Forest Policy and Economics 134 (2022) 102624

Contents lists available at ScienceDirect

Forest Policy and Economics

journal homepage: www.elsevier.com/locate/forpol

Using machine learning to identify incentives in forestry policy: Towards a new paradigm in policy analysis

Daniel Firebanks-Quevedo [a,*], Jordi Planas [b], Kathleen Buckingham [c], Cristina Taylor [d], David Silva [a], Galina Naydenova [b], René Zamora-Cristales [d]

[a] Data Science for Social Good, Chicago, IL 60637, USA
[b] Omdena, 449 Hawthorne Ave #3, Palo Alto, CA 94301, USA
[c] veritree, 1275 Venables St; Unit #230, Vancouver, British Columbia V6A 2E4, Canada
[d] World Resources Institute, 10G St NE, #800, Washington, DC 20002, USA

ARTICLE INFO

Keywords:
Data-science
Economic-incentives
Environmental-policy
Politics
Machine learning

ABSTRACT

As 2021 saw the launch of the United Nations Decade on Ecosystem Restoration, it highlighted the need to prepare for success over the decade and to understand what public economic and financial incentives exist to support sustainable forest and landscape restoration. To date, Initiative 20 × 20, a coalition of 18 Latin American countries, has committed to place 50 million hectares under restoration and conservation by 2030. Understanding the public policies in these countries that turn those commitments into action, however, is very labor-intensive, requiring decision makers to read and analyze thousands of pages of documents that span multiple sectors, ministries, and scales that lie outside of their areas of expertise. To address this, we developed a semi-automated policy analysis tool that uses state-of-the-art Natural Language Processing (NLP) methods to mine policy documents, assist the labeling process carried out by policy experts, automatically identify policies that contain incentives and classify them by incentive instrument from the following categories: direct payments, fines, credit, tax deduction, technical assistance and supplies. Our best model achieves an F1 score of 93–94% in both identifying an incentive and its policy instrument, as well as an accuracy of above 90% for 5 out of 6 policy instruments, reducing multiple weeks of policy analysis work to a matter of minutes. In particular, the model properly identified the relative frequency of credits, direct payments, and fines that exist as the primary policy instruments in these countries. We also found that tax deductions, supplies, and technical assistance are much less used among most of the countries and that, oftentimes, the policy documents describe economic incentives for restoration in vague and intangible terms. In addition, our model is designed to constantly improve its performance with more data and feedback from policy experts. Furthermore, while our experiments were run on Spanish policy documents, we designed our framework to be widely scalable to policies from different countries and multiple languages, limited only by the number of languages supported by current multilingual NLP models. Using a standardized approach to generate incentives data could provide an evidence-based and transparent system to find complementarity between policies and help remove barriers for implementers and policymakers and enable a more informed decision-making process.

1. Introduction

As concerns of the global impacts from climate change grow, there has been an increased involvement of the private sector and civil society in environmental policymaking (Cockli and Moon, 2020; Scoones, 2016), along with an increase in the sharing of policy ideas across the world (Affolder, 2019). Just as these trends continue to develop, it becomes imperative that we strive for reliable, reproducible, rapid and scalable environmental policy analysis. While there have been techniques developed to assist forest policy research such as scenario analysis, behavior models, and decision support systems, the experience with using these tools is limited (Garcia-Gonzalo and Borges, 2019), and none of them deal with analyzing the actual text from policy documents in an automated and scalable manner. In addition, forestry policy in

* Corresponding author.
E-mail addresses: dafirebanks@gmail.com (D. Firebanks-Quevedo), jordi@madeingame.cat (J. Planas), kathleen@veritree.com (K. Buckingham), rene.zamora@wri.org (R. Zamora-Cristales).

https://doi.org/10.1016/j.forpol.2021.102624
Received 8 February 2021; Received in revised form 18 August 2021; Accepted 8 October 2021
Available online 11 November 2021

regions like the European Union is "fragmented" and "weakly institutionalized", manifested in overlapping and conflicting objectives regarding forests and forest ecosystem services (Elomina and Pülzl, 2021). Since policy analysis is a complex process that involves searching for and retrieving specific information from large amounts of documents, we believe that it would benefit from automation and the recent advances in artificial intelligence technologies. In fact, we believe that such innovations have the potential to benefit both the policy analysis and policy making processes.

For laws to become effective, regulations must be written to operationalize legal mandates. These regulations usually depend on ministries and government agencies, which have some flexibility to modify them if required, without having to change the law. This offers a great opportunity to identify bottlenecks that affect the effectiveness of policy instruments in supporting restoration. It is in this space where Natural Language Processing (NLP) can provide positive results. NLP refers to a broad collection of techniques that aim at the automated processing of human-generated texts. NLP can play a critical role by enabling policy analysts to rapidly identify large amounts of strategic information, prioritize deep analysis in areas where policies may complement each other, and find other areas where policies may conflict in the nexus of economic development and environmental sustainability.

Currently, government officials are working at full capacity with limited resources, which translates into limited time for analysis and improvement. A common response to this issue in Latin America, for example, is the outsourcing of policy analysis to consultants with the notable exception of Chile, which employs policy analysts in its Office of Agrarian Studies and Policies. However, even with this office dedicated to policy analysis, reports on agricultural programs lag implementation by an average of 3 years (Oficina de Estudios y Políticas Agrarias, 2021). An alternative to overcome this challenge that has been utilized routinely in health policy is the reliance on commissioned rapid reviews, which provide policymakers with relevant summaries of research and take an average of 5–12 weeks to complete (Tricco et al., 2017). Furthermore, most of the operational legal initiatives are tied to annual budgets which implies that when policy analysis is performed at this level of granularity, it can quickly become outdated. Automation of the most time-consuming activities using NLP may help with both adaptive management and keeping an updated stream of information for timely analysis.

Climate change has elevated the importance of Nature-based Solutions (NbS) for adaptation and mitigation. In particular, forest and landscape restoration presents opportunities to increase resilience, support food and water security, protect biodiversity, and foster livelihood development. As 2021 saw the launch of the United Nations Decade on Ecosystem Restoration, it highlighted the need to prepare for success over the decade and to understand what public economic and financial incentives exist to support sustainable forest and landscape restoration. To date, Initiative 20 × 20, a coalition of 18 Latin American countries, has committed to placing 50 million hectares under restoration and conservation by 2030 (Initiative 20x20, 2020). A significant percentage of land will be restored into agroforestry, reforestation, asssited natural regeneration and sustainably managed secondary forests. Understanding the incentives within public policy instruments in these countries that turn those commitments into action, though, is very labor-intensive, requiring decision-makers to read and analyze thousands of pages of documents that span multiple sectors, ministries, and scales that lie outside of their areas of expertise. In order to highlight incentives across these areas, the paper presents a case study of policies across Mexico, El Salvador, Chile, Peru, and Guatemala, with focus on Chile and El Salvador. These countries were part of a Restoration Policy Accelerator Program to accelerate solutions to bottlenecks on incentives for NbS (WRI, 2021).

The case study uses NLP to mine policy documents and to classify incentive instruments. While the case study focuses on forest and landscape restoration in particular, the application of NLP to policy can be transferred to other forestry and land use applications. The models identified most of the relevant policies in a matter of minutes and performed well at identifying the relative frequency of policy instruments. Both the models and human analysts found that in the five Latin American countries assessed, direct payments, credits and fines were the primary policy instruments used; while tax deductions, supplies and technical assistance were infrequently used. Moreover, policy documents tended to describe economic incentives for restoration in vague and intangible terms. Using a standardized approach to generate incentives data could provide an evidence-based and transparent system to find complementarity among policies, help remove barriers for implementers and policymakers, support the need for clearer and more straight-forward terminology and enable a more informed decision-making process.

1.1. The opportunity for Machine Learning in policy analysis

Machine Learning (ML) is "a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy" (IBM 2020). Natural Language Processing (NLP) uses rule-based and machine learning techniques to process and analyze large amounts of text by algorithmically identifying patterns in a subset of either labeled or unlabeled data to enable the filtering and extraction of information from documents (Liddy, 2001). Labeled datasets are collections of data that a human analyst has annotated as whether they belong to a certain category or not. For example, in this paper individual sentences were labeled as referring to economic incentives. This collection of sentences labeled as being incentives together with a collection of sentences labeled as non-incentives were used to train the ML models which learn the differences between the two types of sentences. Later, the model is able to classify unlabeled (non-annotated) sentences into the predefined categories (incentives and non-incentives).

