Domain-Knowledge-Based Classification of Biodiversity Conservation Actions Based on Project Reports*

Kanyao Han, Rezvaneh Rezapour, Katia S. Nakamura, Dikshya Devkota, Daniel C. Miller, and Jana Diesner

University of Illinois at Urbana-Champaign

Extracting and classifying information relevant to a certain application domain are key tasks for evaluating actions and outcomes of projects. Oftentimes, these actions and outcomes are explicitly or implicitly mentioned in professionally-generated reports and publications. Retrieving such information from texts can be solved via inductive (bottom-up) and deductive (top-down) approaches, where human coders closely read documents, and annotate them based on indicators they find in the data (inductive) or instances of theory such as predefined categories (deductive) (Rezapour et al., 2020). Even though often considered as gold standard, manually extracting and classifying information can be time-consuming, cumbersome, and inefficient, especially for tasks that require domain expertise or contextual in-depth reading. To address this limitation, researchers have developed and applied computational solutions, e.g., machine learning and rule-based methods, to speed up the process of locating and categorizing instances of actions, outcomes, impact, and many other categories in text data (Töpel et al., 2017). Current computational solutions for this purpose require large amounts of labeled data to be able to exploit the contexts and underlying structure of the text data, which is necessary to produce reliable results. However, in real-world settings, such as small businesses and municipal administrations, sufficient amounts of annotated data might not be available, and generating them is constrained by a lack of expert human coders and resources to have them do these annotation tasks. Our study addresses this issue by proposing a semi-automated solution that brings together domain expertise and computational modeling to classify project reports based on small annotated data-sets.

Our data-set consists of project reports of funded work in the area of international biodiversity conservation. While funders have invested billions to preserve natural resources, especially in economically poor but biodiversity-rich countries, the long-term effectiveness of these investments is not always clear (Rana and Miller, 2019). To trace the impact of these projects, it is important to first understand what conservation actions have been undertaken. For identify such actions, we used the International Union for Conservation of Nature's (IUCN) categories [1] as listed below:

> (1) Land/Water Management, (2) Species Management, (3) Awareness Raising, (4) Law Enforcement and Prosecution, Livelihood, (5) Economic and Moral Incentives, (6) Conservation Designation and Planning, (7) Legal and Policy Frameworks, (8) Research and Monitoring, (9) Education and Training, (10) Institutional Development

*This work was supported by the John D. and Catherine T. MacArthur Foundation. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the MacArthur Foundation.

[1] Actions Classification (Version 2.0, Version 1.0):
https://cmp-openstandards.org/library-item/threats-and-actions-taxonomies/

	1	2	3	4	5	6	7	8	9	10
IK + DT	91.4	**95.7**	**78.8**	100	81.7	73.4	**90.2**	**91.5**	94.3	86
IK + EK + DT	**95.6**	**95.7**	77.4	100	**82.8**	**80.3**	**90.2**	**91.5**	**95.8**	**87.4**

Table 1: Classification accuracy of IUCN Categories 1-10 based on two models (values in percent, excluding Category 4 due to data sparsity)
IK = Initial Keywords; EK = Extended Keywords; Decision Tree

To computationally classify project reports, we trained models to assign one or multiple IUCN categories to new and unseen reports based on only 71 annotated reports. Our framework consists of five steps: (1) creating/obtaining initial keyword lists, (2) extending the initial lists via word embeddings and input from domain experts, (3) computing frequencies of keywords and their co-occurrence in context as features, (4) building decision trees (DT) for categorization, and (5) evaluating prediction performance. To obtain the initial keyword lists (1), we asked the domain experts in our team to create a list of keywords for each of the listed IUCN categories. To refine and expand these lists (2), we used a Word2Vec embedding model, which returned semantically related words for each keyword in our initial list. Instead of directly using these related words, the domain experts carefully selected relevant ones based on their domain knowledge. Since the labeled data consisted of only 71 reports, we leveraged the frequencies of the refined keywords as a feature (3), and trained a simple decision tree model to assign outcomes to each report (4). Table 1 shows the classification accuracy using k-fold cross-validation (k = 4) (5). The results suggest that our framework, which combines information provided by domain experts with automated methods for seed list expansion, word embedding, expert input, and supervised learning, can achieve 77.4% - 95.8% of classification accuracy across 10 IUCN categories. Our method was not able to classify category 4 due to data sparsity.

The results suggest that our proposed domain-knowledge-based approach may help to speed up categorization tasks, which frees up time from domain experts for other tasks. While the manual labeling of these reports took about one hour per document, our human-in-the-loop approach can assist experts in labeling documents in a more timely manner. Overall, our approach shows how combining human expertise with natural language processing methods can help to quickly and fairly accurately label documents for project actions.

In our future work, we plan to use the detected conservation actions to analyze the evolution of outcomes and impact. The findings from this work can provide a more detailed and comprehensive understanding of the objectives and impact of funded research.

References

Rana, P. and Miller, D. C. (2019). Machine learning to analyze the social-ecological impacts of natural resource policy: insights from community forest management in the indian himalaya. *Environmental Research Letters*, 14(2):024008.

Rezapour, R., Bopp, J., Fiedler, N., Steffen, D., Witt, A., and Diesner, J. (2020). Beyond citations: Corpus-based methods for detecting the impact of research outcomes on society. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6777–6785.

Töpel, M., Zizka, A., Calió, M. F., Scharn, R., Silvestro, D., and Antonelli, A. (2017). Speciesgeocoder: fast categorization of species occurrences for analyses of biodiversity, biogeography, ecology, and evolution. *Systematic Biology*, 66(2):145–151.