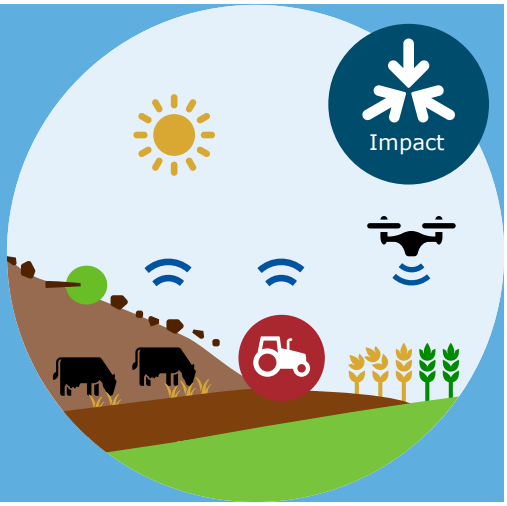
Machine learning to identify good practices with Nature-based Solution

Scaling nature-based solutions for adaptation

Emerging DS/AI methods



WAGENINGEN
UNIVERSITY & RESEARCH

Data Driven Discoveries in a changing climate (D3C2)

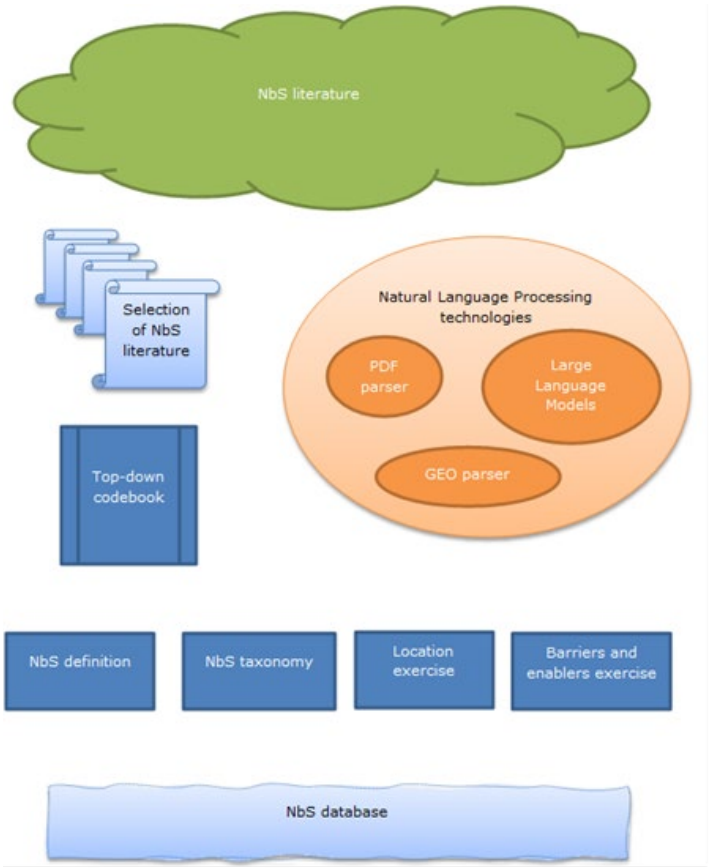
Objective: This study aimed to support scaling of Nature-based Solution (NbS) for adaptation by using machine learning tools. We explored a combination of technologies from the realm of Natural Language Processing (NLP), like a PDF parser, transformers, ClimateBert, and named entity recognition (NER). By doing so, we aimed to extract and classify NbS as well as their associated barriers and enablers for implementation from text, and store this in a database.

Activities

We composed a corpus of unstructured content representative for the topic of NbS. A first set came from the literature about dedicated to NbS in the IPCC 6th assessment report. This set was not helpful in creating a good definition of Nature-based Solutions. One of the possible causes might be that these publications are rather aimed at underpinning the topic discussion in the IPCC assessment report rather than focussing on NbS as such. We subsequently collected around 20 papers to experiment with different methods and approaches. This new set, selected through relatively simple queries performed on Scopus, appeared to be better applicable for our research.