Machine Learning techniques have been previously used for policy analysis with the aim of impact assessment. Rana and Miller (2019) have highlighted the opportunities for integrating ML into forestry and conservation practices. By relating social-ecological system variables with long-term ecological outcomes they identify key factors conducive to policy success and variability in outcome pathways. But automation of legal text understanding is also important for successful forestry policy analysis at scale, which is where NLP techniques should be deployed. While the case study illustrated in this paper focuses on restoration policy analysis, as previously mentioned, the tools can be applied to forestry policy and land use policy more broadly. After all, some of the main barriers to restoration policy analysis are not unique to the field. An over-abundance of dense documents and time constraints are characteristic of many projects that have warranted NLP interventions. Two sectors — healthcare and legal research — are unsurprising pioneers in using NLP technologies. Both fields routinely deal with high volumes of text data, are under high time pressure to deliver, and often involve decisions which can have huge ramifications.

Firstly, in healthcare, NLP technology is routinely used for a wide variety of tasks including clinical decision support (CDS) (Kohn et al., 2014), patient record analysis (FitzHenry et al., 2013), and patient risk evaluation (Bedi et al., 2015). Oncology, for example, is a rapidly changing field, so health care providers must constantly educate themselves about new studies and clinical end points. Since the US healthcare system transitioned to electronic health records (EHRs), there has been an abundance of electronic data. What remains unchanged is the lack of time, especially in a field which deals with matters of life and death. NLP has been seen to expedite the collection of relevant lung cancer patient outcomes and contribute to an evidence-based "learning healthcare system." For example, models can extract relevant case outcomes from radiologic reports in 10 minutes, a task which the investigators estimate would take a human curator 6 months (Kehl et al., 2019). Basically, NLP allows analysts to identify and summarize data that can be used for

evidence-based decision-making.

Secondly, legal research products utilizing NLP have also emerged in recent years (Haney, 2020). One legal product, Smart Code, developed by Bloomberg, extracts relevant paragraphs from laws that include certain statutes or rules. This application uses some ML algorithms to classify legal documents. Its performance is comparable to fully manually annotated systems when the queries are relatively simple but underperforms with increasing query complexity (Zarin, 2017).

While policy analysis in the forestry domain is in its infancy (Taylor et al., 2020; Brandt, 2019) many lessons from these sectors can be applied to the policy field. When comparing the healthcare and legal systems, the difference in the raw data greatly influences the NLP methodologies or 'pipelines' of the two use cases. While humans can understand unstructured information, computers need to be trained to understand unstructured data. The healthcare system, for example, is increasingly adopting the use of electronic records. This provides a structured and relatively uniform database of raw data for using NLP techniques. In the legal system, however, long and complex documents in different electronic formats are commonplace. Healthcare electronic records use the same structured forms for millions of documents, but legal documents have unstructured prose, tenses, and formulations for the same meaning. Thus, there is a need to preprocess policy data using human annotation, through the development of training data. Due to the complexity of legal texts, most NLP approaches rely on human annotation. For this reason, computers still cannot mimic the expertise of a legal analyst.

In a seminal paper of NLP applied to forestry policy analysis, Cheng et al. (2018) developed an application to improve document search and document synthesis. They used the NLP techniques available at the time which relied on statistical analysis of word-word co-occurrence. However, transformer models (Vaswani et al., 2017) were introduced in the NLP landscape around the same time, radically changing the way of solving many NLP tasks. Transformers, which will be described in more detail in Section 2.3, currently represent the state-of-the-art in NLP techniques because they create a better representation of the semantics of each word by taking its context into account, something that previous models were not able to capture. Yet even with transformer models it is almost impossible to reach the level of performance of human specialists in policy analysis. However, there is scope to use cutting-edge NLP technologies to help policy analysts beyond the power of search engines like Google (Wiedemann et al., 2019). For instance, NLP methods that do not require human annotation (unsupervised methods) have seen increasing usage in policy analysis tasks such as identifying shifts in policy agenda (Greene and Cross, 2015) and understanding topical focuses of government petitions (Hagen et al., 2015). Moreover, NLP presents significant opportunities for speeding up the process of policy analysis compared with human analysis. Humans would not be able to cover such a wide range of documents in a limited time (Grimmer and Stewart, 2013).

1.2. Restoration policy analysis

Policy analysis in forestry and restoration is essential in order to understand whether the costs outweigh the benefits — in short, whether there are financial and economic incentives for landowners or landholders to restore land. However, a key challenge for restoration policy analysis is the term itself. While forestry may be a relatively well-defined sector, Forest and Landscape Restoration (FLR) is a nebulous term and each country has different interpretations based on their mitigation/adaptation priorities. Although FLR has been in wide use, at least since the onset of the Bonn Challenge[1], the broad term invites multiple definitions and interpretations of what is and is not included under the scope. Moreover, new terms such as Nature-based Solutions (NbS) are increasingly being used and have overlapping mandates with FLR. FLR is cross-cutting across ministries, land, and mandates. Understanding clearly what financial and economic incentives exist in policy under the ministries or mandates of Forestry, Agriculture, Mining, Land and Environment departments is challenging.

Policy is also a broad term that invites obfuscation. Essentially, policy instruments are used as tools of governance (Howlett and Ramesh, 1993). When legislation is written, governments choose among a range of instruments for the government to meet their goals at the lowest cost (Richards, 2000). The policy analysis targeted in this study deals with financial and economic schemes employed by national and sub-national governments that incentivize non-state actors to engage in restoration activities. However, states can deploy a wide variety of methods to create financial and economic incentives. In this study, we refer to these methods used by governments to implement policy as 'policy instruments' and have organized them using the following classifications:

- Direct payment: A direct cash payment given to program participants in exchange for an activity that restores an ecosystem service such as water, carbon or soil. t
- Credit: A loan which finances restoration activities, or a form of insurance which guarantees that a debt will be paid if financial difficulties arise
- Tax deduction: The reduction of tax liability in exchange for the provision of ecosystem services
- Technical assistance: The provision of technical training or access to expert assistance in the realm of forestry, agriculture, or agroforestry
- Supplies: Material support in the form of tools, plants, seeds, equipment, or other infrastructure
- Fine: A penalty requiring payment to the government in order to compensate for environmental damage

In collaboration with experts in forestry policy, we choose these classifications based on our knowledge of existing restoration incentive programs in Latin America and similar classification systems (OECD, 2017). In recent years, countries like Guatemala have developed high-profile direct payment for restoration actions like PROBOSQUE and PINPEP that have received scholarly attention (vonHedemann, 2020). Many other Latin American countries also have major direct payment programs which act as key forest restoration incentive schemes — from Mexico's Programa Apoyos para el Desarrollo Forestal Sustentable (Comisión Nacional Forestal, 2020) to Chile's Native Forest Law (Ministerio de Agricultura de Chile, 2008). Other financial instruments include credits and tax deductions, the latter of which were among the most popular instruments for restoration incentives in Latin America in the 1980s and 1990s. However, tax deductions have since fallen by the wayside (in all study countries besides Peru) due to the lack of efficacy and weak monitoring of performance and impacts (Jaramillo and Kelly, 1999). With tax instruments, it was often the case that program beneficiaries would plant a tree in order to receive the tax benefit but not care for or maintain the tree beyond that point — which is not the case with other mechanisms, in which the provision of benefits is often contingent on tree survival or other environmental factors. Other instruments like technical assistance and supplies are not financial in nature but still provide economic benefit through in-kind support. For example, Mexico's Sembrando Vida program provides a monthly payment to farmers along with specialized training and plant materials to beneficiaries managing agroforestry systems (Diario Oficial de la Federación, 2020). Fines also function as powerful incentive tools, while they can disincentivize activities that are counteractive to restoration. For example, in El Salvador, some municipalities require the payment of a fine for each tree that a citizen cuts (República de El Salvador, 2016). Fines such as these that are imposed due to land use changes for ecologically harmful activities serve as disincentives to degradation and, therefore,

[1] The Bonn Challenge was initiated in 2011 by IUCN to restore 150 million hectares of land by 2020.

are incentives for conservation.

Through Initiative 20x20, the Latin American region has prioritized FLR as a key strategy to mitigate Greenhouse Gas (GHG) emissions and adapt to climate change effects. Different countries have designed and implemented restoration strategies that include policy instruments to support restoration. However, policies on FLR often lack clear strategies to enable public investments in restoration and remove barriers to facilitate private finance. This is in part due to the lack of complementarity of agriculture and environmental policies and the need for immediate solutions instead of long-term, sustainable strategies (Kissinger et al., 2015). When assessing policies across ministries, incentives are assessed at the ecosystem or functional level. Opportunities for agricultural crops, forestry plantations, or mining land rehabilitation are assessed separately. Therefore, a more holistic approach is needed. In this paper we present a way to automate the most time-consuming activities in policy analysis which may help in overcoming this problem.

