesa et al. (2020) investigated a potential role of Artificial Intelligence (AI) in achieving sustainability. AI could contribute to sustainable development in various ways from improving information support and automating decision making to increasing efficiency of production and addressing global challenges. A consensus-based experts' opinion elicitation process was conducted whether AI could play a role of enabler or an inhibitor of attainment of specific targets. According to the study, 82% of participants confirmed AI's enabler role in attainment of all targets of SDGs. However, the concerns were raised on the lack of regulatory insights safeguarding transparency, safety, and ethical standards.

To fulfill UN's sustainability plan to reach sustainable society and environment within 2030, multi-disciplinary efforts are required to transform the development process. European Union Global Sustainability Reporting Directive expanding the requirements to submit reports on the sustainability compliance to multi-national companies operating in European Union increases significantly the volume of unstructured documents that must be analysed for sustainable decision making (EU, 2022). Natural Language Generation (NLG) could be one of the major powerful tools that aids in this transformation process and addresses concerns expressed by research communities. Application of NLP techniques for analysis of documents in the area of Sustainability is expected to uncover important information on the attainment of the SDG standards and to become an important tool for sustainability assessment. However, methodologies for NLP application to this domain are yet to be developed. In addition, the multi-disciplinary nature of this important domain makes the investigation of pertinence of language models tested on benchmark data sets an interesting problem.

The current study aims at investigating of applicability of Natural Language Generation (NLG) approach, namely abstractive summarization, for identifying representation of SDGs in scientific publications indexed by Scopus Elsevier database. The study is an initial step in developing an NLP-based framework for evaluation of attainment of SDGs based on documents written in human language.

2. BACKGROUND

Relevance of text documents to various SDGs was investigated using systemic approaches, bibliometric analysis, and NLP techniques. Allen et al. (2020) conducted systemic review of recent scientific publications and national practices to identify scientifically based approaches to sustainable operations. Although the systemic approach is a basis for all comprehensive studies, when it is applied to a problem manually, the number of information sources and documents used in the analysis are limited due to human ability to grasp the information. This explains the interest of the research community to automate processing of textual and unstructured data and extract important information quickly from the large collections of documents.

SDGs were articulated to overcome low-dimensional representation of sustainability aspects. Each SDG is multifaceted. Therefore, frameworks for sustainability assessment of undertakings require thorough analysis of SDGs interlinkage. The interdependencies of SDGs were investigated based on supervised machine learning algorithms. To apply supervised classification, keywords were identified in the descriptions of SDGs, and text documents were assigned classes according to similarity between the keywords from a goal and the document (Hajikhani and Suominen, 2018).

Network analysis was also acclaimed as a tool to investigate SDGs interrelations. Bellantuono et al. (2022) developed three conceptual networks on interdependencies of SDGs based on publications, communications on Twitter, and based on country achievements. Kiriu and Nozaki (2020) conducted text analysis based on pre-defined keyword hierarchy on the set of Corporate Responsibility Reports and proposed the score to evaluate an organization performance based on word embedding techniques. Keyword-based analysis was applied to social media data to map public opinion to SDGs (ADB, 2022). Smith et al. (2021) combined network analysis with NLP word embedding to investigate interdependencies of SDGs in UN reports. The study showed significant deviation in representing environmental area and revealed hidden interrelations between SDGs from different areas.

All these studies demonstrated usefulness of NLP approaches to processing corpora of documents related to sustainability area. However, the investigated techniques did not include approaches considering contextual interpretation of words.

3. METHODOLOGY

Natural Language Generation (NLG) is devoted to approaches to create human-like written text automatically. Text-to-text generation models process existing human written text and produce the new text with similar meaning. Text summarization, machine translation, text simplification, and automatic grammar correction are examples of the most popular text-to-text NLG tasks. Text summarization is aimed at producing a short text that represents the meaning of the original one correctly. Nowadays two approaches to the text summarization exist: extractive and abstractive. In extractive summarization, the most representative sentences from the original text are selected. Abstractive summarization generates new sentences and phrases to reflect the meaning of the text. Given that SDGs are articulated in natural language in sentences of different length, comparison of their descriptions with summaries of relevant text documents is expected to identify and quantify relevance of documents to each SDG and its targets in more comprehensive way compared to the traditional keyword search. The study is based on abstractive summarization and follows methodological framework presented in Figure 1.

