Received: 1 September 2021 | Revised: 8 February 2022 | Accepted: 27 August 2022

DOI: 10.1002/asi.24714

RESEARCH ARTICLE

JASIST WILEY

An expert-in-the-loop method for domain-specific document categorization based on small training data

Kanyao Han[1] | **Rezvaneh Rezapour**[2] | **Katia Nakamura**[3,4] | **Dikshya Devkota**[3] | **Daniel C. Miller**[3,4] | **Jana Diesner**[1]

[1]School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

[2]College of Computing and Informatics, Drexel University, Philadelphia, Pennsylvania, USA

[3]Department of Natural Resources and Environmental Science, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

[4]Keough School of Global Affairs, University of Notre Dame, South Bend, Indiana, USA

Correspondence
Kanyao Han, School of Information Sciences, University of Illinois at Urbana-Champaign, 501 E. Daniel St., Champaign, IL 61820, USA.
Email: kanyaoh2@illinois.edu

Funding information
John D. and Catherine T. MacArthur Foundation

Abstract

Automated text categorization methods are of broad relevance for domain experts since they free researchers and practitioners from manual labeling, save their resources (e.g., time, labor), and enrich the data with information helpful to study substantive questions. Despite a variety of newly developed categorization methods that require substantial amounts of annotated data, little is known about how to build models when (a) labeling texts with categories requires substantial domain expertise and/or in-depth reading, (b) only a few annotated documents are available for model training, and (c) no relevant computational resources, such as pretrained models, are available. In a collaboration with environmental scientists who study the socio-ecological impact of funded biodiversity conservation projects, we develop a method that integrates deep domain expertise with computational models to automatically categorize project reports based on a small sample of 93 annotated documents. Our results suggest that domain expertise can improve automated categorization and that the magnitude of these improvements is influenced by the experts' understanding of categories and their confidence in their annotation, as well as data sparsity and additional category characteristics such as the portion of exclusive keywords that can identify a category.

1 | INTRODUCTION

Categorizing documents according to domain-specific ontologies and taxonomies to enrich information that domain experts and practitioners can analyze is a common task for working with text data (Grimmer & Stewart, 2013). In recent decades, automated categorization models have been increasingly utilized in a variety of domains such as environmental science (Borg et al., 2021) and social science (Zhang & Pan, 2019a, 2019b) to minimize the amount of labor and time needed for labeling large-scale text data. We have access to (highly) accurate models for tasks such as positive and negative sentiment prediction (Yadav & Vishwakarma, 2020; Zhang & Pan, 2019a), latent topic discovery through topic modeling (Roberts et al., 2013), and classification of online posts with regard to collective action (Zhang & Pan, 2019a), Covid-19 (Lu et al., 2021), and hate speech (Schmidt & Wiegand, 2017), among many others. However, few automated models or methods are available when the categorization of documents requires in-depth domain expertise of the data or task. In view of this, recent literature has explored whether

and how we can leverage automated categorization methods to extract domain-specific information from text corpora (Monroe, 2019; Oliver, 2019).

One common approach to address this issue is to utilize "expert-in-the-loop" models, in which domain expertise is incorporated into the structure of categorization models. More specifically, such models are either developed using features (e.g., keywords) provided by domain experts (Gerla et al., 2019; Huang & Lu, 2010; Rezapour et al., 2019; Ryazanov et al., 2021; Wang, 2014), or calibrated and enhanced based on experts' feedback and evaluation of categorization results (Guo et al., 2016; Holzinger et al., 2017; Straub, 2021). However, as shown in prior literature, in some tasks, initial features provided by domain experts, that is, lists of keywords, can be incomplete and/or imprecise and thus require extensive expansion and modification (Gharibshah, 2020; Haj-Yahia et al., 2019). In this paper, we work towards filling this gap by presenting an automated categorization method in which we leverage a combination of unsupervised feature mining and supervised categorization techniques to enhance the expertise provided by domain experts and integrate that into document categorization models.

We bring this methodological work to the domain of biodiversity conversation, where categorizing documents by conservation actions can support the more effective allocation of resources and improved evaluation of social-ecological impact (Hayward, 2011; Miller, 2014; Waldron et al., 2017). More specifically, in this project, our collaborators from the domain of environmental science have partnered with the John D. and Catherine T. MacArthur Foundation to assess the social-ecological impact and outcomes of its funded conservation interventions over 40 years, and to identify the effects of long-term financial support for biodiversity conservation. A prerequisite for this impact assessment task is to categorize funding-related documents based on a classification schema of conservation actions known as the "International Union for Conservation of Nature (IUCN)",[1] which was created based on widely accepted theories and practices in conservation science (Salafsky et al., 2008). Since the documents of interest in this project are technical and lengthy, contain information irrelevant to impact assessment (e.g., replicated contents, project costs, and thank-you emails), and require in-depth reading based on domain expertise, our collaborators need to spend more than 1 h to annotate each document with respect to the IUCN categories. Therefore, automated categorization models are needed that can assist in efficiently labeling documents for further analysis at scale. Also, these models need to be built based on a small amount of training data due to the cost of data annotation.

In view of the domain experts' need for automated categorization methods that do not rely on large-scale annotated data, our paper aims to answer three research questions:

- First, what are the specific computational and methodological needs in domains where categorizing documents requires domain expertise and in-depth reading, and what are difficulties with applying existing, state-of-the-art categorization methods in these domains?
- Second, how can we design an expert-in-the-loop method that integrates domain expertise, unsupervised feature mining, and supervised categorization to categorize domain-specific documents based on a small amount of annotated data (in our case, 93 documents) such that we can minimize the amount of labor and time needed for manual annotation?
- Third, under what conditions can domain expertise enhance and improve automated categorization accuracy? The findings can inform researchers to further advance expert-in-the-loop categorization models for other domain-specific data and tasks.

2 | OVERVIEW ON CATEGORIZATION METHODS AND THEIR LIMITATIONS

Text categorization is typically performed by using a top-down (deductive or theory-driven) or bottom-up (inductive or data-driven) approach (Rezapour et al., 2020). With the top-down (theory-driven) approach, the primary task is to assign predefined labels to each textual unit of analysis, such as words, paragraphs, or documents. These labels can stem, for example, from prior theories and existing category schemas. This approach is particularly suitable when studies are driven by theory instead of data, and thus usually use predefined concepts and categories (Bhattacherjee, 2012; Lazer et al., 2009; Mazzocchi, 2015). The most common approach to automated top-down categorization is supervised machine learning, including classic feature-based learning (Alzamzami et al., 2020; Huang & Lu, 2010; Rezapour & Diesner, 2017) as well as more modern deep learning (Liu et al., 2021; Zhang & Zhang, 2020). For training such models, various characteristics of text data, such as lexical, syntactic, probabilistic distribution, and word embedding features, are widely and successfully used. For example, Liu et al. (2018) trained logistic regression (LR), support vector machine (SVM), and convolutional neural network (CNN) based classifiers with tf-idf and word embedding features to assign predefined privacy policy labels to sentences and segments of web privacy policy statements. Also, Huang and

Lu (2010) computed novel probabilistic distribution features and lexical features to build LR models for labeling scientific publications with Medical Subject Headings (MeSH) categories (Lipscomb, 2000).

With the bottom-up (data-driven) approach, the primary task is to quantitatively or qualitatively explore the data to identify latent categories directly from the data, and then trace these categories back to textual units of analysis for categorization (Bernard et al., 2016). This approach is particularly suitable when no well-defined categorization schemas exist for a dataset or research question. Latent categories can be identified by using unsupervised learning methods to uncover latent semantic topics in text collections. A common method to this end is topic modeling, which usually leverages Bayesian generative methods to uncover latent themes as lists of terms, which represent topics contained in text data (Blei et al., 2003). After arranging identified relevant topics into a category schema, each unlabeled document is then labeled with one or multiple categories in the schema based on a substantial presence test (typically according to the proportions of representative terms in the document) (Debortoli et al., 2016; Suominen & Toivanen, 2016). Another approach to bottom-up categorization is supervised learning based on human annotated ground truth data. For example, Rezapour et al. (2020) asked human annotators to closely read a sample of sentences from project reports and label them with any types of social impact they saw in the data. These impact labels were then grouped and synthesized into a fine-grained impact category schema. The researchers then built a supervised learning model with features such as tf-idf, parts of speech, and domain-specific information to categorize unseen documents based on this novel schema with near 80% accuracy.