2. Methodology

Our primary objectives for this study are to set up a software tool to (1) identify whether a sentence in a policy document mentions incentives, and (2) identify the incentive instrument associated with each incentive. To this end, we frame the problem as a *text classification* problem, where the NLP model learns to classify text into categories. The workflow is composed of five main steps (Fig. 1). These include:

- Scraping, to automatically retrieve policy documents from the internet
- Preprocessing, in order to transform the text written in the documents to computer readable data
- Data labeling, combining both human and computer efforts to constantly increase the size of our labeled dataset
- Classification, which is the end goal of the project, using the labeled data to fine-tune a model
- Evaluation of model performance with standard statistical metrics such as the F1 score

It is important to note that the core of our methodology is a sequential classification system in which we first develop a binary classifier to assess whether a sentence refers to an incentive or not. Then, sentences which are labeled as being incentives are processed by a second (multiclass) classifier to assign them to a specific policy instrument. All coding scripts and notebooks have been written in Python 3, and are publicly available in the GitHub repository of the project.

2.1. Data sourcing and processing

In total we sourced around 1375 documents from the five countries of our case study, and we referenced the specific numbers and sources in the appendix Table A3.

2.1.1. Repository of official documents from El Salvador

This project was discussed with officials from the Ministry of Environment of El Salvador (MARN). El Salvador officials provided a collection of documents that they found to be relevant to land restoration policies in their country so that we would be able to compare the automated data-gathering method (scraping) with a manual process of searching specific databases.

2.1.2. Scraping

The process of extracting policy documents started by automatically scraping official websites using a collection of custom algorithms and search terms relevant to landscape restoration. For this purpose, the Python text-mining Scrapy library (Kouzis-Loukas, 2016) was utilized to download and store documents in an Amazon Web Services (AWS) S3 bucket for further processing. This, along with the document metadata collected during scraping, allowed for the creation of a comprehensive database of traceable policy documents for use further down the pipeline.

To retrieve policy documents from Chile, we crawled the website of the Library of the National Congress (Ley Chile, 2020). The policy documents from El Salvador were obtained by scraping the website of the Supreme Court of El Salvador (Centro de Documentación Judicial, 2020).

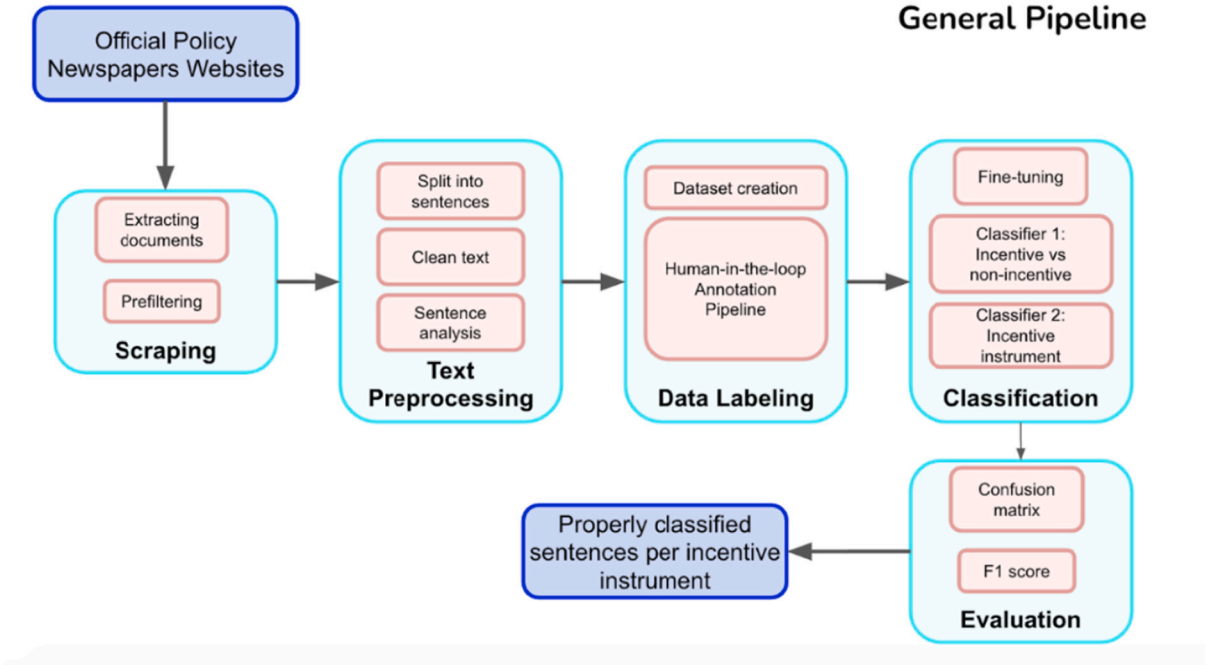

Fig. 1. Data pipeline from the source in official websites to the retrieval of incentives and classification of incentive instrument.

2.1.3. Pre-processing

Generally, a policy document file can be a Word document (.docx) or PDF (.pdf) and is in text or image format. If the document file is not an image of a scanned document, the conversion to computer-readable text is straightforward, as was the case with the policies from Chile. However, policies from El Salvador were PDF files containing scanned images from the original, paper-printed documents. In order to transform text within an image to computer-readable text, optical character recognition (OCR) tools must be used. The transformation of PDF scanned documents was performed using the OCR engine of Tesseract OCR (Smith, 2007). There are two main challenges in the pre-processing step. The first challenge is to partially keep the structure of the documents in their main sections and the second is to retrieve the sentences in the document. Keeping the structure of the document is important for optimizing information retrieval downstream. To retrieve the sentences is paramount, as sentences will be the basic unit of the NLP modeling. Both were performed using custom scripts, one for the Chilean documents and one for the documents from El Salvador.

2.2. Data labeling

In machine learning, computer models learn from examples that can be labeled or unlabeled. A labeled data point is a sentence associated with a particular label. In this study, each sentence is given two labels: one to tag it as being an incentive or not and the other to tag it with a specific incentive instrument. In Table 1, there is an example of some sentences that were classified as referring to incentives or not and as referring to credit instruments or not. The process of looking at data and learning to do a task is called *training*.

State-of-the-art NLP models are available mostly *pre-trained*, meaning that they have been usually trained with unlabeled text from huge corpora such as the entirety of the English Wikipedia (2500 M words) or even a large set of documents and websites on the internet. Pre-trained models are very powerful and can be used off-the-shelf for many tasks, but in order to adapt them for improved performance in a specific domain they must be fine-tuned. Fine-tuning a pre-trained model for classifying FLR incentives involves training it on additional policy-related labeled data.

Using the right training data is essential for good model performance. However, labeling training data is difficult, time consuming, expensive, and requires the effort of human annotators with specific domain experience and training. Therefore, we devoted special attention to the building of training datasets by setting up an assisted labeling technique and comparing the performance of different annotators.

Table 1
Example of two sentences that are incentives, one of them contains credit as the instrument.

Sentence	Does it contain an incentive?	Is credit the incentive instrument?
El Servicio podrá conceder los siguientes préstamos reajustables en Unidades de Fomento a sus afiliados, cuando sus recursos lo permitan.*	Yes	Yes
Propiciar el desarrollo de los asentamientos poblados de menor tamaño, favoreciendo la provisión de bienes y servicios de calidad en aquellos de mayor densidad, mejorando el uso y la ocupación regular del territorio.**	No	No

* The Service can grant the following re-adjustable loans in Development Units to its members, when resources allow.

** Promote the development of smaller towns, favoring the provision of quality goods and services in those with greater density, improving the use of and regular activity in the territory.

In this paper, we use two datasets for model training, constructed using two different methodologies. The first dataset was obtained by pure human analysis, which we will refer to as our *hand-picked dataset*. The second dataset was built by a combination of automatic computer processing and human curation, denoted the *human-in-the-loop (HITL) dataset*.

2.2.1. Hand-picked dataset

The initial batch of training data was created by a policy analyst who read over 3000 pages of policy documents — all relating to financial and economic incentives and restoration — from the governments of Chile, El Salvador, Guatemala, Mexico, and Peru and manually labeled the textual excerpts using the categories listed in Table 2.

Of the more than 3000 pages of policy that the policy analyst examined, a 48-page subset (<2%) was reviewed by a Latin American restoration policy expert to ensure the accuracy of the labeling process.