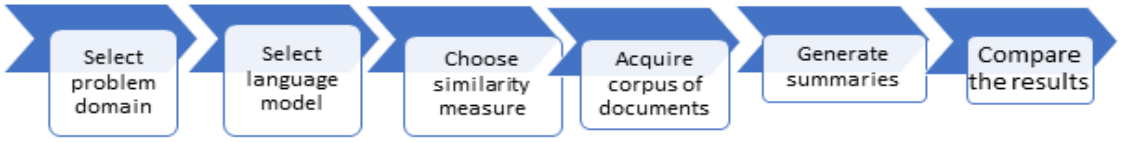

Figure 1. Framework for document analysis using text summarization

3.1. Language models

Language models depict occurrences and co-occurrences of words or sequences of words in texts using statistical or machine learning techniques. With the evolution of artificial neural networks (ANN) into deep leaning tools, language models that are based on deep neural networks dominated the NLP area. Vaswani et al. (2017) proposed to apply Transformer architecture of ANNs for natural language modeling. Deep learning models built based on this architecture accept a sequence of elements and transform it into another sequence. Due to the complexity of the models and high computational cost of their training on data from an investigated domain, the models are pre-trained on large corpora of documents to reproduce general language structure and word distributions.

There are several Transformer-based language models available as open-source software. Six models were initially considered in the study: BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2019), T5 (Liu et al., 2021), GPT-2 (Radford et al., 2019), and XLNET (Yang, 2019). Model implementations were selected from the library of HuggingFace platform. Firstly, the models were evaluated based on the required computational resources. All these models were pre-trained on large corpora of general purpose text documents. Given the multifaceted nature of the Sustainability concept that covers all important aspects of human society, the assumption was made that all these models are suitable for text-generation NLP tasks in the selected problem domain. GPT-2 and XLNET appeared to be more memory demanding than other models for application on a single GPU. The four other models were used in preliminary computations. According to the BLEU similarity measure BART and T5 models outperformed BERT and RoBERTa on the corpus of 1634 publications randomly selected from Scopus database. Therefore, these two models were selected for further investigation.

3.2. Similarity measures

NLG model performance is evaluated by comparing the model produced text with the known human written text also called target text. BLEU Score proposed by Papineni et al. (2002) is one of the commonly used metrics for NLG tasks. BLEU score measures Precision. With respect to NLP tasks, Precision refers to the portion of words in the generated text that are also present in the target text. It varies between 0 for absolutely dissimilar words and 1 for identical words. This score was used at the stage of model selection.

However, just word counting may not be sufficient for accurate extraction of semantics of the texts. The latter can be depicted using word embedding techniques. Word2Vec is one of the popular word embeddings which represents semantics and relationships between words. The embedding maps words into a numeric vector

space. Word similarity is measured by the cosine distance between corresponding vectors in this space (Mikolov et al., 2013). Similarity score calculated based on Word2Vec embedding, which also varies within the range from 0 to 1, was chosen to evaluate relevance of the selected documents to the target texts.

3.3. Building a corpus

The study requires to assemble a corpus of documents pertinent to sustainability domain. To ensure the credibility of the dataset, Scopus Elsevier database was chosen as the source of data because it indexes expertly curated scientific documents. 988 peer-reviewed scientific publications which appeared over the period from January 2022 to May 2023 were extracted from this database on February 13, 2023, using a single keyword 'Sustainable Development Goal'. A publication was selected if the keyword was found in its title, authors' list of keywords, or abstract. The preliminary bibliometric analysis was conducted on the metadata of extracted papers. Figure 2 presents top sources of publications contributing to the corpus the highest number of text documents. The chart was developed using Microsoft Power BI. Figure 3 shows Word cloud of publications' keywords identified by their authors. It was produced using wordcloud and matplotlib Python packages. Analysis of authors' keywords confirms that topics dominating in the corpus are relevant to the sustainability domain and cover all three pillars of the Sustainability concept: environment, economy, and society. Abstracts of extracted papers formed the corpus of investigated peer-reviewed publications.

