The Journal of Open Source Software

SDGdetector: an R-based text mining tool for quantifying the efforts toward SDGs

Yingjie Li [1,2], **Meng Cai** [3,4], **Veronica F. Frans** [1], **Yuqian Zhang** [1], and **Jianguo Liu** [1]

1 Center for Systems Integration and Sustainability, Michigan State University, East Lansing, MI 48823, United States **2** Natural Capital Project, Stanford University, Stanford, CA, 94305, United States **3** School of Planning, Design and Construction, Michigan State University, East Lansing, MI, 48824, United States **4** Institute for Traffic Planning and Traffic Engineering, Technical University of Darmstadt, Darmstadt 64287, Germany

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

The global interest in moving towards a sustainable future has grown exponentially at all levels. The United Nations' Sustainable Development Goals (SDGs) adopted by world leaders in 2015 provided an integrated framework to track progress toward sustainability (UN, 2019). Textual data, such as public statements posted on websites, organization reports, and scientific publications, provide a rich source for evaluating the planned and ongoing efforts, as well as achievements towards sustainability. However, no computational tool exists to date that is able to accurately and efficiently identify SDG-related statements from these large amounts of text data. To fill this gap, we developed the **SDGdetector** package in R (R Core Team, 2021) to map textual data to specific goals and targets under the UN SDG framework for quantitative analysis. This is the first open-source, high-resolution, and high-accuracy analytic package that can identify which and how many SDG goals and targets are declared in any type of text-based data. This package thus enables a unique way to monitor individuals' and organizations' commitments and efforts towards advancing the 17 SDGs and 169 associated targets.

Statement of need

The Sustainable Development Goals (SDGs) agenda, adopted by all United Nations Member States in 2015, provides a shared blueprint for nations, cities, corporations, research institutions, and individuals to track and plan their contributions to social, economic, and environmental transformations (UN, 2019). Although considerable efforts and contributions have been made, half of the 231 indicators listed in the global indicator framework for SDGs lack either established methodology or available data on measuring and implementing the goals (https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/). As a complement to the widely used statistical data, the unstructured text provides a rich and important data source to narrow the data gap. For example, by identifying SDGs commitments and contributions in text from the legally binding corporate annual reports, one can evaluate which SDGs and to what extent the corporation is moving towards (Li et al., 2022). Manually reviewing and matching text to specific SDGs or targets can be extremely time-consuming and costly. In addition, though conventional manual coding may achieve high accuracy, it faces precision issues because of intercoder reliability challenges. This issue especially stands out when dealing with textual big data and with the objective to classify and map the massive data into tens and hundreds of topic categories (e.g., 169 SDG targets). To address these challenges, we assembled and refined six databases on SDG search queries (Bautista-Puig & Mauleón, 2019; Duran-Silva et

al., 2019; Jayabalasingham et al., 2019; Schubert, 2020; UN, 2019; Vanderfeesten et al., 2020; Wulff & Meier, 2021) and developed the **SDGdetector** package to automate the text analysis process in a text mining approach (Figure 1). Our integrated database is unique because it is by far the only available one that can be used to detect SDG-relevant statements at the SDG target level. In combination with this database, the text mining approach, an artificial intelligence (AI) technology, enables us to use natural language processing to transform the unstructured text in documents into normalized and structured data suitable for analysis. After repeated validation and calibration, this package has achieved high accuracy in detecting SDG-related statements in textual data ($>$ 75.5%, measured by the alignment between the R package results and experts' manually coded results). This package has been used in large-scale research projects in the field of corporate sustainability and urban science (Cai et al., 2022; Kassens-Noor, 2022; Li et al., 2022).