It is important to note that we did not use the labels in the *Land use type* category in the modeling phase because as a minimal viable product (MVP), we centered this study on the incentives and their typology independently of the target land type. As per the *Impact on restoration* category, we decided to categorize all sentences in this dataset as *Incentives*. Intentionality or the fact that a certain incentive is a disincentive in some contexts are complex issues that require further treatment. A disincentive is always an incentive in a competing or conflicting area. For example, facilitating credit for the mining industry might be a disincentive for land restoration. Identifying disincentives for restoration as the topic focus could be built out later.

2.2.2. Human-in-the-loop dataset

Human-in-the-loop annotation describes an interactive process where an ML system gives some initial predictions on the sentence labels and the human expert corrects them if they are wrong (Fig. 2) (Klie et al., 2020). It is a form of assisted labeling that speeds up the labeling task as the human analyst is not given a set of documents to read, but rather a preselected list of sentences with a label to be confirmed.

For each policy instrument, we manually choose five sentences from the hand-picked dataset that were representative of the instrument. We then modified them by trimming words that, being out of the scope of the policy instrument, could mislead the model. With the 30 sentence queries, we used S-BERT (Reimers and Gurevych, 2019), a pre-trained model described in Section 2.3, to search similar sentences in a database containing more than 116,000 sentences from 4700 policy documents from Chile and 348 policy documents from El Salvador. The output of the ML system is the list of the original sentences in the sentence database ranked by cosine similarity to the query.

We took only the top 1000 sentences from the output of each query, as they have the highest sentence similarity with the query. These sentences were given the same label as the query sentences used to fetch them. Finally, we retrieved only the sentences that were common to the five queries for each policy instrument, yielding six lists containing between 450 and 600 sentences which were then sent to multiple human labelers (or raters) to confirm and curate the automatic labeling.

2.2.3. Experimental setup

Labeling data is not always an easy task. Particularly when it comes

Table 2
Categories used to label the incentive sentences in the hand-picked dataset.

Policy Instrument	Land use type	Impact on restoration
Direct payment	Forest	Incentive
Tax deduction	Agriculture (Crop)	Incentive (Intention)
Credit/guarantee	Agriculture (Pasture)	Disincentive
Technical assistance	Mangrove	Disincentive (Intention)
Supplies	Peatlands	
Fines	Grasslands	

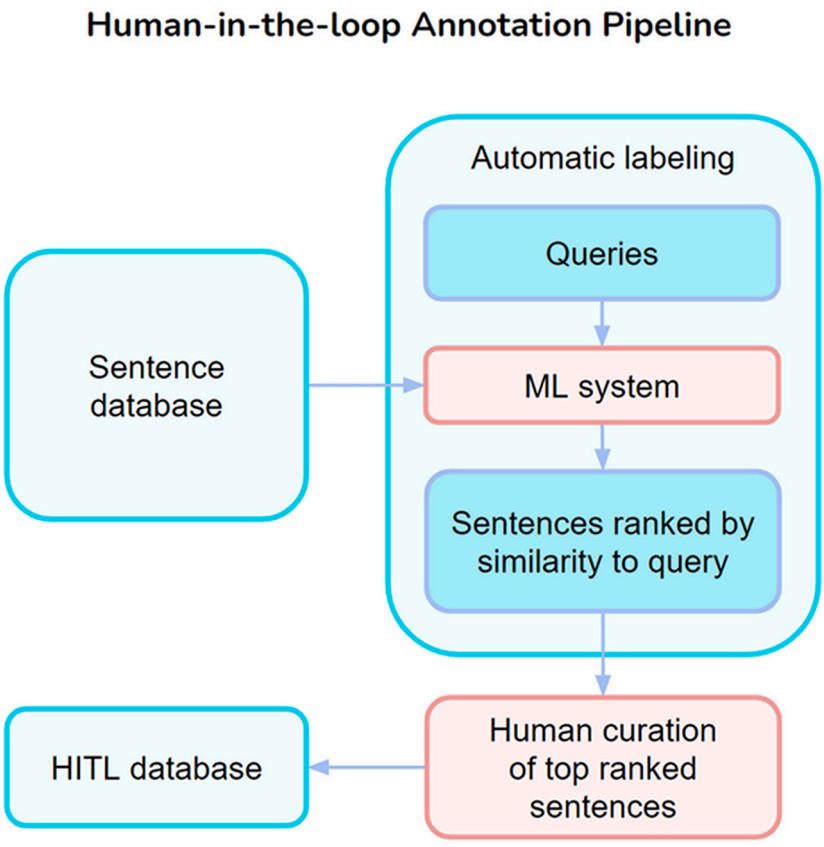

Fig. 2. Assisted labeling pipeline - where human and machine interact to create a set of labeled sentences.

to natural language processing and more specifically to policy analysis, labeling involves a certain degree of subjectivity. Furthermore, labeling is the main contribution of stakeholders in NLP projects. For example, if an organization in the forestry domain wants to use the tools developed in this paper, its main role would be to supply the model with good quality labeled data. Thus, we have defined an experimental setup to explore whether there exists some labeling strategy that leads to optimal results from the NLP model.

We analyzed three factors: i) Inter-rater differences among labelers, ii) merging the datasets coming from different labelers, and iii) combining the HITL dataset with the hand-picked dataset.

In order to observe the impact of inter-rater differences among human labelers on model performance, the HITL dataset was given to 3 different labelers with different degrees of expertise and language skills, with no defined criteria to evaluate besides the label meaning. After manual curation, the labelers confirmed on average 48% of labels as being incentives and 28% of labels as belonging to some of the policy instruments. These percentages are a proxy of the model performance before fine-tuning. To broadly assess the level of labeling agreement among raters, inter-rater reliability was calculated using Fleiss's Kappa — a measure of the agreement between more than two raters that naturally controls for chance. The consensus is that values between 0.40 and 0.75 represent a fair to good level of agreement beyond chance alone (Green, 1997). For our scenario, the raters' judgement of whether the sentence contained an incentive yielded a 0.497, whereas their judgement of the incentive instrument yielded a 0.565. In the appendix, we provide an in-depth analysis of the HITL datasets curated by the three labelers.

We also evaluated the impact of merging the HITL datasets from the different raters on the overall classification model performance. We performed three different merges of the dataset: Merge1 in which we retrieved the label of a sentence if at least one of the curators labeled it as positive; Merge2 in which we retrieved the label of a sentence if at least two of the curators labeled it as positive; and Merge3 in which we only retrieved sentences with full agreement among the three raters.

Finally, we also combined the hand-picked dataset and the HITL dataset in different ways. For the binary classification we tried two different configurations which were:

- HITL/train&test: Fine-tuning the model with 80% of the HITL dataset (training set) and evaluating the classification power with 20% of the HITL dataset (testing set).
- HITL+hand-picked/train&test: Merging the two datasets, using 80% of the merge for fine-tuning the model (training set) and evaluating the classification power with 20% of the merge (testing set).

For the multiclass classification, we tried a third combination.

- HITL/train&hand-picked/test: Fine-tuning the model with 100% of the HITL dataset (training set) and evaluating the classification power with 100% of the hand-picked dataset (testing set).

In total, we explored 12 experimental setups for the binary classification and 18 for the multiclass classification.

2.3. Classification

Once we had training data, we proceeded with the original goal — text classification. Recall that our main task was divided into two subtasks: Determine whether a sentence mentions incentives and identify the incentive instrument if applicable (Fig. 3).