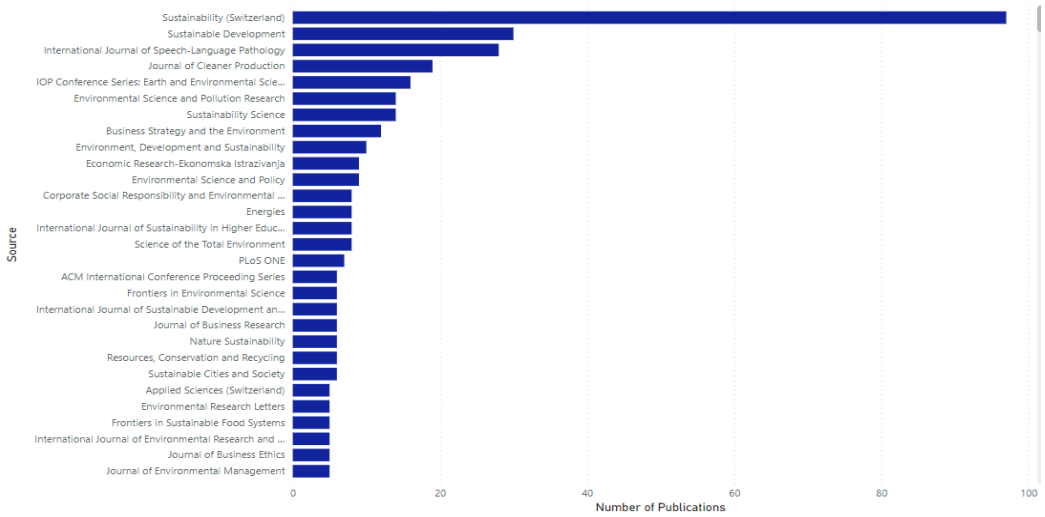

Figure 2. Distribution of documents in the corpus between publication sources

Figure 3. The most frequent publication keywords

4. RESULTS AND DISCUSSION

The corpus of abstracts was pre-processed according to the language model's requirements. Each abstract was considered as a short text, which was fed into two selected models for summarization. Each SDG short description was downloaded from the United Nations web site. These descriptions were also pre-processed using standard NLP techniques to make them suitable for semantic similarity calculations. For each obtained summary, its similarity to each of the 17 SDGs was calculated based on Word2Vec score. Aggregated results of computational experiments are presented in Table 1.

Table 1. Similarity of the corpus of documents to SDGs

Goal	BART			T5		
	Min	Average	Max	Min	Average	Max
Goal 1	0	0.516	0.675	0.115	0.504	0.663
Goal 2	0	0.52	0.787	0.089	0.49	0.771
Goal 3	0	0.517	0.682	0.065	0.509	0.812
Goal 4	0	0.509	0.782	0.053	0.488	0.738
Goal 5	0	0.387	0.7	0.038	0.383	0.737
Goal 6	0	0.44	0.744	0.006	0.43	0.799
Goal 7	0.007	0.527	0.852	0.048	0.495	0.732
Goal 8	0	0.569	0.774	0.11	0.534	0.796
Goal 9	0.033	0.53	0.834	0.093	0.483	0.767
Goal 10	0.013	0.461	0.657	0.096	0.46	0.658
Goal 11	0	0.477	0.774	0.071	0.438	0.668
Goal 12	0.027	0.475	0.817	0.059	0.456	0.704
Goal 13	0.032	0.517	0.747	0.108	0.508	0.707
Goal 14	0.067	0.412	0.706	0.054	0.396	0.701
Goal 15	0.047	0.428	0.704	0.149	0.406	0.657
Goal 16	0	0.44	0.685	0.121	0.411	0.619
Goal 17	0.052	0.572	0.849	0.117	0.52	0.795

Scores calculated for BART generated summaries showed a wider range of values compared to those generated by the T5 model. It seems that BART is less sensitive to the topics that are outside the scope of the text although might have some relevance to the text. The presence of topics marginally related to 9 sustainable goals was not recognized by BART in some papers, while T5 generated summaries of the same texts that allowed to identify relevance to these SDGs. At the same time, the minimal similarity scores are very low for all SDG, indicating that some publications present studies that are more focused on particular aspects of sustainability. These results from both models are consistent.

