The Journal of Open Source Software

SDGdetector: an R-based text mining tool for quantifying efforts toward Sustainable Development Goals

Yingjie Li [1,2,3], **Veronica F. Frans** [1,4,5], **Yongze Song** [6], **Meng Cai** [7,8], **Yuqian Zhang** [1,2], **and Jianguo Liu** [1,2]

1 Center for Systems Integration and Sustainability, Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI 48823, United States **2** Environmental Science and Policy Program, Michigan State University, East Lansing, MI 48823, United States **3** Natural Capital Project, Woods Institute for the Environment, Stanford University, Stanford, CA, 94305, United States **4** Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, MI 48824, United States **5** W.K. Kellogg Biological Station, Michigan State University, Hickory Corners, MI 49060, United States **6** School of Design and the Built Environment, Curtin University, Perth, WA, 6102, Australia **7** School of Planning, Design and Construction, Michigan State University, East Lansing, MI, 48824, United States **8** Department of Civil and Environmental Engineering, Technical University of Darmstadt, Darmstadt 64287, Germany

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

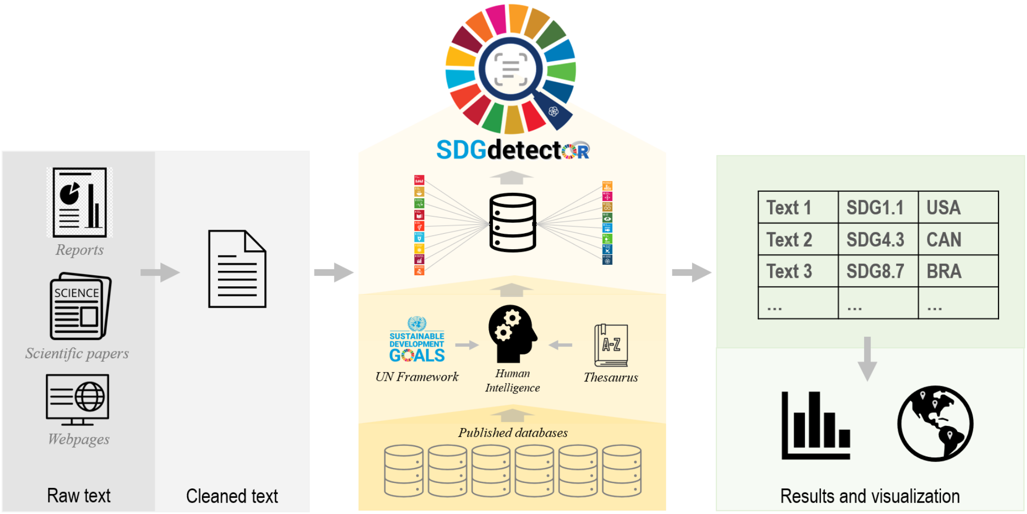
Submitted: 01 January 1970
Published: unpublished

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, provide an integrated framework to track progress toward sustainability (UN, 2019). Textual data, such as public statements posted on websites, organization reports, and scientific publications, is a rich source for evaluating the planned and ongoing efforts, as well as achievements towards sustainability. However, no computational tool exists to date that can 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 analytical package that can identify which and how many SDG goals and targets are declared in any type of text-based data frame or corpus. 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 to use existing statistical data for SDG assessments, half of the 231 indicators listed in the global indicator framework for SDGs lack either established methodologies or available data for measuring and implementing the goals (https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification). As a complement to the commonly used statistical data, textual data (e.g., websites, organization or government reports, and scientific publications) are rarely considered but show great potential for becoming a rich and important data source to narrow this existing SDG data gap (Cai, 2021; Chang et al., 2021). For example, by identifying SDG commitments and contributions in text from legally-binding corporate annual reports, one can evaluate which SDGs are being mentioned (directly or indirectly) and to what extent corporations are moving towards them. Or, published research papers could also be evaluated to link research institutions' commitments to SDG progress. Manually reviewing and matching text corpora 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 is especially an issue when attempting to objectively classify and map massive data into tens and hundreds of topic categories (e.g., the 169 SDG targets). To address these challenges, we developed the ***SDGdetector*** package, which automates the text analysis process via a text mining approach (Figure 1).



Text 1	SDG1.1	USA
Text 2	SDG4.3	CAN
Text 3	SDG8.7	BRA
...	...	...