Since this paper aims to design and evaluate an expert-in-the-loop method to build models that automatically categorize text data from the domain of biodiversity conservation, where scholars and practitioners have been widely using the IUCN classification schema as the theoretical framework for analyzing conservation actions, we also use the IUCN schema and consequently a top-down approach in this study.

Highly accurate, state-of-the-art categorization methods used for the top-down approach usually rely on large-scale, annotated training data with high-dimensional embedding feature sets (often hundreds of black-box dimensions computed by word embedding methods such as word2vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019). To get data annotated, researchers have been leveraging crowdsourcing through platforms such as Amazon Mechanical Turk. However, due to concerns about the reliability of annotations provided by non-experts on crowdsourcing platforms (Dow et al., 2012; Eickhoff, 2018), this solution is mainly suitable for straightforward annotation tasks that require little to no domain expertise or specialized training. Prior work on successfully crowdsourcing the annotation of sufficiently large volumes of training data can be divided into two groups. First, tasks where workers assign a valence value (typically positive, negative, or neutral) to reviews (typically of consumer products), tweets, and other types of information and digitized artifacts (Founta et al., 2018; Kim et al., 2019; Kumar et al., 2018). Instances of this task are sentiment analysis and opinion mining. The texts to labels are typically short and easy to comprehend, and the labels to assign are intuitively understandable. The hand-annotation process is further aided by providing clear guidelines, including definitions and examples for instances of categories. Second, tasks where workers locate and label the effect or their impression of pieces of text data. These annotations can often be done based on intuition and common sense, such as identifying topics (e.g., politics, technology, entertainment, and finance) from (social) media articles (Budak et al., 2016).

Nonexpert crowdsourcing annotations are only of little help when tasks require the consideration of context, holistic in-depth reading, and domain expertise (Barbier et al., 2012; Xintong et al., 2014). In these cases, annotation requires specialized training and expertise, such that assigning this task to nonexpert workers for large-scale annotations might not be feasible and/or affordable (Rezapour et al., 2020). As a consequence, one might not be able to obtain a sufficiently large amount of annotated training data for domain-specific document labeling tasks. Since state-of-the-art automated categorization methods with high predictive power and accuracy, such as embedding and graph-based deep learning models, usually require large amounts of training data (Chuang et al., 2014; Huang & Lu, 2010), building classifiers based on small amounts of annotated data and/or domain-specific categorization schemas has been a challenge.

Given the limitation of crowdsourcing, recent studies have also developed data augmentation and transfer learning methods, which mitigate problems due to a lack of annotated data. For data augmentation, existing texts are modified algorithmically or via human-crafted rules, the "newly" generated texts inherit the label(s) of the original data, and these data are then added to the given corpus of annotated data. For example, Wei and Zou (2019) found that new data created by applying sentence-level synonym replacement, random synonym insertion, random swap of word positions, and random word deletion can improve classification accuracy for five standard text classification tasks. Another method to this end is

back-translation, where existing, labeled texts in language A are translated into language B and then back into language A, resulting in additional labeled data (Edunov et al., 2018; Xie et al., 2019). In addition to data augmentation, another approach to model building with small-scale data is transfer learning, where an existing model trained on one annotated dataset (source dataset) is adapted to another dataset with light to no annotation (target dataset). Nowadays, the most widely used method for transfer learning is model fine-tuning, which trains a new model on a target dataset by adjusting parameters in a pretrained model obtained from a larger annotated source dataset (Croce et al., 2020; Rietzler et al., 2020). Adapting pretrained models across domains is also called domain adaptation. Current domain adaptation methods usually require datasets from more than one domain to have some similar characteristics, including but not limited to text structures and word or sentence distributions (Kadar & Iria, 2011; Vedula et al., 2019). The majority of data augmentation and transfer learning methods are developed and tested based on well-structured and/or easy-to-understand texts from a few genres or domains, mainly news (Croce et al., 2020; Sun et al., 2019), legal documents (Elnaggar et al., 2018; Wei et al., 2018), products reviews and descriptions (Rietzler et al., 2020; Vedula et al., 2019; Wei & Zou, 2019), Q&A data (Croce et al., 2020; Sun et al., 2019; Wei & Zou, 2019), online encyclopedia such as Wikipedia (Ibrahim et al., 2018; Xie et al., 2019), and biomedical publications (Lee et al., 2020; Sun et al., 2019; Sun & Yang, 2019; Vedula et al., 2019). However, little is known about the performance of these methods when working with unstructured and lengthy texts from other genres and domains as well as domain-specific categorization schemas. Moreover, progress with transfer learning, for example, for the genres of news, user-generated content, and academic reports, stems largely from leveraging existing models and datasets, such as pretrained BERT models for news (Devlin et al., 2019) and biomedical texts (Lee et al., 2020) as well as a variety of benchmark datasets. Low-resourced fields may have fewer to no pretrained models and benchmark datasets to leverage. In addition, the documents in our dataset (project proposals and reports) are not well-structured and require holistic reading based on domain expertise to identify categories. Thus, it is difficult to adopt existing domain adaption and transfer learning methods for projects. This situation might generalize to other situations when the existing pretrained models and annotated benchmark datasets are created based on nonoverlapping categorization schemas, text collections, and/or research questions (Salganik, 2019).

In the next sections, we describe our data, classification schema, categorization methods, and findings. We do not aim to degrade the value of current data augmentation and transfer learning techniques, and fully acknowledge possible improvements that those methods may bring. Our goal is rather to introduce a method for situations where there are light to no computational resources in terms of pre-trained models and annotated data such that data augmentation and transfer learning might not be suitable.

3 | DATA

3.1 | IUCN classification schema

The International Union for Conservation of Nature (IUCN) classification schema was developed by the Conservation Measures Partnership in conjunction with the IUCN organization to provide a comprehensive classification of all conservation actions (Salafsky et al., 2008). These conservation actions "are interventions that need to be undertaken to help improve the conservation status of the taxon being assessed."[2] The IUCN classification schema consists of 10 broad categories that represent strategies, interventions, activities, responses, and measures related to conservation actions: (1) land/water management, (2) species management, (3) awareness raising, (4) law enforcement and prosecution (5) livelihood, economic, and moral incentives, (6) conservation designation and planning, (7) legal and policy frameworks, (8) research and monitoring, (9) education and training, and (10) institutional development.[3] This schema is used by practitioners and researchers around the world to label and study conservation goals, actions, and outcomes, and to enable a better understanding of the scope and impact of conservation work (Leberger et al., 2020).

3.2 | Data and annotation

For this project, we use a corpus of project-related documents from the domain of biodiversity conservation. These conservation projects were funded through a MacArthur Foundation program that ran from 1979 through 2018. This program provided support to domestic and international organizations from across the world with the purpose of advancing research and practices protecting nature. However, little is known about the long-term outcomes of this program. To address this gap, we aim to find key factors associated with the impact and outcomes of conservation interventions, and to identify

the effects of long-term financial support for biodiversity conservation based on project proposals and reports.

When assessing and evaluating the outcomes of conservation projects based on text reports, it is helpful to first categorize each document or project into IUCN categories. Each project in our dataset is represented by two documents: a funding proposal, which grantees submitted to apply for support for their projects, and the grantee's final report on their work for a given project. Most projects have both proposals and projects, while a few only have proposals. Therefore, our corpus includes more than 2,000 projects and less than 4,000 documents. Different from some other grant proposals and reports in academia, which can be highly structured, the documents in our dataset do not have a uniform structure. Given that the projects in our dataset are from various fields, including physical environment protection, social advocacy, and capacity building, and from different countries and points in time (over a span of 40 years), the authors of these documents (grantees) can be assumed to have written and organized their proposals and reports to some degree based on community-defined and self-defined criteria, norms, and traditions. Furthermore, considering that these documents are lengthy, technical, and unstructured, it takes even domain experts more than 1 h to read a document to identify the best fitting IUCN categories. Therefore, we aim to develop an automated categorization method that supports domain experts in this process.