It can be observed that summaries generated by different models resulted in different similarity scores. Nevertheless, some patterns were identified. Analysis of maximal similarity of summaries generated by both models confirms that all 17 SDGs found their reflection in the recent scientific publications. For each SDG there are some documents with high similarity scores which means that this SDG was in the focus of a study. The consistency of both models in pointing to SDGs with high scores also supports this conclusion.

It is worth noting that the maximal similarity score calculated based on BART model was achieved for Goal 7 "Ensure access to affordable, reliable, sustainable and modern energy for all", while T5 resulted in maximal score on Goal 3 "Ensure healthy lives and promote well-being for all at all ages". The second high score was registered for BART's summaries on Goal 17 "Revitalize the global partnership for sustainable development". T5 results point to Goal 6 "Ensure access to water and sanitation for all". These suggest that extreme values of the score may not be very informative for identifying dominating topics and research gaps in the domain because they may correspond to a relatively few papers included in the corpus.

Using similarity scores between individual summaries and SDGs, goals that the most relevant and the least relevant to each document from the corpus were identified. The numbers of the most relevant papers and the

least relevant papers from the corpus may provide an insight on SDGs representation in the corpus. This distribution of papers between SDGs is presented in Table 2.

Table 2. The most and the least relevant papers

Goal	Papers with min similarity		Papers with max similarity	
	BART	T5	BART	T5
Goal 1	4	0	109	101
Goal 2	1	0	19	33
Goal 3	0	0	84	150
Goal 4	1	2	37	45
Goal 5	455	369	3	7
Goal 6	58	57	11	14
Goal 7	0	0	56	44
Goal 8	0	0	145	148
Goal 9	6	12	11	12
Goal 10	28	21	12	42
Goal 11	18	23	0	0
Goal 12	15	22	19	26
Goal 13	1	0	95	122
Goal 14	248	278	11	16
Goal 15	107	130	8	10
Goal 16	40	66	0	0
Goal 17	6	8	368	218

The results obtained based on both models are consistent. Minimal similarity according to BART and T5 was determined for the same SDGs. Goal 5 “Achieve gender equality and empower all women and girls” is the least represented in the corpus. Goal 14 “Conserve and sustainably use the oceans, seas and marine resources” is the second least represented. It follows Goal 15 “Sustainably manage forests, combat desertification, halt and reverse land degradation, halt biodiversity loss”. Goals 6 described previously and 16 “Promote just, peaceful and inclusive societies” are also included in the five least relevant goals, but they are ranked differently according to the summaries generated by the two models.

It is worth mentioning that the second high similarity score for T5 results was achieved on a paper relevant to Goal 6 while this goal appeared among the least represented ones. This means, that this goal was represented by papers that are more focused on specific aspects of sustainability, and the number of these papers in the corpus is relatively low.

Similar consensus was observed on the top five SDGs reflected in the corpus. Goal 17 is the most represented in the corpus. Goals 1 “End poverty in all its forms everywhere”, 3 “Ensure healthy lives and promote well-being for all at all ages”, 8 “Promote inclusive and sustainable economic growth, employment and decent work for all”, and 13 “Take urgent action to combat climate change and its impacts” lead the list. However, they are ranked differently. According to BART, the order of the top SDGs is Goal 8, Goal 1, Goal 13, and Goal 3, while T5 results can be ordered as Goal 3, Goal 8, Goal 13, and Goal 1.

