Tracing Policy Implementations – A Data Science Approach

Date: 23.11.2021
Authors: Jonathan Doorn, Emily Robitschek, Raphael Sgier, Paul Türtscher
In collaboration with: *Deutsche Gesellschaft für internationale Zusammenarbeit (GIZ)*
Developed in the context of Hack4Good, 4th Edition

Executive Summary / Abstract

In global politics, nations constantly publish commitments to international agreements, for instance to implement policies to mitigate the impacts of climate change. However, tracking and tracing of the actual policy implementations by these countries is currently time consuming and not systematically done, hampering accountability and sustainable progress. Therefore, GIZ requires a tool to quickly and automatically screen large amounts of documents (legal texts, speeches, tweets, ...) for finding and contextualizing actual implementations of a nation's commitments and policy priorities.

Expected impact

The approach we developed is simple and flexible. It consists of tools that are robust, powerful, and easy to use. The developed framework **can facilitate a more rapid, automated as well as customizable and comprehensive tracking of policy across nations and potentially over time.** Hereby, **the cumbersome task of screening manually dozens of documents can be streamlined, regardless of the traced policy** as the definition of such is user dependent.

Approach

The developed toolbox has two required inputs, a **folder/s of text documents** to be screened and the corresponding **keywords** of a) **nationally determine contribution (NDC) policy terms** which should be found in the given documents, and b) **keywords grouped by topics that are relevant** to the policy question of interest to help track coherence and implementation of the policy priorities (e.g. words related to the sustainable development goals or "SDGs"). To automate the process of finding both the policy goals (NDCs in this case) and context in which they are discussed and implemented (primarily aided by the topic keywords), our method provides a combination of search tools, including a deep analysis tool and a faster simple keyword matching method.

Details Deep Analysis
The deep analysis will embed a list of keywords through a neural network, creating vectors to represent each keyword. A passage of text will be split up into N-Grams, which are all possible consecutive N-length strings present in a document. These N-Grams are then

embedded. The vectors of the N-Grams are correlated with the keyword vectors through the cosine similarity method to calculate the scores. N-Grams which surpass a certain score threshold will be tagged as matches to the subject area of the keywords. This method is useful in broadening the given set of keywords, and is particularly relevant and helpful when the search priority is not to miss relevant terms, as in the case of finding policy goal related terms. Additionally, a similar analysis method can be used to compare the topics of passages.

Details Simple Keyword Matching

This method relies on spacy phrase matching and re searching for more exact keywords in the documents, and is useful when search precision is more important than breadth. The NDC keywords can be found using phrase matching or with a simple comparison of the terms and tokens from the document. The intuitive visualizations of the results allow the analyst to quickly understand which documents of many might contain the most relevant information to answer their question, and prioritize sections within documents to explore how countries are implementing policies and if they are doing so in a way that is aligned with the sustainable development goals. The analyst can even see what kinds of keywords they might be missing in the context of high-scoring passages (see example in Figure 1) and immediately update their search terms to improve the method.

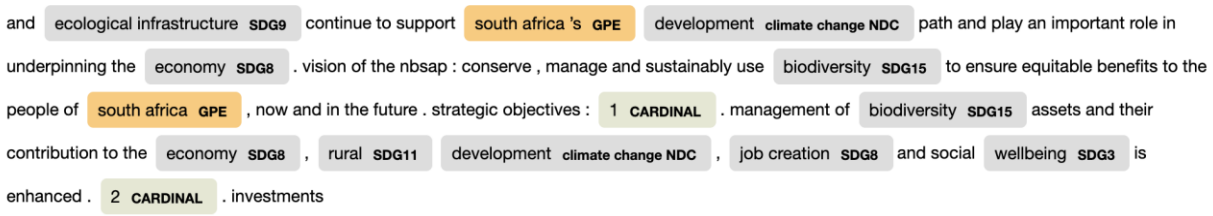

Figure 1: A rendered section of a policy document from South Africa that scored highly for several SDG categories

Finally, regardless of search method, the keyword counts and categories from each document can be scored in terms of relevance and coherence to have a quantitative measure for finding relevant documents and sections.

Difficulties, Limitations & Risks

The main issue consists of missing or unusable data. Depending on the targeted nation, not enough official policy documents are available, or the documents cannot be transformed into usable txt-format. Furthermore, dependent on the nation, more or less documents in English language are available. Our framework is currently only tested on English texts and an extension to include the possibility to analyze documents in other languages needs to be implemented and tested. Furthermore, the current status of the framework only includes a small set of keywords, which depends solely on the user. An extension and automation of the keyword set should be implemented in future versions. Moreover, the current version of the framework is missing a user interface for easy utilization from people not familiar with the code. As a general remark, we would like to stress that nevertheless tracing policy implementations in texts is very important, it does not replace the actual tracing on the ground of the corresponding nations.

Results & Deliverables

Several methods of finding keywords in the documents are provided, each with their own strengths and weaknesses, and can be systematically applied across a whole range of documents (Figure 1) as well as to passages within them (Figure 2).