The first step in model building is to manually annotate data for training and testing. We had two domain experts from the field of environmental science do that based on the IUCN Classification schema, which contains detailed definitions, descriptions, and examples for each category.[4] The annotation process was independently created by the domain experts without our help, and is similar to common practices in their daily research work. Each document in our dataset can contain one or many IUCN categories, and thus the experts decided to assign all corresponding categories to each document. In other words, this is a multilabel categorization task. Moreover, since the domain experts found that identifying IUCN categories in a project proposal or report requires holistic reading of the whole document, they used documents as the unit of analysis rather than sentences or paragraphs. Before annotating the dataset with IUCN categories, the domain experts conducted a four-round pilot study to determine the most appropriate annotation process that results in the highest annotation agreement:

1. In the first round, the domain experts randomly selected 15 documents (a mixture of proposals and projects). Then they each, on their own, identified the three most important IUCN categories per document through in-depth reading and rank-ordered them. Given that many documents contain more than three categories and some categories are of similar importance in a document, the domain experts found that their perceptions of the top three categories varied significantly (a Kappa score of 0.45) from each other. Therefore, they created a revised annotation process
2. In the second round, they drew another random sample of 15 documents. Then each on their own used the same annotation process as above and also recorded their confidence in their label assignment. They then discussed the annotated categories that they were not confident about and modified their annotation after discussing the annotated categories that they were not confident about. This task resulted in a Kappa score of 0.69, which still did not reach the experts' expectations. They then defined a third annotation process
3. In the third round, the experts drew another random sample of 10 projects, identified all relevant IUCN categories per document regardless of their importance, and finally independently chose the top three categories per document. This round resulted in a Kappa score of 0.42, which is lower than the scores from the first two rounds. In view of this, the domain experts did not continue to record their annotation confidence because they assumed that they would not be satisfied with the Kappa score even if they modified annotation after a discussion based on their confidence. Therefore, they designed a fourth annotation process
4. The experts, each on their own, simply identified all IUCN categories that were present in a document without further ranking them. This annotation process resulted in a Kappa score of 0.85.

At this point, the experts were confident that the fourth process would lead to the labels with the highest agreement, and thus adopted the fourth process to label a new random sample of 93 documents (i.e., proposals and reports) from our corpus as the ground truth dataset for model training and evaluation. They divided this sample into two approximately equally sized parts. Each expert labeled one part on their own. They then discussed the labels which they were not confident about and adjusted them accordingly. Figure 1 shows the highly imbalanced distribution of labels across the annotated sample. Some categories, including 1 (land/water management), 2 (species management), and 4 (law enforcement and prosecution) are present in less than 10% of the documents, while others, including 8 (research and monitoring) and 10 (institutional development) are present in more than

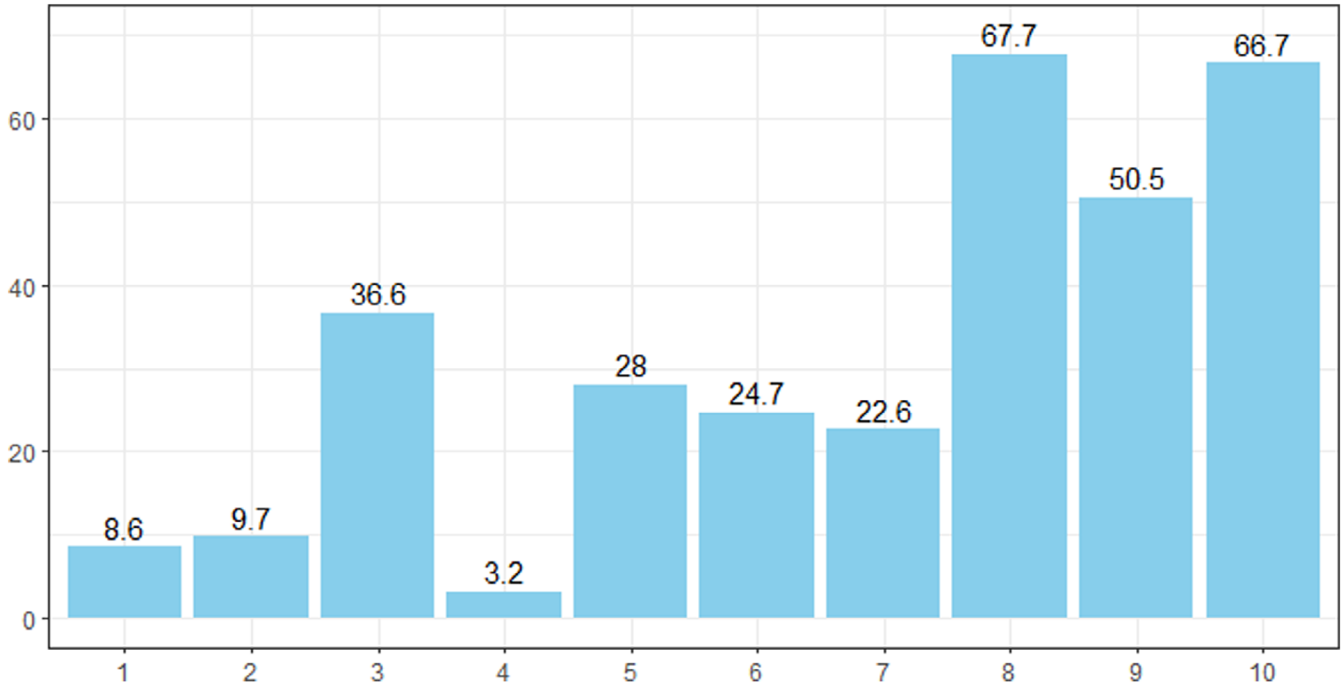

FIGURE 1 Percentages of occurrence of IUCN categories 1–10 in our annotated sample (values in percent, each document may contain multiple categories). For example, 8.6% of documents were labeled with Category 1, and 67.7% of documents were labeled with Category 8.

66% of the documents. For the majority of the documents (76 out of 93), more than one category (nine categories at most) was assigned, and every category co-occurred with any other categories at least once.

Only 93 documents were annotated because the process of labeling and consensus construction among the domain expert coders was time-consuming (more than 1 h per document). Even though annotating more documents might have been possible, it would have violated our goal of supporting rather than burdening domain experts. We also frequently checked in with the human coders to ensure that they were still comfortable with the procedures and methods we proposed in later stages other than the annotation. After the annotation was completed, we conducted a survey to elicit the human coders' perception of their knowledge of the IUCN categories, and analyzed the relationship between their self-reported domain knowledge, annotation confidence, and the prediction results (see Method and Results). We did this to better understand whether insecurities in human judgment and expertise translate into lower prediction results.

To prepare the annotated documents for model training, we converted the original PDF files into text files using a Python library. Spot checking some documents, we found some irrelevant types of information, such as email correspondence, cost reports, and names of participants, which we manually cleaned up in our labeled sample. For analysis, we decided to use both the messy (raw converted text files) and the manually cleaned data to investigate how the noisy information impacts prediction accuracy, and, by extension, whether practitioners need to spend time and effort on sanitizing the data when using our categorization method.