State-of-the-art NLP relies on a powerful and efficient neural network architecture called the Transformer (Vaswani et al., 2017). The Transformer is based solely on the attention mechanism, getting rid of any

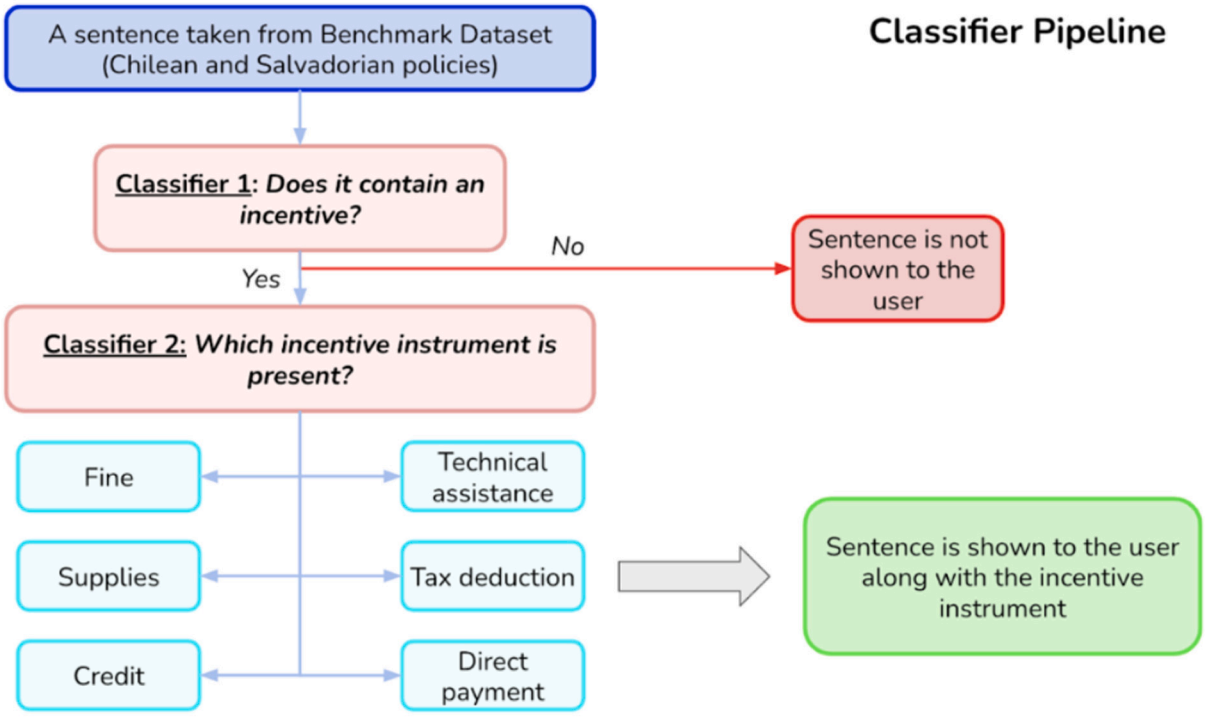

Fig. 3. Classification pipeline for 1 policy sentence that every filtered sentence from a document goes through.

recurrence or convolution operations present in other models, thus being parallelizable and requiring significantly less time to train. The current best performing models for several NLP tasks like text classification use some variation of a Transformer, largely because self-attention is a theoretically and practically stronger technique to process text than rule-based and traditional sequential models are (Minaee et al., 2021). Language models like BERT–*Bidirectional Encoder Representations from Transformers*– (Devlin et al., 2019) and OpenAI's GPT (Brown et al., 2020) leverage this architecture, making up a large part of the modern NLP landscape by providing an off-the-shelf, powerful way to create state-of-the-art models for a wide range of tasks and domains without needing to modify the original architecture.

In this paper, the model used for text analysis was Sentence-BERT (S-BERT), a pre-trained model that is key in both our assisted labeling and classification pipelines. S-BERT is a model trained on two NLI datasets (Bowman et al., 2015; Williams et al., 2018) for the problem of determining whether two sentences are contradictory, neutral or entailments to each other. To do so, it learns to create a good numerical representation of sentences in some number of dimensions called an *embedding*. We chose this model because it allows us to derive semantically meaningful sentence embeddings that can be compared using similarity measures in a fast and efficient manner, contrary to its antecedent BERT.

After embedding the labeled sentences, we tested two classification approaches for each sub-task: classify a sentence based on its similarity with the label and train a classifier to predict a sentence label based on its embedding.

2.3.1. S-BERT for representation and classification

We know that S-BERT is good at creating sentence embeddings and that with enough fine-tuning it can create embeddings that put similarly labeled sentences close in distance to each other. Unfortunately, the model is not good at creating an embedding for the label itself (i.e. "direct payment") as it is not a sentence but rather a set of one or two words. This is an issue if we are ultimately trying to measure how close the label and a sentence embeddings are to each other. Therefore, following a few-shot learning approach (Davidson, 2020), we added an extra step that mitigates this problem by using a word-based embedding model and a translation from word embeddings to sentence embeddings. As a result, we classify sentences by calculating the distance between the sentence embedding and the embeddings for all labels, choosing the label that is the closest to the original sentence.

2.3.2. S-BERT for representation with a separate classifier

Embedding distance between texts is one of many ways to determine the label of a given sentence, and we decided to explore other techniques that have been widely used in the classification world. For our study, we decided to use Support Vector Machines (SVM) (Gunn, 1998), Random Forests (RF) (Breiman, 2001) and Gradient Boosting Trees (GB) (Friedman, 2001). Specifically, we chose to use the embeddings created by S-BERT and feed them as input features for another classifier to determine the sentence labels. Unlike S-BERT, the other classifiers do not learn to create a better representation of the sentence but rather find the best mapping between a sentence representation and its label.

2.4. Evaluation

To evaluate the myriad of different model configurations, we defined an evaluation process suitable to our data and problem. We used a confusion matrix, which provides information on the type and amount of classification errors (false positives, false negatives) a classifier makes, and the *F1-score* of the classifier as a single measure to assess its quality. The F1-score—or harmonic mean between precision and recall—was chosen because it balances the proportion of correctly labeled sentences from all sentences *predicted as belonging* to a category, and the proportion of correctly labeled sentences from all the sentences that *truly belong* to that category. This allows us to take in account the class imbalance in the data when assessing the classifier performance.

3. Results and discussion

3.1. Manual analysis vs automated analysis

While our methodology can be used to analyze forestry policy from any country at the regional or local level, we focus on applying it to the 5

Latin American countries from our case study. In this section, we make a comparison between a manual and automated approach for different steps of our methodology, in order to test the potential of a computer-assisted system in the identification of land restoration incentives.

3.2. Data sourcing

Assessing the soundness of the automatic approach was done by comparing the results of the scraping with the pool of documents used by experts (Table 3). In their practice, members from El Salvador's Ministerio del Medio Ambiente y Recursos Naturales (MARN) provided documents from the following areas:

- Environmental Laws
- General Laws
- International Treaties
- Technical documents and norms
- Draft legislation (Anteproyectos)

The automatic approach scraped the official legislation in the manner described in Section 2.1.2. The automated scraping returned almost six times more results than the manually collected documents, bringing in two additional categories:

- Documents from different ministries relating to industry and agriculture, which may have an impact on restoration. They may contain instruments promoting the respective sector, thus creating disincentives for land restoration.
- Local-level documents issued by municipal authorities outlining environmental laws compliance and use of terrain rules within their remit. Beyond giving the status of environmental regulation within these areas, capturing and analyzing such documents may highlight successful local initiatives to serve as examples.

The feedback from MARN was that these categories were relevant and awareness of them would be beneficial. The scraping also had the following advantages:

- Returns traceable documents; avoids scraping duplicates and incomplete documents.
- Catalogues metadata, including authors, date, title and description of the documents.

Table 3
Comparison of the document retrieval capacity of a panel of human experts and an automated process.

	Human expert*	Automated process**
Environmental laws	Yes	Yes
General laws	Yes	Depending on search parameters
International treaties	Yes	Depending on search parameters
Technical documents and norms	Yes	No***
Draft Legislation	Yes	No****
Sector incentive documents	No	Yes
Local-level documents	No	Yes
Number of documents	64	367

* This process was performed by officials in the Ministry of Environment and Natural Resources from El Salvador.
** The automated process was done by web page scraping as described in Section 2.1.2.
*** Technical documents and norms can be automatically retrieved by scraping designated sites.
**** Draft legislation is not publicly available before official publication.

- Option to work with valid legislation only (legislation in force), regardless of the year of publication.
- Flexibility: it has the scope to expand the reach and include other types of documents.
- Speed: thousands of documents can be retrieved in a matter of minutes.

3.3. Overview of restoration incentives in Latin America

In manually reviewing policy documents, we found that the 5 study countries used similar policy instruments to incentivize restoration activity. The most common instruments used to promote landscape restoration were direct payments, credits, and fines. Most of the study countries had at least one major incentive program that provided direct payments for ecosystem services to program participants. Guatemala, for example, has two active incentive programs that reward participants with direct payments: PROBOSQUE for landowners with legal titling and PINPEP for small landholders with non-legal tenure (Congreso de la República de Guatemala, 2010, 2015). Aside from direct payments, in-kind support is also a common form of economic incentive. Sembrando Vida, for example, is a major program in Mexico that promotes human wellness for farmers and incentivizes the establishment of agroforestry systems. In addition to granting direct payments of $5000 MX pesos/month (US $250/month), Sembrando Vida supplies program beneficiaries with plants, tools, a yearlong training program, and access to specialized technicians (Diario Oficial de la Federación, 2020). Examples of other financial incentive instruments such as tax deductions and credits can be found across the study countries, though in fewer instances than direct payments.