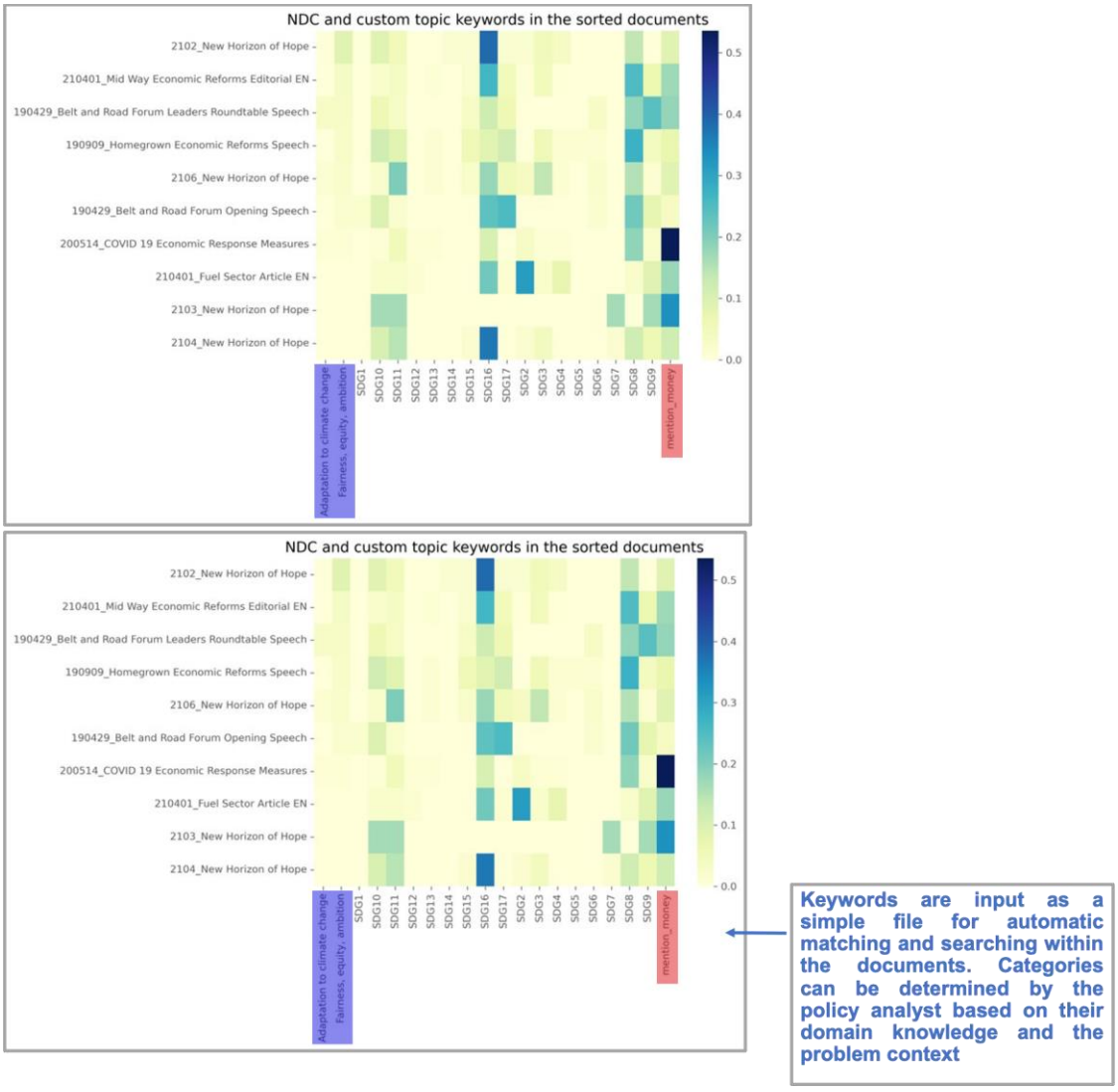

Figure 2: Heatmaps of several policy documents (rows) from South Africa (top) and Ethiopia (bottom) with the NDC terms (in blue) found through deep analysis and the SDG terms (columns) including a custom topic category (in red) extracted with the fast keyword search method. Darker colors represent higher scores for that category.

Our deliverable includes analysis libraries and examples to process pools of text documents. Different documents are analyzed for the presence of keywords that are associated with a policy framework given by the user. The results are then compiled into visualizations and scores to indicate how important documents in a document pool and sections of a document are to gain insight on framework-related impact.

Recommendations & Conclusion

Advise on the actual results implementation across the NGO (e.g. who is required to be involved in the implementation, capacity building required, how often certain parameters should be reviewed) as well as future possible developments (e.g. different algorithmic approaches, different data sets/sources, different technologies etc). Target 5 lines

One can easily improve the interpretability of and extend the existing functionality of the architecture to other policy frameworks by broadening the pool of search documents and adding keyword files specific to theses frameworks. The choice of keywords and topics is easily the most important factor (in addition to document quality) for the approach laid out here. Ideally, these keywords should carefully (and comprehensively) chosen to reflect what the policy analyst is looking for. The coherence and relevance scores should be updated with a better normalization method, weighting across topic categories, and compared across a known set of documents to validate their meaning. Additional to the developed architecture, we would recommend to try an unsupervised method to complement the highly supervised one laid out here, and perhaps to utilize an intelligent search framework such as Jina or Haystack to implement a question-answer search engine. Finally, the need to create a user interface is apparent. It would allow analysts to manage document pools and keyword files and to look into the results of their analysis more easily.