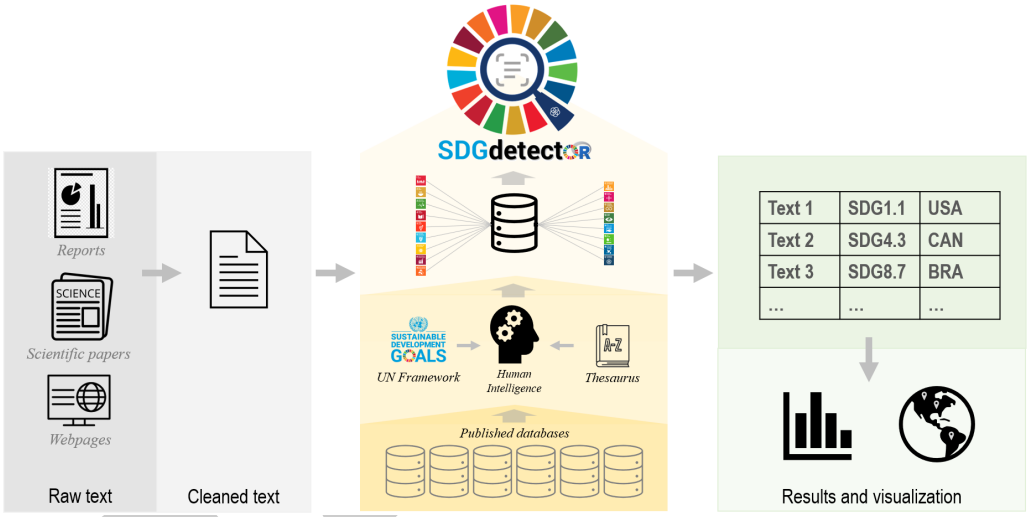

Figure 1. Flowchart for identifying SDG-related statements from textual data.

Functionality

SDGdetector is an R (R Core Team, 2021) package that provides functions for three main tasks:

(1) detecting whether a reported action aligns with any specific Goals (among the 17 SDGs) and Targets (among the 169 targets) under the Global indicator framework for Sustainable Development Goals (UN, 2019). The unit of text can be a clause, a sentence, or a paragraph. For the best accuracy, we suggest users split a large chunk of text into sentence or clause levels for analysis.

(2) estimating the priorities of sustainability contributions by counting how frequently a particular Goal or Target is mentioned in the text report.

(3) detecting which countries or regions are mentioned along with the SDG statements. For global studies, this function provides a means to show where the SDG efforts could be possibly implemented or have been planned.

The package is based on the tidyverse (Wickham et al., 2019) framework and is available on GitHub https://github.com/Yingjie4Science/SDGdetector. This lightweight package has great potential to be useful in many disciplines with objectives to identify which SDGs and to what extent an entity is putting effort into (Cai et al., 2022; Li et al., 2022). The associated lexicon database can be also used for developing similar applications in Python or other programming languages.

The Journal of Open Source Software

Acknowledgements

The authors acknowledge contributions from UN Global Sustainability Index Institute (UNGSII) Foundation during the genesis of this project. We thank Racheline Maltese for her input in developing the SDG search terms in the early stage. This work was funded by the National Science Foundation (grant numbers: DEB-1924111, OAC-2118329).

References

Bautista-Puig, N., & Mauleón, E. (2019). Unveiling the path towards sustainability: Is there a research interest on sustainable goals? *ISSI*, *II*, 2770–2771. ISBN: 978-88-338-1118-5

Cai, M., Li, Huang, H., Decaminada, T., & Kassens-noor, E. (2022). *Sustainable development goals in smart city implementation* [In revision].

Duran-Silva, N., Fuster, E., Massucci, F. A., & Quinquillà, A. (2019). *A controlled vocabulary defining the semantic perimeter of Sustainable Development Goals*. Zenodo. https://doi.org/10.5281/zenodo.3567769

Jayabalasingham, B., Boverhof, R., Agnew, K., & Klein, L. (2019). *Identifying research supporting the United Nations Sustainable Development Goals*. *1*. https://doi.org/10.17632/87txkw7khs.1

Kassens-Noor, E. (2022). *Urban scAInce: Explaining why and how cities transform through artificial intelligence* [In revision].

Li, Y., Frans, V., Zhang, Y., Cai, M., & Chen, R. (2022). *Global Business Giants' Commitment to Sustainable Development Goals* [In revision].

R Core Team. (2021). *R: A language and environment for statistical computing*. https://www.R-project.org/

Schubert, G. (2020). *Scientific publications on sustainable development*. Stockholm University Library.

UN. (2019). *Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development*. https://unstats.un.org/sdgs/indicators/indicators-list/

Vanderfeesten, M., Otten, R., & Spielberg, E. (2020). *Search Queries for "Mapping Research Output to the Sustainable Development Goals (SDGs)" v5.0*. Zenodo. https://doi.org/10.5281/zenodo.3817445

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wulff, D. U., & Meier, D. S. (2021). *text2sdg: Detecting UN Sustainable Development Goals in Text*. Zenodo. https://doi.org/10.5281/zenodo.5553980