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

The SDGdetector package was developed by (1) compiling six existing 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); (2) reviewing all SDG targets and indicators (UN, 2019) to manually refine and update the search terms to create query dictionaries at the levels of the 17 SDGs and the 169 SDG targets (which correspond to the 231 SDG indicators); (3) manually assessing and improving the accuracy of these queries using thousands of randomly-selected statements from real-world corporate annual reports across multiple iterations; and (4) turning these queries into a lexical database for text mining across large bodies of text and tabulating the matched SDGs and SDG targets.

SDGdetector is a unique tool because it is by far the only one available that is equipped with a database for detecting SDG-relevant statements at the target level. We are aware of another useful R package (*text2sdg*)(Wulff & Meier, 2021), which mostly uses single words as search terms and was designed to only map text to SDGs at the goal level (coarser resolution). Our search queries in the comprehensive database further considered sentence structure to reduce noise hits, and can capture hits at both goal and target levels. 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 within documents into normalized and structured data suitable for analysis and visualization. After repeated validation and calibration, this package has achieved high accuracy in detecting SDG-related statements within textual data ($>$ 75.5%, measured by the alignment between the R package results and four experts' manually-coded results; see the "Accuracy Evaluation" section on GitHub for more information). Complete data and code necessary for reproducing this accuracy evaluation can be found on our GitHub repository under the `./docs/accuracy_evaluation/` folder. Other similar tools, such as the *text2sdg*, however, did not report any accuracy evaluations.

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 them. This package can be used in large-scale research projects in the field of corporate sustainability and urban science. It can also be used in systematic reviews and syntheses of published literature and patents. The associated lexical database embedded within this R package can be also used for developing similar applications in Python or other programming languages.

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).

(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.

Usage

(1) Data preparation. Textual data can come from a variety of sources, such as PDF files, HTML webpages, TXT, or Microsoft Word documents. 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. Users can use our function *pdf2text()* or self-defined functions to extract textual data from PDF files, clean the text, split the text into sentences, and format the data in a dataframe.

(2) Detect SDG goals and targets. The input can be a single sentence, or a dataframe that contains many rows of sentences. If the input is a dataframe, users should designate which column to be used for SDG detection.

```r
# load package
library(SDGdetector)

# a string as the input
text <- 'our goal is to mitigate climate change, end poverty, and reduce
  inequality globally'
SDGdetector(x = text)

# a dataframe as the input
df <- data.frame(col = c(
  'our goal is to end poverty globally',
  'this product contributes to slowing down climate change'))
SDGdetector(x = df, col = col)
```

In addition to the lexical database included in the ***SDGdetector*** package, users can also add customized search queries to the lexical database.

```r
# A list of terms used to determine whether a sentence relates to SDG efforts
terms_new <- c("improve", "farmer", "income")
```

```
# Use *AND* operator to combine the terms and generate a customized search query
# (or called a matching pattern);
# then add the query to the existing lexical database
add_sdg_pattern(sdg_id = 'SDG1_2', x = terms_new, operator = 'AND')
```