4 | METHODOLOGY

Given the small size of our annotated data (93 annotated documents), we cannot use advanced categorization models with complex features due to the high risk of overfitting. Therefore, a simple classification model with a few well-defined features seems most appropriate. In view of this, we develop a computer-assisted approach to provide the domain experts, who were also the annotators, with (refined) keywords through word embeddings for feature engineering and model building. As Figure 2 shows, our feature engineering and model building workflow consists of the following steps: (a) eliciting initial keyword lists from the domain experts, (b) using word embeddings to extract words, which are semantically similar to the initial keywords, from the annotated documents, (c) asking the same domain experts to select and refine the keywords produced in the first two steps, (d) calculating frequencies of the keywords from the third step and their co-occurrence in context as features, and (e) building and evaluating supervised models. We

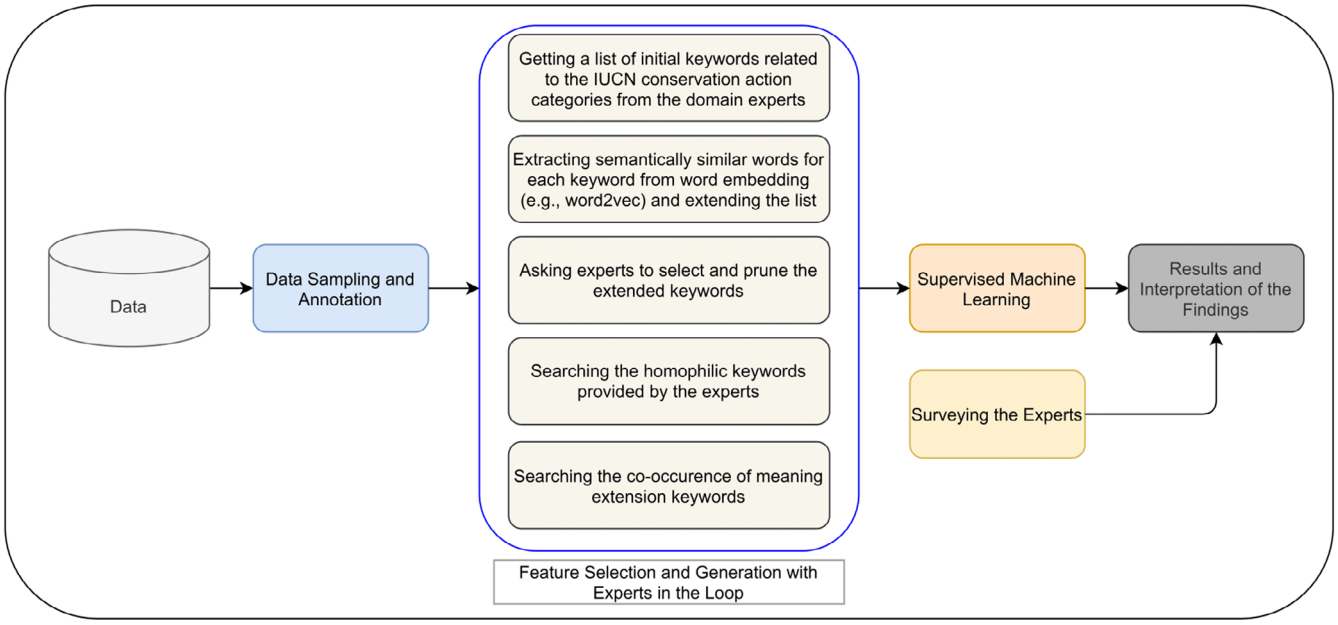

FIGURE 2 Categorization workflow

treated this multilabel classification task as 10 binary classification tasks, that is, we created one set of features for each of the 10 categories, and then leveraged five-fold cross-validation to train one binary classification model for each of the 10 classes.

4.1 | Keyword selection

4.1.1 | Initial domain-related keywords

We first asked the domain experts to provide a list of keywords that can be potentially used as features for predicting each of the 10 IUCN categories. The keywords could be proposed based on the experts' knowledge of the categories and domain, and/or from any resources that the experts found useful for this task. We told them that the proposed keywords had to be as exclusive and insightful as possible per each category. The domain experts first listed all keywords that came to their minds (they gave similar sets of keywords) and then modified or removed some of them after discussion. In the following sections, we use "initial keyword(s)" to exclusively refer to the words obtained in this step. Table 1 shows a list of five examplary keywords for each IUCN category. Among the lists of keywords they provided for the 10 categories, Category 7 (Legal and Policy Frameworks) had the lowest number of suggested keywords (#6), while Category 9 (Education and Training) and Category 10 (Institutional Development) received the highest number of keywords (#17).

4.1.2 | Keyword extension and refinement using word embedding

The initial keywords resulting from the abovementioned process represent features containing domain expertise that we can leverage for model building. While these keywords provide valuable insights, we also saw limitations with them: they seemed incomplete (with respect to covering a complete list of insightful and exclusive keywords for category prediction), imprecise (to exclusively identify an IUCN category), and not contextualized (more based on prior knowledge than their actual use in the proposals and reports). Searching the annotated documents for the initial sets of keywords, we identified two problems: First, the frequency of some of these keywords was low across all annotated documents (close to zero). This could be because some keywords came from the experts' theoretical assumptions, not the data. Second, some keywords provided by the experts represented words that are common in project proposals and reports, such as "management," "funding," "species," "protected area," and "conservation." While such terms are crucial for some categories, such as Species Management and Institutional Development, given the genre of the corpus, they may also occur in documents related to other categories; hence they are not discriminatory for a classifier.

To remedy these shortcomings, we leveraged word embeddings to (a) help the experts refine their selected keywords and (b) capture contextual information from the annotated dataset. Similar to recent studies in public health (Dai et al., 2017; Nikfarjam et al., 2015) and online

TABLE 1 List of IUCN categories and five initial exemplary keywords per category that were provided by domain experts

Category	IUCN conservation actions	Examples of initial keywords
1	Land/water management	Habitat, landscape, restoration, corridor, forestation
2	Species management	Reintroduction, captive, breeding, prioritization, ecological function
3	Awareness raising	Campaign, public awareness, media, workshop, community outreach
4	Law enforcement and prosecution	Justice, punish, offender, arrest, patrol, illegal
5	Livelihood, economic and moral incentives	Alternative livelihood, economic development, integrated, ecotourism
6	Conservation designation and planning	Protected area, national park, priority area, creation, zoning
7	Legal and policy frameworks	Legal mechanism, legislation, policy, law, influence policy
8	Research and monitoring	Database, inventory, survey, rare plant, scientific research
9	Education and training	Degree, capacity building, training, university, college
10	Institutional development	Infrastructure, operation, funding, finance, networking

content categorization (Haj-Yahia et al., 2019), leveraging word embeddings helps to extend word features by looking for similar words rather than directly vectorizing the entire document for classification. Since our goal with using word embeddings was to assist researchers who have limited technical expertise to extend and refine keyword lists, we selected Word2vec (Mikolov et al., 2013) instead of more advanced embedding models, such as ELMo (Peters et al., 2018) or BERT (Devlin et al., 2019). More specifically, ELMo and BERT will generate multiple vectors for a word in multiple sentences and thus a plethora of word vectors for the whole set of initial keywords. Since manually choosing relevant words from simplified vector spaces created by Word2Vec was already a challenging (but manageable) task for our domain experts, especially for categories about which they had less knowledge or felt less confident, we assumed that it may be even more difficult for them to deal with more complex vector spaces created by ELMo or BERT.

To create word embeddings, we removed stop words and punctuation, lemmatized the texts, and then trained a unigram Word2vec model. We extracted the top 50 similar words for each initial keyword that the domain experts provided. To check whether the embeddings trained on our annotated sample of 93 texts was robust, we trained another embedding model based on another sample of around 100 documents from our corpus, which resulted in 50 top words that were highly similar to the initial keywords according to the feedback from the domain experts. The experts then selected (combinations of) words from the first embeddings that seemed most relevant. We refer to the 10 lists of words (for 10 categories, respectively) selected from the embeddings as "extended keyword lists." After a manual inspection of the extended keyword lists, the experts found that these words can be divided into two groups: The first group consists of homophilic keywords, which belong to the same type of event or thing as the initial keywords. For example, the initial keyword "income" (from Category 5) has several homophilic keywords among the top 50 words in the embedding, such as "revenue," "saving," and "price." The experts carefully selected most relevant keywords from these automatically retrieved homophilic keywords, which we later used for feature engineering and model building. The second group consists of meaning-extension keywords, which capture contextual information that further specifies how an initial keyword was actually used in the data and thus extends the meaning of this word from its literal meaning. For example, the term "livelihood" was found (by the experts) to be associated with "adaptive," "promote," and "landscape." It is worth noting that meaning-extension can be either direct or indirect: "adaptive" and "promote" are direct meaning-extension keywords because they directly specify the type of livelihood or action related to livelihood. "Landscape," on the other hand, is an indirect meaning-extension keyword as it indicates that the keyword "livelihood" occurs in the context of "landscape" and should be understood as such in the project proposals and reports. If there is no such context, or "livelihood" occurs alone or alongside dissimilar words, the mere occurrence of "livelihood" might not be strong enough to suggest Category 5 as a label for that document. The experts carefully selected important meaning-extension keywords from the word embeddings outputs. However, as opposed to homophilic keywords, they labeled these keywords mainly as "keyword 1 + keyword 2" or "keyword 1 + keyword 2 + keyword 3," that is, co-occurring skip-grams instead of unigrams. The domain experts also explained that these skip-grams should co-occur within a sentence, paragraph, or section of a

document in any sequence. There can be other irrelevant words in between these meaning-extension keywords, such as "keyword 2 + irrelevant words + keyword 1 + keyword 3." In addition to these co-occurring skip-grams, the experts also found a few bigram keywords without any intervening words. Finally, the experts removed or altered some words from the initial keyword lists to reduce redundancy and refine the appropriateness of the keywords. As the result of this step, we obtained two sets of keywords: homophilic keywords (unigrams or bigrams) and meaning-extension keywords (n-grams or skip-grams). These keywords represent domain expertise features enhanced by unsupervised data mining.