Through a closer look at El Salvador and Chile, we can see that a manual review of policy documents reflected the prevalence of direct payment schemes to incentivize restoration in both countries (Fig. 4). However, the hand-picked data set did not reflect other trends which we know to be true, such as the prevalence of fines in El Salvador. To correct this unbalanced data and compare differences between manual curation and automatic curation, we used the HITL dataset for large-scale quantitative analysis. This dataset was not built for the purpose of quantifying policy instruments, but it can be taken as a proxy of a large-scale search on official documents for the two countries. The differences between the countries are apparent (Fig. 5). While Chile employs a range of instruments, El Salvador's policies are mainly prohibitive, as fines feature extensively in the local-level policies. Credits and direct payments do not appear to be employed in El Salvador to the same extent as in Chile.

3.4. Effect of labeling strategy on model performance

The influence of the labeling strategy was evaluated on both the binary and multiclass classification. The classification was done using S-BERT for representation and classification as explained in Section 2.3.1.

The maximum F1 differences among datasets labeled by different annotators are 0.25 for the HITL dataset alone and 0.14 for the merged dataset, with average F1 differences among raters being 0.17 and 0.09, respectively. The fact that the differences among raters decrease in the merged dataset comparison is consistent with the fact that merging each HITL dataset with the hand-picked dataset introduces a buffering effect on the labelers' influence.

The average F1 for all the conditions analyzed with only the HITL dataset was 0.72, while the average score for the merging of HITL and hand-picked datasets was 0.75, just a slight increase that can be explained by the hand-picked dataset containing only sentences labeled as incentives (Table 4). Finally, the average of all F1 scores for the labelers was 0.73 while the average F1 scores of all merges was 0.75.

We also evaluated how the labeling strategy influenced the F1 score in multiclass classification. Here, we were able to use the hand-picked dataset just for evaluation, so we could also test the consistency of

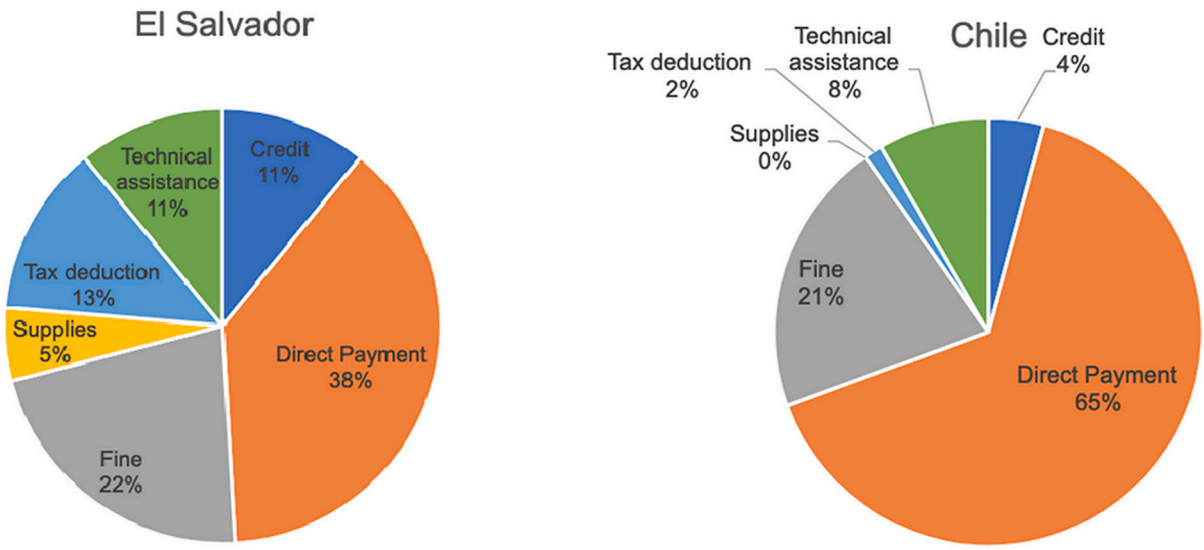

Fig. 4. Distribution of sentences refering to incentive instruments in policy documents from El Salvador, on the left, and Chile on the right. Data obtained after manually processing a collection of 5 documents per country.

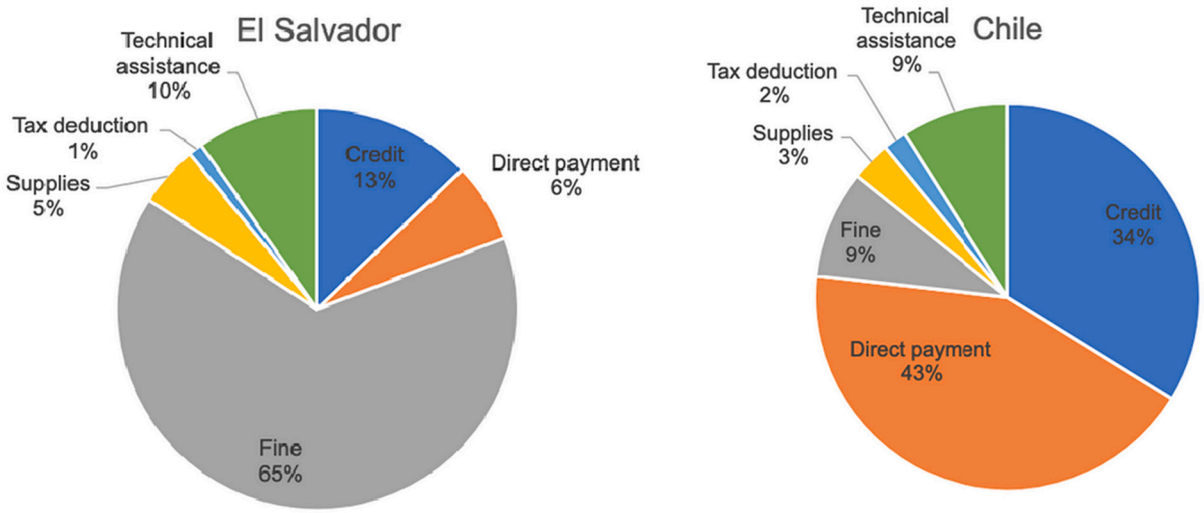

Fig. 5. Distribution of sentences refering to incentive instruments in policy documents from El Salvador, on the left, and Chile on the right. Data obtained after automatically processing 4700 documents from Chile and 360 from El Salvador.

Table 4
F1-score of S-BERT for binary classification per dataset type. Effect of different labelling strategies on incentive/not incentive classification.

		Train/test combination	
		HITL/train&test	HITL + hand-picked/train&test
Labeling type	Labeler1	0.56	0.77
	Labeler2	0.81	0.81
	Labeler3	0.73	0.67
	Merge1	0.78	0.68
	Merge2	0.83	0.76
	Merge3	0.60	0.82

Table 5
F1-score of S-BERT for multiclass classification per dataset type. Effect of different labelling strategies on incentive instrument classification.

		Train/test combination		
		HITL/ train&test	HITL/train&hand-picked/test	HITL + hand-picked/train&test
Labeling type	Labeler1	0.68	0.57	0.64
	Labeler2	0.57	0.66	0.82
	Labeler3	0.62	0.65	0.78
	Merge1	0.69	0.66	0.75
	Merge2	0.60	0.67	0.84
	Merge3	0.65	0.59	0.75

each rater's labeling against an external dataset. We observe (Table 5) that the F1 score for labelers 2 and 3 is better when the model is tested with the hand-picked dataset than when tested with a subsample of the corresponding HITL dataset. This might be because the fine-tuning was performed with the full HITL set instead of 80%, but it clearly indicates that the training on the full HITL dataset yields good performance on an independent testing dataset. For Labeler1, the results are not aligned with those from the other two labelers, potentially due to a smaller number of samples labeled by Labeler1 (Appendix).

The averages of the differences among labelers in multiclass classification were 0.07, 0.06 and 0.12 for the three train/test set

configurations.

The average F1 score of labelers by each of the three train/test combinations are 0.62, 0.63, and 0.75 and the average F1 score of merges by each of the three train/test combinations are 0.65, 0.64, and 0.78, respectively.

Overall, these results indicate that the labeling strategy is influential because the variations in model performance might be higher than 0.2 F1 points among different labelers. In this study, Labeler2 was the most consistent labeler in both the binary and in the multiclass classifications. It is important to note that Labeler2 was the most experienced of the three in the forestry domain. Combining the labeling of different labelers provides small but consistent improvements of model performance. We also observed that combining the HITL dataset and the hand-picked dataset consistently yielded the best results in terms of model performance.