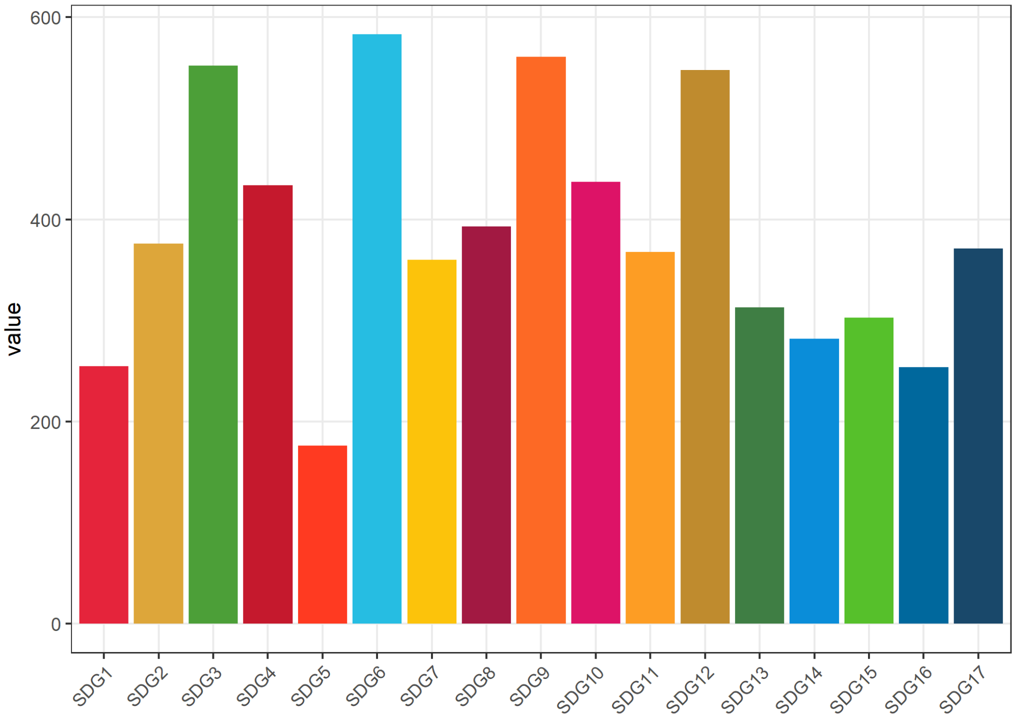
(3) Detect countries or regions. To understand where the SDG efforts are implemented or planned, users can use the function *detect_region()*. The result will return a list of country names in the ISO 3166-1 alpha-3 – three-letter country codes format.

```
text = 'China and USA devoted the largest efforts on solar energy'
detect_region(x = text)
```

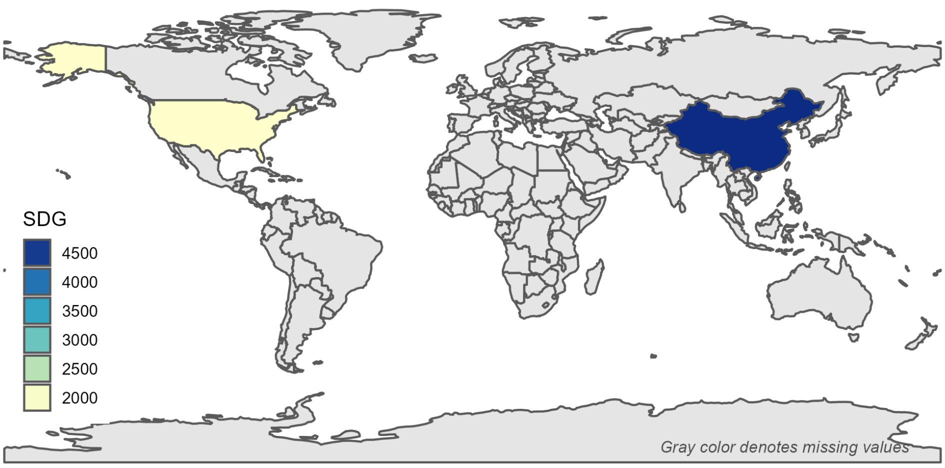
(4) Results and visualization. Users can summarize the detected SDG goals and targets by report (or organization), by Goal, and by region. The package provides methods for visualizing the SDG frequency, via its family of *plot_sdg_** functions. For instance,

```
data("sdgstat")
df <- sdgstat

# plot SDG on a bar plot
plot_sdg_bar(data = df, sdg = SDG, value = Value)
```



```
# plot SDG by country on a map
plot_sdg_map(data = df, sdg = SDG, value = Value, country = Country, by_sdg = F)
```



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 during the early stages of this project. This work was funded by the National Science Foundation (grant numbers: DEB-1924111, OAC-2118329). VFF was supported by the National Science Foundation Graduate Research Fellowship Program (Fellow ID: 2018253044) and the Michigan State University Enrichment Fellowship.

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. (2021). Natural language processing for urban research: A systematic review. *Heliyon*, *7*(3), e06322. https://doi.org/10.1016/j.heliyon.2021.e06322

Chang, T., DeJonckheere, M., Vydiswaran, V. G. V., Li, J., Buis, L. R., & Guetterman, T. C. (2021). Accelerating Mixed Methods Research With Natural Language Processing of Big Text Data. *Journal of Mixed Methods Research*, *15*(3), 398–412. https://doi.org/10.1177/15586898211021196

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

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. https://www.su.se/polopoly_fs/1.530251.1607009534!/menu/standard/file/sdg-publikationer-2010-2019_gabor_rev3.pdf

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