4.2 | Classification models

4.2.1 | Feature engineering

We created two sets of keyword features: First, the frequency of homophilic keywords, measured as the frequency per keyword per document. Second, the frequency of meaning-extension keywords, measured as the co-occurrence of meaning-extension keywords, that is, the frequency of two and three meaning-extension words that co-occurred within a window size of 15 and 30, respectively.[5] We also created two sets of aggregated features by calculating the total frequency of homophilic keywords and of meaning-extension keywords. While aggregating individual features are often not a rigorous approach as it can cause information loss and introduce noise and biases if a few individual features are poor predictors, the small size of our annotated dataset may not allow for large feature sets in models. Therefore, we built two sets of models for each categorization task: one based on individual keyword features, and one based on aggregated keyword features. We then selected the best performing model. In addition to the keyword features, our models also include the geographical region (continent) of each project and the type of the document (proposal or report) as two additional features.

4.2.2 | Classifier

Given the small size of our annotated data, we selected decision trees (DT), random forests (RF), and support vector machines (SVM) as classifiers. For each dataset (messy and clean) and each set of features (individual and aggregated), we trained one model per classifier and category via five-fold cross-validation. The hyperparameters in these models, including but not limited to loss functions, kernel functions, penalty methods, splitting strategies, and/or feature weights, were tuned via grid search. We found that the best prediction for each category was usually achieved by RF or SVM rather than DT. A probable explanation for that might be that the small size of the training data led to overfitting (Ying, 2019), which DT is least capable of overcoming this problem (James et al., 2013). For each category, we will only present the result from the best model.

4.2.3 | Comparison

We implemented the following models to enable comparative evaluations of model performance.

- **Dummy classifier:** This simply classifies all documents as the majority class in the data. This is a widely used baseline when the annotated dataset is highly imbalanced (Agarwal et al., 2019). Any other classifier is expected to not perform worse than this baseline.
- **Meta-information:** This only considers document-level features, including the geolocation (continent) and document type feature (proposal or report). This baseline measures performance in lack of any domain-knowledge feature, that is, the initial and extended keywords selected by domain experts.
- **Meta-information + initial keywords:** These models contain both document-level features and initial keyword features. By comparing their performance to the meta-information models, we can test the isolated gain in performance due to the initial domain expertise provided by domain experts, which may still be incomplete, imprecise, and not contextualized.
- **Meta-information + initial keywords + extended keywords:** These models contain document-level features, initial keyword features, and extended keyword features. By comparing them to the meta-information + initial keywords model, we can test the isolated gain in performance due to domain expertise enhanced (extended) by an unsupervised machine learning method, namely word embeddings. This enhanced domain expertise is a combination of domain expertise, interpretable data mining results, and contextualized information from the documents.

4.3 | Survey

Since we hypothesize that domain knowledge improves categorization accuracy, we asked the expert annotators the following questions:

- How confident were you in your decisions about categorizing documents?

TABLE 2 Experts' confidence or familiarity with categories ranging from −1 (*very unconfident/unfamiliar*) to 1 (*very confident/familiar*)

Category	Annotation	Initial keywords	Keywords extension	Familiarity with the category	Sum
1	−0.5	0.25	0	0.5	0.25
2	−0.5	0.25	−0.25	0.5	0
3	0	0	0	0.5	0.5
4	−1	0	0	0	−1
5	−0.75	0.5	0.5	1	1.25
6	0.5	0.5	0.75	1	2.75
7	−1	0	−1	0	−2
8	0.5	0.75	0.5	1	2.75
9	0.5	0.5	0.75	0.75	2.5
10	0.5	0.5	−1	0.5	0.5

- How confident were you in the list of initial keywords you provided per category?
- How confident were you in your selection of additional keywords from a given list (which was generated with word embedding methods)?
- How familiar were you with the domain knowledge per category?

They were asked to choose a score (including decimals) from −1 (very unconfident/unfamiliar) to 1 (very confident/familiar). They were also encouraged to leave additional comments on their experience with their decision-making processes and expertise.[6] We aggregated the survey results by calculating the sum of these scores for each category to evaluate the annotators' overall knowledge of each category. We did not share the automated categorization results with them prior to this survey. Table 2 shows the survey results. The scores per annotator per category vary, indicating that some categories were easier for them to handle than others. For example, the overall score for Category 7 (Legal and Policy Frameworks) is the lowest (−2). We assume that this is because this category requires more expertise in the domain of law and legal studies than in environmental science. It is worth noting that the aspects of domain expertise reflected by the different confidence/familiarity metrics in the table are heterogeneous and not always positively correlated. Therefore, a simple sum of these metrics might be misleading in certain situations, while it can provide a rough understanding of the overall domain expertise in other situations. For example, for Category 5, the scores of "initial keywords," "keywords extension," and "familiarity with the category" are high, but the score of "annotation" is very low. Adding these scores will dilute the large difference between them. Therefore, in the following section, we also investigated the individual metrics when we found the sum (overall knowledge) to be misleading.

5 | RESULTS

Table 3 and Figure 3 show the accuracy, F1 score, and ROC-AUC curve for the classifiers we trained to predict the IUCN Categories 1–10. In the following discussion, we will use the term "performance" to refer to both overall accuracy and the ability to distinguish between classes. The prediction results can be divided into four groups: highly imbalanced classes (pink), lack of sufficient domain expertise (green), sufficient domain expertise & balanced classes (blue), and special categories (yellow).

- **Categories in pink columns (1, 2, and 4) - highly imbalanced classes:** The binary class distributions for Categories 1, 2, and 4 are highly imbalanced (see Figure 1) in our annotated data. For example, we only have three documents for Category 4. Training a model on too small training data is highly likely to lead to overfitting. As mentioned in the Related Work section, for larger, that is, medium-to-large-sized data, a variety of resampling and data augmentation methods (Dıéez-Pastor et al., 2015; Ibrahim et al., 2018) can help to balance the distribution of data points across categories, but this is not applicable for the data in the pink colored categories due to their size. Moreover, this sparsity may also contribute to the experts' lack of confidence in and knowledge about a category. In fact, our experts told us that they were not able to be sufficiently familiar with these categories because they were too rare in our datas.
- **Categories in green columns (3 and 7) - insufficient domain expertise:** The domain experts

TABLE 3 Accuracy and F1 score (in parentheses) of prediction results for Categories 1–10 (values are in percent. When a model simply classifies all documents as the majority class, the F1 score was manually labeled as 0.00 to indicate the low quality of this model