Results from both models imply that goals 3, 7, and 8 were reflected in all extracted publications at least to some extent. Therefore, they may represent the current trend in research. There were no papers with maximal similarity score for Goals 11 and 16 found in the corpus by both models. This means that the corpus does not have a paper with the clear focus on this SDGs.

5. CONCLUSION

The study confirmed the effectiveness of NLG models for automated analysis of textual data in a multi-disciplinary domain. It resulted in an approach to automatic processing of a corpus of text documents aiming to identify relevance of documents to SDGs using quantitative analysis of semantic similarity scores. As opposed to multiclass classification, the approach is based on abstractive text summarization and application of Word2Vec word embedding for calculation of semantic similarity of each document to each SDGs. Quantitative analysis of the obtained scores enables the drawing of conclusions on the representation of SDGs in the corpus. These conclusions help to reveal the most popular topics in the area of sustainability, provide insights on the topic distributions among the publications and identify gaps in the selected research area that require attention of research community. At the same time, further improvement of quantitative analysis of sematic similarity scores can generate more interesting and detailed information. This expansion is a subject of further research.

The approach can aid in bibliometric analysis to complement its standard techniques or in selection of the most relevant papers for reading in the sustainability area and in the analysis of corporate documents. The approach can be applied to different problem domains and corpora of documents extracted from various sources provided that corresponding target texts are available.

ACKNOWLEDGEMENTS

The authors are grateful to anonymous reviewers for their thoughtful suggestions and helpful comments on the manuscript improvement. This study was partially funded by Academic Excellence Fund, York University, Canada.

REFERENCES

Allen, C., Metternicth, C., Weidmann, T., 2021. Priorities for science to support national implementation of the sustainable development goals: A review of progress and gaps. Sustainable Development, 29: 635–652.

Asian Development Bank (ADB), 2022. Mapping the public voice for development. Natural Language processing of social media text data. A Special Supplement of Key Indicators for Asia and the Pacific. http://dx.doi.org/10.22617/FLS220347-3

Bellantuono, L., Monaco A., Amoroso, N., Aquaro, V., Lombardi, A., Tangaro, S., Bellotti, R., 2022. Sustainable development goals: conceptualization, communication and achievement synergies in a complex network framework. Applied Network Science, 7:14. https://doi.org/10.1007/s41109-022-00455-1.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

European Union, 2022. Sustainable economy: Parliament adopts new reporting rules for multinationals. https://www.europarl.europa.eu/news/en/press-room/20221107IPR49611/sustainable-economy-parliament-adopts-new-reporting-rules-for-multinationals. Downloaded on 13.03.2023

Hajikhani, A., Suominen, A., 2018. The Interrelation of Sustainable Development Goals in Publications and Patents: A Machine Learning Approach. 1st Workshop on AI + Informetrics - AII 2021. https://ceur-ws.org/Vol-2871/paper15.pdf.

Kiriu, T., Nozaki, M., 2020. A text mining model to evaluate firms' ESG activities: An application for Japanese firms. Asia-Pacific Financial Markets, 27:621–632

Lewis, M., Liu, Y., Goyal, N., Ghazvini Nejad, M., Mohamed, A., Levy, O., Zettlemoyer, L., 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461.

Liu, F., Shakeri, S., Yu, H., & Li, J., 2021. EncT5: Fine-tuning T5 encoder for non-autoregressive tasks. https://arxiv.org/abs/2110.08426.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., & Chen, D. et al., 2019. RoBERTa: A Robustly optimized BERT pretraining approach. https://arxiv.org/abs/1907.11692

Mikolov, T., Chen, K., Corrado, G., & Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. DOI: 10.3115/1073083.1073135.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Smith, T.B., Vacca, R., Mantegazza, L., Capua, I., 2021. Natural language processing and network analysis provide novel insights on policy and scientific discourse around Sustainable Development Goals. Scientific Reports, 11:22427.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Vinuesa, R., Azizpour, H. & Leite, I., 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233 (2020). https://doi.org/10.1038/s41467-019-14108-y

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. Retrieved 12 June 2022, from https://arxiv.org/abs/1906.08237