3.5. Classification optimization

In the previous section, we used distance between sentence embeddings and a proxy of the label embeddings to classify sentences as explained in Section 2.3.1. However, this is not necessarily the best classification method for all types of text. Therefore, we explored whether classification performance could be improved using different classification tools commonly used in the data science literature as explained in Section 2.3.2.

We decided to try the new classification tools on the Labeler2 and Merge2 datasets and check only the performance after merging the HITL and the hand-picked datasets because we had the best overall performances with these conditions in the previous analysis.

As we can see in Table 6 and Table 7, these classifiers achieve better performance than the embedding distance method. In particular, the F1 for Labeler2's best results increased by 0.13 for the binary classification and 0.10 in the multiclass classification. A very similar trend was observed by the Merge2 approach where F1 increased by 0.18 and 0.08 for binary and multiclass classification, respectively. In this case, Merge2 shows small but consistent F1 scores when used for training the model, confirming that the strategy of combining data labeled by different labelers is worth implementing.

Looking in detail at the accuracy of the model in the confusion matrices shown in Fig. 6, we see that for the binary classification, the presented model has higher precision than recall, and as such is better at predicting non-incentives than incentives. In future work, special attention should be put into the reduction of these false negatives rather than false positives, as the latter can be filtered out at later stages of the text processing.

With regards to the classification of policy instruments, we see in Fig. 7 that for Labeler2, the values are evenly distributed along the diagonal — meaning that the model can predict the true labels in 5 out of 6 instruments with more than 90% accuracy. At the same time, the values out of the diagonal are mainly zeroes and never higher than 0.1, which means that the rate of misclassification is very low. For Merge2, which has an average F1 very close to Labeler2 as shown in Table 7 (0.93 and 0.92, respectively), some of the accuracy values in the diagonal are above 0.95. Even if the values outside the diagonal are mainly zeroes, some of them are slightly above 0.1. In summary, the labeling performance of Labeler2 is more balanced than that of Merge2 and, therefore, better.

Table 6
F1-score of binary classifiers in Merged-datasets. Effect of different classification models on incentive/not incentive classification.

Model	Labeling type	HITL + hand-picked/train&test
SVM	Labeler 2	0.89
	Merge2	0.93
RF	Labeler2	0.90
	Merge2	0.94
LightGBM	Labeler2	0.89
	Merge2	0.94

Table 7
F1-score of multiclass classifiers in Merged-datasets. Effect of different classification models on incentive instrument classification.

Model	Labeling type	HITL + hand-picked/train&test
SVM	Labeler 2	0.92
	Merge2	0.89
RF	Labeler2	0.93
	Merge2	0.92
LightGBM	Labeler2	0.93
	Merge2	0.92

4. Conclusions and future directions

This paper presents a novel use of NLP in the domain of forestry. The case study of FLR highlights the complexities of interdisciplinary challenges inherent in land use policies and land use governance. To expedite a global assessment of incentives within forestry, land-use, and FLR, utilizing Machine Learning and Natural Language Processing is essential. Since forestry departments and governments often do not have resources to continually assess policies and conduct widespread comparative studies, processes such as those outlined in the paper can be utilized to accelerate learning, build an evidence base, and provide a foundation for an assessment of effectiveness. Moreover, even with adequate resources, forest policy is difficult to understand. This methodology can provide a way to quantify traditionally qualitative approaches and bring greater accountability to instrument types.

4.1. Recommendations

As this project is a first step in exploring the uses of NLP in the field of restoration policy, we purposefully selected broad classification categories. However, more specific categorization would likely yield more granular findings in the future. For example, before deciding to focus solely on incentive classification, we defined a landscape classification schema where we divided the label "Forest" into 2 categories, "Protected" and "Non-Protected." Many policy documents, however, do not clearly distinguish between protected and non-protected forests, which led us to use the broader classification of "Forest." With more time and progress on the automated side, making these distinctions could be more feasible in the future. A network analysis, for example, could aid policy analysts in determining which documents implement a specific law that refers to protected forests. Being able to clearly identify relationships among documents would improve the accuracy of the classifications and, in turn, the quality of policy analyses.

Expanding the understanding of incentives to include disincentives and perverse incentives would also be beneficial. Firstly, understanding disincentives to restoration can enhance our understanding of what makes restoration incentive programs work. Disincentives should be interpreted as measures that disincentivize the uptake of a policy goal, so disincentives in the context of this study would be measures that incentivize activities which are counteractive to restoration. For example, tax deductions for exploratory oil drilling may encourage behavior that is destructive to landscapes and therefore counteractive to restoration. Land that could have otherwise been conserved or used in a more ecologically friendly manner would instead be exploited for oil exploration due to this disincentive to restoration. Any benefit from an ecologically harmful activity that outweighs the benefits of restoration, therefore, is considered a disincentive. Secondly, perverse incentives refer to incentive policies that produce unintended consequences. Because perverse incentives require knowledge of policies' outcomes, it is not possible to identify them based on the policy document alone. In the future, a review of policy analyses and program outcomes may allow our analysis of perverse incentives as well.

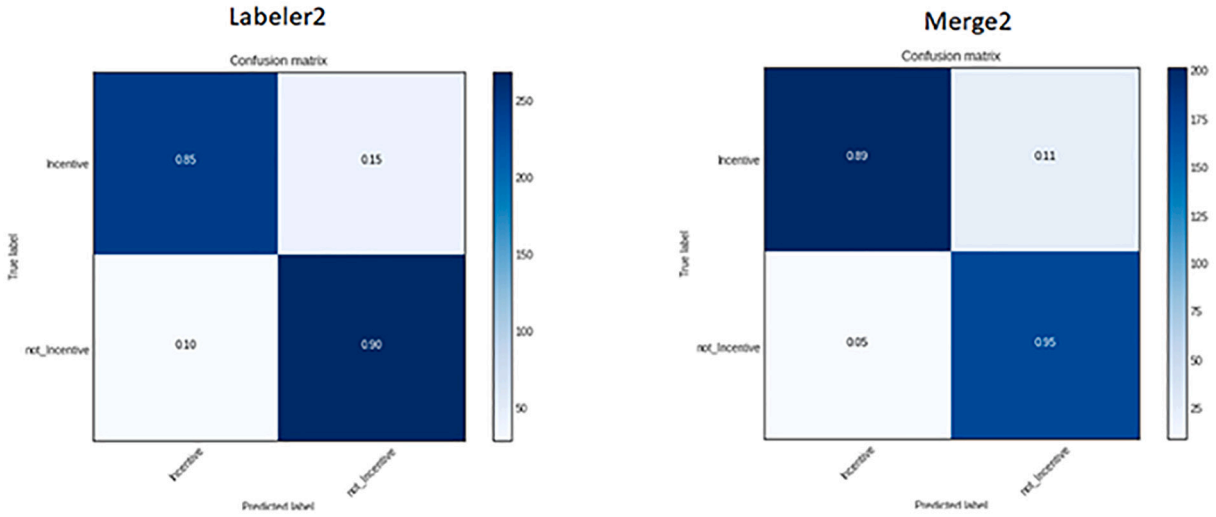

Fig. 6. Evaluation of the random forest binary classifier. Confusion matrix containing accuracy per label corresponding to the Labeler2 dataset on the left and the Merge2 dataset on the right.

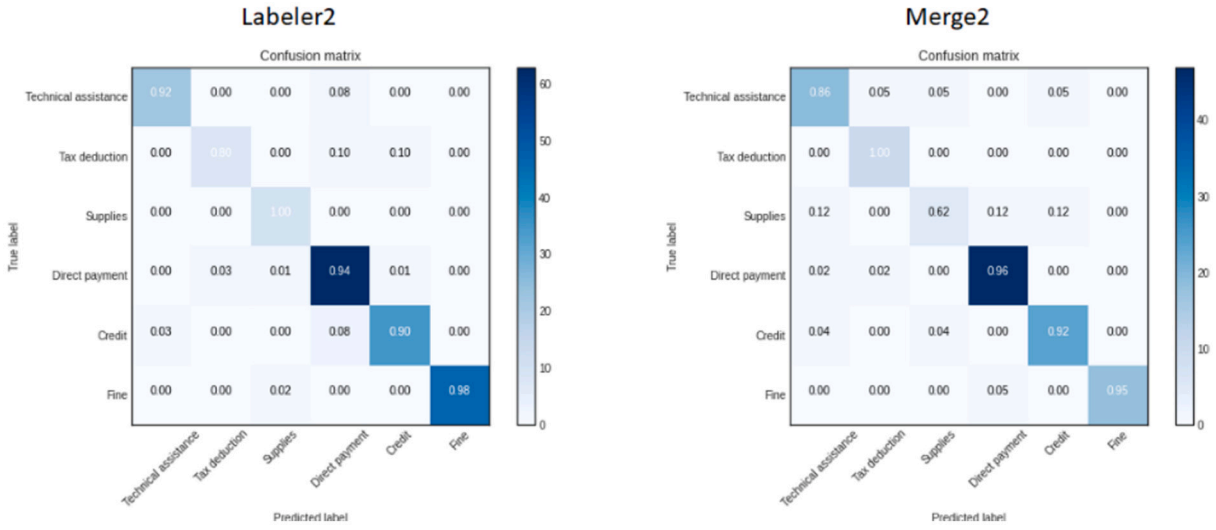

Fig. 7. Evaluation of the random forest classifier for multiclass classification. Confusion matrix containing accuracy per label corresponding to the Labeler2 dataset on the left and the Merge2 dataset on the right.