	1	2	3	4	5
Dummy classifier	91.3 (0.00)	90.2 (0.00)	63.4 (0.00)	96.8 (0.00)	72.0 (0.00)
Meta	91.3 (0.00)	90.2 (0.00)	63.4 (0.00)	96.8 (0.00)	72.0 (0.00)
Meta + IK (m)	91.3 (0.00)	90.2 (0.00)	66.8 (27.7)	96.8 (0.00)	72.0 (0.00)
Meta + IK + EK (m)	91.3 (0.00)	90.2 (0.00)	66.8 (27.7)	96.8 (0.00)	**72.8 (0.10)**
Meta + IK (c)	91.3 (0.00)	90.2 (0.00)	66.8 (27.7)	96.8 (0.00)	72.0 (0.00)
Meta + IK + EK (c)	91.3 (0.00)	90.2 (0.00)	66.8 (27.7)	96.8 (0.00)	**72.8 (0.10)**
	6	**7**	**8**	**9**	**10**
Dummy classifier	75.2 (0.00)	77.4 (0.00)	67.7 (0.00)	50.5 (0.00)	66.7 (0.00)
Meta	75.2 (0.00)	77.4 (0.00)	67.7 (0.00)	62.5 (59.5)	66.7 (0.00)
Meta + IK (m)	79.6 (31.1)	81.9 (31.4)	**73.3 (80.0)**	72.0 (68.9)	69.9 (81.0)
Meta + IK + EK (m)	**85.1 (54.0)**	81.9 (31.4)	72.9 (79.2)	**76.3 (75.4)**	69.9 (81.0)
Meta + IK (c)	79.6 (33.1)	81.9 (31.4)	67.7 (80.2)	74.2 (75.1)	67.9 (74.3)
Meta + IK + EK (c)	**87.2 (59.7)**	81.9 (31.4)	68.4 (80.1)	**78.4 (79.0)**	**72.0 (81.4)**

Note: Colors: pink (columns 1, 2, 4) = highly imbalanced data distributions; green (columns 3, 7) = experts' lack of confidence or familiarity regarding IUCN categories; yellow (columns 8, 10) = special category characteristics such as special topics and their associated keywords; blue (columns 5, 6, 9) = the best condition where data distributions were balanced, and experts had a good understanding of categories. Meta = meta-information (no domain expertise); IK = initial keywords (initial domain expertise); EK = extended keywords (enhanced domain expertise via word embedding); m = messy data; c = clean data. The bold values show the best result for each category.

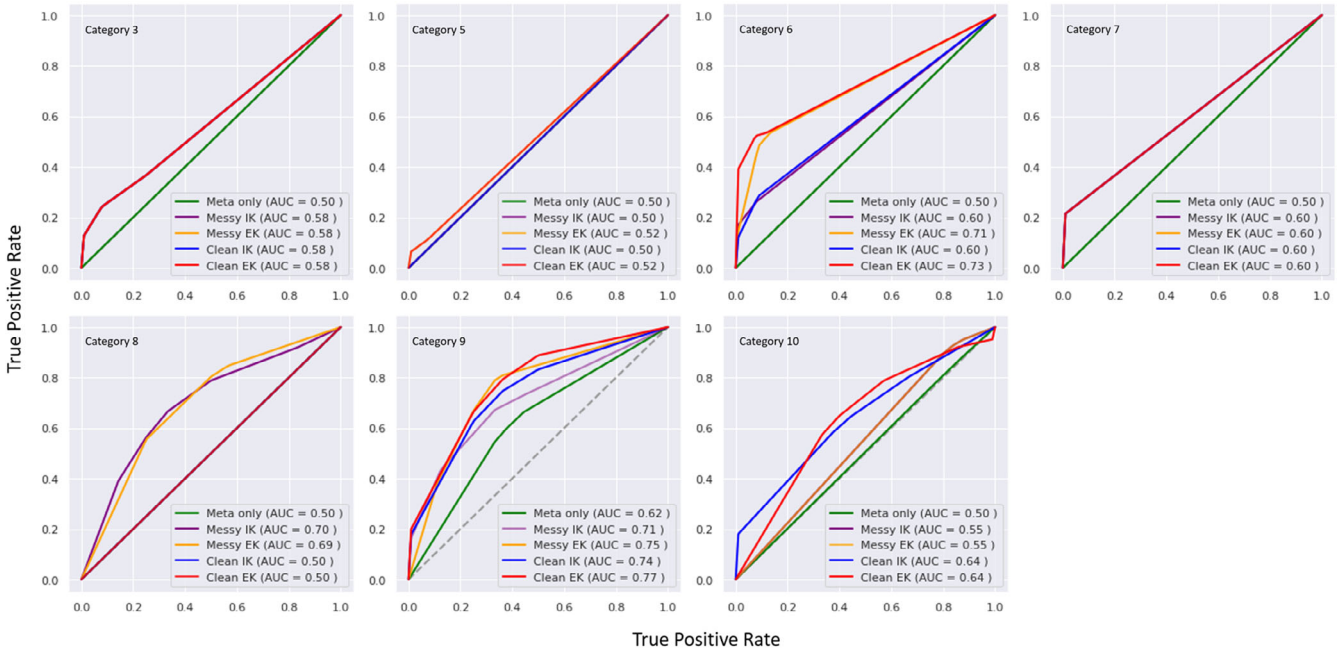

FIGURE 3 Average ROC curves (based on five-fold cross-validation) for all categories except 1, 2, and 4, where all documents were classified as the majority class, and thus none of the features can improve performance (Areas under the ROC curves are shown in legends. A few curves overlap, e.g., when adding new features into models does not improve performance).

reported not having a good understanding of Categories 3 and 7. The distributions of those categories are not highly imbalanced. The prediction results show that the initial set of keywords provided by the experts can improve classification accuracy, but the additional enhancements we tested (extended keyword features) led to no significant further improvements. In other words, there is no significant difference between the

results for models with basic versus extended keyword sets for both messy and clean data. A possible explanation for this result is that keyword extension through word embeddings requires more domain expertise than the initial keyword selection because the experts had to consider more complex contexts represented by the word embedding. In other words, the experts may fail to extract useful keywords from word embeddings due to their lack of domain expertise in certain subdomains. Despite this issue, the prediction models for these categories perform better than those for the categories in the pink columns in terms of F1 scores, and the improvements are brought by the initial keywords.

- **Categories in blue columns (5, 6, and 9) - sufficient domain expertise and balanced classes:** This is the best situation for small-sized data, that is, the distributions are balanced, and the experts are said to have a good overall knowledge of these categories. As the results for Categories 6 and 9 in Table 3 show, compared with using only meta-information (geolocation information and document type), using the initial keywords improves prediction accuracy by 4.4 and 9.5% in the messy data, and the extended keywords further improve this accuracy by an additional 5.5 and 4.3% in the messy data. The F1 scores and ROC-AUCs also largely improve. However, the results for Category 5 are different (only a slight increase), even though the domain experts reported having a good overall knowledge of this category (score of 1.25). In fact, the difference in scores between Categories 5 versus 6 and 9 might stem from the differences in confidence about annotation. Moreover, the lack of the experts' confidence in annotation may have resulted in potentially less reliable annotated data and lower performance of the model. Overall, the results in the blue columns suggest that domain expertise, when used as a feature, is associated with improvements in classification results for both clean and messy data in small text collections where the class distribution is not highly imbalanced. However, difficulties with annotation might nullify these improvements.
- **Category 8 (yellow) - highly exclusive keywords:** This category (research and monitoring) is special because it contains too many highly exclusive keywords, although it looks similar to Categories 6 and 9 in terms of class distribution and knowledge scores. While in Categories 6 and 9, the models with extended keywords and clean data performed better than the models with initial keywords and messy data, the clean data model for Category 8 performs worse than the one with clean data, and the extended keywords do not bring much of an improvement over the model only with initial keywords. There are two possible explanations for these observations. First, Category 8 is the only IUCN category that is related to scientific research, while the other nine categories are more applied. Furthermore, some of the annotated documents (i.e., funding reports) in our sample data that include this category are scholarly work and academic papers. Regardless of the content of these academic papers that may or may not be directly relevant to core topics in proposals and reports, these papers might contribute unique, discriminatory keywords, such as "model," "research," "methodological," and "collect data," as features. Hence, the clean data, from which we removed academic papers and only kept funding proposals and reports, lack these discriminatory keywords. Second, the experts who annotated the data and selected the keywords are environmental scientists from an academic institution and thus might be most familiar with this category. The initial keywords they selected already contained precise information as features, such that the extended keywords might not provide enough additional information to significantly improve prediction.
- **Category 10 (yellow) lack of exclusive keywords:** This (institutional development) is another special category without enough exclusive keywords. Our experts shared with us their concerns about this category since the generated keywords were general and nonexclusive. Many keywords associated with this category, such as "fundraising," "management," and "operation," can also occur in other categories. Indeed, our experts reported they had an overall good knowledge of this category and were able to propose initial keywords, but they were not confident about the usefulness of these keywords for automated categorization. Additionally, they found that most words returned via word embeddings were also nonexclusive and that it was difficult for them to extract additional useful keywords (see the confidence score of keyword extension in Table 2). In other words, annotating for this type of category needs more holistic, in-depth reading, and the extended keyword feature was only of little help. Not surprisingly, the performance of our domain-knowledge-based method in Category 10 is worse than that for Categories 6 and 9.