As a next step in the research, we explored the definition of the domain and its semantics using a top-down as well as a bottom-up approach. In the top-down approach we asked domain experts to structure the relevant terminology around NbS in a taxonomy. This was documented in what we have called a 'Code Book for Nature-based Solutions'. For the bottom-up approach, a keyword extraction algorithm was used to identify the most important keywords for the collection of documents in the literature.

We have investigated various AI technologies and their ability to retrieve the desired information from the



---

corpus. Due to the limited size of this project and our struggles with the fuzzy nature of semantics, it was not feasible to collect large scale data for training and validation of the AI solutions investigated. Nor was any of such data already available prior to the project. The lack of training data also meant that a traditional machine learning approach, where a dedicated algorithm is trained for a particular task, could not be investigated. Various alternatives were explored instead.

A text mining framework was used to obtain access to the contents of the documents in the corpus. The general task of identifying the desired knowledge on NbS and its barriers and enablers was broken down into multiple subtasks to be individually performed by AI:

- identifying the NbS itself, including taxonomic classification;
- resolving the geographical location of the NbS;
- associating barriers and enablers to the NbS.

Achievement

The codebook, and particularly specific fragments describing targeted categories, has been used in initial experiments to find and analyse scientific publications. During these explorations it became apparent that the terminology surrounding NbS is rather unformalised and that the boundaries between taxa can be unclear. This is challenging for experts working in NbS. It is even more challenging for AI, because AI requires a well-defined task to begin with, and requires much more contextual information and advanced mechanisms for interpreting these inputs to perform that task.

As an alternative to the above approach, we explored Google's Bard generative AI to propose a taxonomy for NbS, thereby effectively leveraging contextual information on NbS in Bard's training data.
The result shows resemblance to the codebook manually conceived but appears to follow more strictly defined taxa.

About identifying the NbS itself, and taxonomic classification: inspired by the taxonomy provided by Google's Bard, we used additional prompts to Bard to obtain descriptions and further hierarchy for a part of the taxa. We obtained a taxonomic classifier able to predict the most likely taxonomic labels for each document.

Understanding the location of the NbS was a challenge. Consequently, multiple attempts have been undertaken: Among the implemented models, ChatGPT consistently yields the most accurate results, demonstrating a superior understanding of both questions and topics. This model excels in comprehending and responding effectively to queries, particularly when multiple countries are involved or when the solution is approached from a broader perspective, encompassing entire countries or continents. Notably, ChatGPT identifies and associates correct countries and, where applicable, even cities. However, ChatGPT also has disadvantages, notably hallucination (generating made-up results that resemble real results) and problems with data security and copyrights.

As a first crude attempt a rule-based tagger was developed in the Spacy framework to recognise terms signalling barriers and enablers being mentioned. On top of this, a simple algorithm was written to detect within sentence co-occurrence between any of these signalling terms and the previously identified NbS. This results in an excel of which the content could be further analysed. Generative AI was also applied which provides language results that also need further analysis.

Outlook

If terms and conditions are insufficiently clear about data security, only content from papers already in the public domain should be used. Alternatively, models with similar capabilities running on premise instead of remotely can help to overcome this issue. Among the options to be explored are Alpaca/Llama, GPT4all, Ollama framework.

Generative AI often does not provide structured output, but natural language instead. Additional steps are required to translate unstructured output into structured forms from which a database on NbS can be populated. Fine-tuning a large language model to perform the task of delivering structured data directly might be a viable alternative, depending on the difficulties of obtaining appropriate training data.

Quality assessment of the methods reported in this document has only been done through human judgement of a limited number of examples. Proper validation, including appropriate ground-truth data labelled prior to AI execution is recommended before widely adopting any of the methods. Validation is even more important when relying on generative AI methods known to exhibit forms of bias and hallucination.

Deliverables

- A working document of which a final draft is sent together with this summary report.

Lessons learned

- The language around NbS is still too fuzzy for AI to properly analyse it.
- ChatGPT is the most promising tool but has limitations regarding data security, reliability and traceability of results.
- Colleagues working on AI in different science groups know how to find each other. Non-AI colleagues know who can help them with these methods.
- The taxonomies for NbS are interesting.

Contact

Judith Klostermann
Researcher policy processes
judith.klostermann@wur.nl

Wageningen University & Research
P.O. Box 47
6700 AB Wageningen
The netherlands
T +31 317 48 07 00
www.wur.eu