Finally, the comparison of categorization performance between clean and messy data suggests that cleaning data leads to slightly higher accuracy. However, considering the costs of data cleaning and that improvements are not universal, doing data cleaning or not is a trade-off between accuracy and costs.

6 | CONCLUSION AND DISCUSSION

Our work presents a solution for combining domain expertise with NLP methods to develop models that facilitate labeling a corpus from the domain of biodiversity conservation—a low-resourced field in terms of NLP—using a small number of documents annotated by domain experts. The proposed solution may also generalize to categorization tasks where only a few resources exist, lengthy text documents with little structure need to be analyzed, and domain expertise and/or in-depth reading are necessary for labeling training data. Exemplary genres include project documents from private and nonprofit organizations (e.g., project proposals and reports), casual writings (e.g., personal writings and correspondence), and ethnographic notes. In addition, even though domain experts have been analyzing news articles and social media posts, their research questions, analytical frameworks, and categorization schemas often differ from general domain applications, and thus the information they want to uncover is domain-specific. These differences lead to more costly and time-consuming categorization tasks as well as a lack of computational resources (i.e., pretrained models and benchmark datasets) since most existing resources were created to satisfy common rather than domain-specific needs.

We have shown that our proposed solution improves the quality of features by integrating domain expertise rather than optimizing prediction accuracy via novel algorithms. Domain-specific lexicons created by experts have been widely used for incorporating domain expertise into categorization models, and researchers often use domain-specific lexicons as features to categorize documents (Fellbaum, 2010; Huang & Lu, 2010; Rezapour et al., 2019). However, we found that the initial keywords provided by domain experts can be incomplete (w.r.t. covering all concepts and related topics), imprecise (w.r.t. exclusively identifying a category), and not contextualized (more based on prior human knowledge than their actual usage in documents). To enrich and refine feature sets that only contain keywords initially proposed by experts, we leveraged word embeddings. Recent studies have shown how the geometry of vector spaces resulting from these embeddings can be interpreted in terms of complex semantic relations and social meaning (Caliskan et al., 2017; Garg et al., 2018; Kozlowski et al., 2019). Building upon this idea, we asked the experts to select additional keywords from generated vector spaces based on their domain expertise. Our method relies more on the input from domain experts than data mining methods (Haj-Yahia et al., 2019) for feature enrichment for two reasons. First, the benchmark datasets utilized in prior studies are large-scale and content-focused, while our corpus (including both annotated and unannotated data) is small and contains a large proportion of information irrelevant to our study. Embeddings trained based on such data are likely to be imprecise and/or irrelevant. Hence, we leveraged input from domain experts, including their judgment on which similar words are potentially useful, to offset such undesirable impacts. Second, while data mining methods may extract a larger number of extended keywords from embeddings than the number of keywords elicited from experts, small training datasets are unlikely to support that many features. Using our method, experts can focus on selecting a few high-quality features. Our findings suggest a new way to leverage unsupervised learning algorithms when only small amounts of data can be labeled: Experts can utilize their domain knowledge to identify new features or revise previous features based on interpretable unsupervised learning results, regardless of imprecision and/or irrelevance caused by small-scale annotated data and the high dimensionality of the original results.

Our results also reveal a variety of challenges and limitations with our approach. First, data imbalance is a challenge for automated categorization because rare categories lead to overfitting or having to skip a class. Second, the positive correlation between categorization accuracy and experts' understanding of a given category we observed emphasizes the importance of expertise in domain-specific categorization tasks. Third, our results imply that there can be unexpected, undesirable results due to special characteristics of one or more categories, such as large portions of nonexclusive keywords (Category 10) and highly exclusive keywords (Category 8). Finally, our results indicate that difficulties with annotation (Category 5) are associated with failure in model building, even when experts have a solid understanding of a category and the class distribution is not highly imbalanced. Furthermore, annotators can differ in their label assignment, for example, when texts are complex or ambiguous, and meaning is implicit or expressed between the lines. Overall, these limitations caused by a combination of domain and data problems suggest that closer collaborations between domain experts and data scientists are beneficial for future studies.

In summary, our paper (a) illustrates difficulties with building automated categorization models in low-resource fields, (b) describes an automated categorization method that integrates domain expertise, unsupervised feature mining, and supervised categorization based on a small number of documents, and (c) analyzes the relationship between

data characteristics, domain expertise, and categorization accuracy. We hope our work can shed light on a new approach to collaboration between domain experts and data scientists, which emancipates domain experts from costly and time-consuming manual work and avoids burdening them to annotate a larger dataset than they would do in their daily research activities.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support from the John D. and Catherine T. MacArthur Foundation.

ORCID

Kanyao Han https://orcid.org/0000-0003-2100-8637

ENDNOTES

[1] See the latest version at https://conservationstandards.org/library-item/threats-and-actions-taxonomies/

[2] https://conservationstandards.org/library-item/threats-and-actions-taxonomies/

[3] The IUCN conservation actions classification has a hierarchical structure (three levels). In the highest level, all conservation actions are grouped into three categories: (a) target restoration/stress reduction actions, (b) behavioral change/threat reduction actions, and (c) enabling condition actions. The 10 categories listed in this paper and used for categorization are in the middle level. Furthermore, in the lowest level, the categories are divided into fine-grained subcategories, each providing a comprehensive list of actions.

[4] Code book at https://bit.ly/2ZMLrc2

[5] We tested various window sizes and found the selected ones to produce robust results.

[6] We asked experts about their decision-making processes and expertise twice during this project. The first time was informal, and the second time was a formal survey approved by the IRB at our university. The annotators provided the same answers both times.

REFERENCES

Agarwal, P., Sharma, M., & Chandra, S. (2019). Comparison of machine learning approaches in the prediction of terrorist attacks. In *2019 twelfth international conference on contemporary computing (IC3)* (pp. 1–7). IEEE.

Alzamzami, F., Hoda, M., & El Saddik, A. (2020). Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation. *IEEE Access*, *8*, 101840–101858.

Barbier, G., Zafarani, R., Gao, H., Fung, G., & Liu, H. (2012). Maximizing benefits from crowdsourced data. *Computational and Mathematical Organization Theory*, *18*(3), 257–279.

Bernard, H. R., Wutich, A., & Ryan, G. W. (2016). *Analyzing qualitative data: Systematic approaches*. SAGE.

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices*. University of South Florida.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borg, D., Sestito, G. S., & da Silva, M. M. (2021). Machine-learning classification of environmental conditions inside a tank by analyzing radar curves in industrial level measurements. *Flow Measurement and Instrumentation*, *79*, 101940.

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(S1), 250–271.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Chuang, J., Wilkerson, J. D., Weiss, R., Tingley, D., Stewart, B. M., Roberts, M. E., Poursabzi-Sangdeh, F., Grimmer, J., Findlater, L., Boyd-Graber, J., & Heer, J. (2014). Computer-assisted content analysis: Topic models for exploring multiple subjective interpretations. In *Advances in neural information processing systems workshop on human-propelled machine learning* (pp. 1–9). NeurIPS.

Croce, D., Castellucci, G., & Basili, R. (2020). Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2114–2119). Association for Computational Linguistics.

Dai, X., Bikdash, M., & Meyer, B. (2017). From social media to public health surveillance: Word embedding based clustering method for twitter classification. *SoutheastCon*, *2017*, 1–7.

Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems*, *39*(1), 7–135.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics.

Dıéez-Pastor, J. F., Rodrıéguez, J. J., Garcıéa-Osorio, C., & Kuncheva, L. I. (2015). Random balance: Ensembles of variable priors classifiers for imbalanced data. *Knowledge-Based Systems*, *85*, 96–111.

Dow, S., Kulkarni, A., Klemmer, S., & Hartmann, B. (2012). Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 1013–1022). ACM.

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 489–500). Association for Computational Linguistics.

Eickhoff, C. (2018). Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 162–170). ACM.

Elnaggar, A., Otto, R., & Matthes, F. (2018). Named-entity linking using deep learning for legal documents: A transfer learning approach. arXiv preprint arXiv:1810.06673.

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: Computer applications* (pp. 231–243). Springer.

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth international AAAI conference on web and social media*. AAAI.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), E3635–E3644.

Gerla, V., Kremen, V., Macas, M., Dudysova, D., Mladek, A., Sos, P., & Lhotska, L. (2019). Iterative expert-in-the-loop classification of sleep psg recordings using a hierarchical clustering. *Journal of Neuroscience Methods*, *317*, 61–70.

Gharibshah, J. (2020). *Extracting actionable information from security forums* (PhD dissertation). University of California, Riverside.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Guo, X., Yu, Q., Li, R., Alm, C. O., Calvelli, C., Shi, P., & Haake, A. (2016). An expert-in-the-loop paradigm for learning medical image grouping. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 477–488). Springer.

Haj-Yahia, Z., Sieg, A., & Deleris, L. A. (2019). Towards unsupervised text classification leveraging experts and word embeddings. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 371–379). Association for Computational Linguistics.

Hayward, M. W. (2011). Using the iucn red list to determine effective conservation strategies. *Biodiversity and Conservation*, *20*(12), 2563–2573.

Holzinger, A., Plass, M., Holzinger, K., Crisan, G. C., Pintea, C.-M., & Palade, V. (2017). A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop. arXiv preprint arXiv:1708.01104.

Huang, M., & Lu, Z. (2010). Learning to annotate scientific publications. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 463–471). ACM.

Ibrahim, M., Torki, M., & El-Makky, N. (2018). Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 875–878). IEEE.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kadar, C., & Iria, J. (2011). Domain adaptation for text categorization by feature labeling. In *European conference on information retrieval* (pp. 424–435). Springer.

Kim, J., Amplayo, R. K., Lee, K., Sung, S., Seo, M., & Hwang, S.-W. (2019). Categorical metadata representation for customized text classification. *Transactions of the Association for Computational Linguistics*, *7*, 201–215.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905–949.

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference* (pp. 933–943). ACM.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebaragary, T., Michael, K., Royand, M., & van Alstyne, M. (2009). Computational social science. *Science (New York, N.Y.)*, *323*(5915), 721–723.

Leberger, R., Rosa, I. M., Guerra, C. A., Wolf, F., & Pereira, H. M. (2020). Global patterns of forest loss across iucn categories of protected areas. *Biological Conservation*, *241*, 108299.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, *88*(3), 265.

Liu, F., Wilson, S., Story, P., Zimmeck, S., & Sadeh, N. (2018). *Towards automatic classification of privacy policy text* (Technical Report CMU-ISR-17-118R and CMULTI-17-010). School of Computer Science Carnegie Mellon University.

Liu, J., Singhal, T., Blessing, L. T., Wood, K. L., & Lim, K. H. (2021). Crisisbert: A robust transformer for crisis classification and contextual crisis embedding. In *Proceedings of the 32nd ACM conference on hypertext and social media* (pp. 133–141). ACM.

Lu, Y., Pan, J., & Xu, Y. (2021). Public sentiment on chinese social media during the emergence of Covid19. *Journal of Quantitative Description: Digital Media*, *1*, 1–47.

Mazzocchi, F. (2015). Could big data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO Reports*, *16*(10), 1250–1255.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Miller, D. C. (2014). Explaining global patterns of international aid for linked biodiversity conservation and development. *World Development*, *59*, 341–359.

Monroe, B. L. (2019). The meanings of "meaning" in social scientific text analysis. *Sociological Methodology*, *49*(1), 132–139.

Nikfarjam, A., Sarker, A., O'connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, *22*(3), 671–681.

Oliver, P. (2019). Great methods reveal their own limitations. *Sociological Methodology*, *49*(1), 63–68.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the conference of the North American chapter of the Association Computational Linguistics*. Association Computational Linguistics.

Rezapour, R., Bopp, J., Fiedler, N., Steffen, D., Witt, A., & 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* (pp. 6777–6785). LREC.

Rezapour, R., & Diesner, J. (2017). Classification and detection of micro-level impact of issue-focused documentary films based on reviews. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1419–1431). Association for Computing Machinery.

Rezapour, R., Shah, S. H., & Diesner, J. (2019). Enhancing the measurement of social effects by capturing morality. In *Proceedings*

of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 35–45). Association for Computational Linguistics.

Rietzler, A., Stabinger, S., Opitz, P., & Engl, S. (2020). Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th language resources and evaluation conference* (pp. 4933–4941). LREC.

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: Computation, application, and evaluation* (Vol. 4, pp. 1–20). NeurIPS.

Ryazanov, I., Nylund, A. T., Basu, D., Hassellöv, I.-M., & Schliep, A. (2021). Deep learning for deep waters: An expert-in-the-loop machine learning framework for marine sciences. *Journal of Marine Science and Engineering*, *9*(2), 169.

Salafsky, N., Salzer, D., Stattersfield, A. J., Hilton-Taylor, C., Neugarten, R., Butchart, S. H., Collen, B., Cox, N., Master, L. L., O'connor, S., & Wilke, D. (2008). A standard lexicon for biodiversity conservation: Unified classifications of threats and actions. *Conservation Biology*, *22*(4), 897–911.

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1–10). Association for Computational Linguistics.

Straub, J. (2021). Machine learning performance validation and training using a "perfect" expert system. *MethodsX*, *8*, 101477.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics* (pp. 194–206). Springer.

Sun, C., & Yang, Z. (2019). Transfer learning in biomedical named entity recognition: An evaluation of bert in the pharmaconer task. In *Proceedings of the 5th workshop on BioNLP open shared tasks* (pp. 100–104). Association for Computational Linguistics.

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464–2476.

Vedula, N., Maneriker, P., & Parthasarathy, S. (2019). Bolt-k: Bootstrapping ontology learning via transfer of knowledge. In *The world wide web conference* (pp. 1897–1908). Association for Computing Machinery.

Waldron, A., Miller, D. C., Redding, D., Mooers, A., Kuhn, T. S., Nibbelink, N., Roberts, J. T., Tobias, J. A., & Gittleman, J. L. (2017). Reductions in global biodiversity loss predicted from conservation spending. *Nature*, *551*(7680), 364–367.

Wang, Y. (2014). On a novel cognitive knowledge base (CKB) for cognitive robots and machine learning. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, *6*(2), 41–62.

Wei, F., Qin, H., Ye, S., & Zhao, H. (2018). Empirical study of deep learning for text classification in legal document review. In *IEEE international conference on big data (Big Data)* (Vol. 2018, pp. 3317–3320). IEEE.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 6382–6388). Association for Computational Linguistics.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., & Le, Q. V. (2019). Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848.

Xintong, G., Hongzhi, W., Song, Y., & Hong, G. (2014). Brief survey of crowdsourcing for data mining. *Expert Systems with Applications*, *41*(17), 7987–7994.

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, *53*(6), 4335–4385.

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022.

Zhang, H., & Pan, J. (2019a). Casm: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, *49*(1), 1–57.

Zhang, H., & Pan, J. (2019b). The challenges of "more data" for protest event analysis. *Sociological Methodology*, *49*(1), 76–82.

Zhang, H., & Zhang, J. (2020). Text graph transformer for document classification. In *Conference on empirical methods in natural language processing (EMNLP)* (pp. 8322–8327). Association for Computational Linguistics.

How to cite this article: Han, K., Rezapour, R., Nakamura, K., Devkota, D., Miller, D. C., & Diesner, J. (2023). An expert-in-the-loop method for domain-specific document categorization based on small training data. *Journal of the Association for Information Science and Technology*, *74*(6), 669–684. https://doi.org/10.1002/asi.24714