Finally, our recommendations in the classification realm are twofold. On one hand, we would like to improve the performance of classifiers like Random Forests and Support Vector Machines by tuning model parameters and other common data science optimization procedures. On the other hand, we hope to be able to experiment with more advanced models that have come out recently in the NLP literature, such as GPT-3 (Brown et al., 2020).

4.2. Limitations

Other limitations that we leave for potential future work are:

- The inability of the model to distinguish between intention/plan/general mention. This means that the model may identify an intention of payment to incentivize farmers in a given region, without a concrete program that backs it up (for the time being, we leave this task to the policy analyst). The solution to this problem would also tackle the challenge of determining the currently valid policies and obsolete and superseding documents.
- The model still relies on expert knowledge. We depend on data labeled by policy analysts, which holds true if we were to classify information in a different manner like identifying landscape types in policies. This means investing time and other resources in labeling data - however, until unsupervised methods improve, data labeling is still a worthy investment as it saves reading time in the future, and the knowledge acquired by the model can be built upon later.
- The subjectivity in evaluating incentives. This presents difficulties not only in validating an automated model, but also in understanding incentives in general. Although the demonstrated levels of inter-rater reliability (0.497/0.565) are in the range of what is considered "moderate agreement" (Landis and Koch, 1977) and "beyond chance alone" (Green, 1997), it shows that the agreements of what constitutes an incentive varies greatly across raters. However, the fact that such inconsistency exists in the non-automated approach makes an automated model an acceptable solution, given the time-saving benefits.

5. Conclusion

The introduction of Natural Language Processing techniques to policy analysis is not only important to forest and landscape restoration, but to the policy analysis world in general. It presents a tool that can search for relevant policy instruments across ministries and disciplines, marking the first step in identifying incentives for forestry in an automated fashion. Since forest and landscape restoration involves a mosaic of landscapes across scales with ever-changing regulations, it is difficult for landowners, policy makers, and practitioners to stay up to date with regulations, particularly beyond jurisdictional, land-use, or issue boundaries. The time intensity and resources that are required for reading thousands of documents is prohibitive to conducting large scale policy analysis.

This paper presents a Minimum Viable Product (MVP) that shows great promise in overcoming the time and resource constraints stemming from the analysis of forest and landscape restoration policies across ministries and countries. Several technical tasks were achieved which enabled a contribution to knowledge. Starting with text extraction and scraping of official policy documents in 5 Latin-American countries, we were able to create a labeled sentence dataset from official policy documents. Another key contribution is our Human-in-the-loop annotation pipeline for efficient labelling of new examples and expansion of the training data. We used S-BERT to obtain embeddings that can capture that sentence semantics. Finally, we trained classifiers that can distinguish between incentive and non-incentive sentences, and among policy instrument sentences.

Using an NLP model trained and evaluated on 2 datasets curated by policy experts, the paper presented a case study for policy analysis for El Salvador and Chile. Our best results found that the model could identify incentives in sentences with 85–89% accuracy and distinguish 5 key incentive instruments (direct payment, credit, technical assistance, fines, and supplies) with over 90% accuracy. The model had difficulty identifying 1 instrument (tax deduction), with 80% accuracy since we had fewer data points to both train and evaluate on. Adding to these accuracy levels, our model achieves an F1-score of 93–94% in both identifying an incentive and its instrument among the 6 incentive classes, which means that it is quite remarkable overall despite the imbalance in number of sentences per incentive instrument. Furthermore, our model is designed to constantly improve its performance with more data and feedback from policy experts.

The design of this MVP allows for its extension to policies across multiple countries and languages to enable a uniform, structured approach to policy analysis across countries. Firstly, the speed at which the automated analysis to highlight the existence of incentives for restoration is carried out can save significant time and resources and allow for greater focus to be paid to structured and specific policy analysis. Secondly, although accuracy needs to be increased in some policy instruments, such an approach would enable an understanding of regional and global instruments. Thirdly, this approach allows for creating systematic and quantifiable social data, as it is essential for people to see the benefits from forest and landscape restoration. It could highlight favored instruments in certain regions and identify trends over time. In this way, it could create the foundation for systematic analysis of effectiveness through providing a more quantifiable assessment of policy systems. Trends will be more easily visible regarding whether 'carrots' or 'sticks' are utilized across different geographies. We believe that investing in new computational social science methodologies such as NLP for policy analysis will allow us to create a knowledge and evidence base from which we can advocate for change. With data on frequency of policy instruments and trends across geographies we will be able to understand and advocate for better incentives for forest and landscape restoration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Omdena and DSSG Solve for their collaboration on this project; Karla Posada from El Salvador's Ministry of Environment and Natural Resources; Carolina Dreikorn from the United Nations Environment Programme; Rong Fang and John Brandt from the World Resources Institute; Francisco Pérez Canales and Ignacio Fernández from Omdena for their contributions in adapting Scrapy; Juan Felipe Cerón and Daniel Calle from Omdena for the preliminary work on S-BERT; Matt Sweeney and Brenda Jimenez from DSSG Solve for their contributions during the initial phase of the project; Kai Qi from Brown University for an early version of the policy framework. We thank Good Energies Foundation and HSBC for funding this research project and their support for exploring experimental methodologies to understand economic and financial incentives for restoration. Neither Good Energies Foundation nor HSBC did not have any direct involvement neither in the design of the project nor in the design of the experiments presented in this paper.

Appendix A. Appendix. Labeling differences among policy analysts

After the three labelers independently curated the HITL dataset, 1,373 new sentences were labeled as belonging to one of the six policy instruments considered in this study. In Table A1, there is the comparison of the sentences labeled by each labeler.

Two ways of comparing the labeling of the three labelers were used (Table A2): one where the majority (one half or higher) of the raters agreed with the assigned label, and one where at least one rater agreed with the assigned label (this was deemed necessary to compensate for cautious/conservative labelling).

Of all unique documents that were predicted as likely to contain an incentive by the pre-trained model, 62% were found to contain an incentive; in the case of 87% of them, the incentive type matched with the predicted label (e.g. the document contained credit incentives when it was predicted to contain credit incentives), according to at least one rater (in 82% of the cases, according to the majority of the raters). Additionally, just under half of the documents with incentives were found to contain more than one type of policy instrument.

In terms of unique instances (sentences), 48% of the instances that were predicted as likely to contain an incentive were found to contain an incentive. Compared with the performance at the document level, this reflects the characteristics of legislation texts, where details about the incentives are spread across different paragraphs and the difficulties in categorizing sentences without context. Still, this means that for a significant part of the automatically tagged instances, details could be found exactly in the highlighted sentences, with most of them correctly identifying the policy instrument.

Table A1
Incentive sentences classified in the different incentive instruments after curation of the HiTL dataset.

	Credit	Direct Payment	Fine	Supplies	Tax deduction	Technical assistance
Labeler1	54	16	30	15	14	21
Labeler 2	164	167	196	21	9	12
Labeler3	135	136	268	17	29	69

Table A2
Performance of automatic tagging compared to categorization by experts on instances ($n = 3089$), which represent sentences or part of sentences from pre-filtered documents ($n = 633$) with similarity score of >0.5 to pre-defined example query sentences ($n = 5$).

	Unique Inputs	N/% of inputs categorised as incentive (any type, any rater)	N/% of inputs categorised positively for type of incentive (at least one rater)	N/% of inputs categorised positively for type of incentive (majority of raters)	N/% of inputs categorised as containing incentive of different type
Instances	3089	1486	953	854	533
(of all instances)	*100%*	*48%*	*31%*	*28%*	*17%*
(of positively categorised instances)	–	*100%*	*64%*	*57%*	*36%*
Documents	633	391	341	320	195
(of all documents)	*100%*	*62%*	*54%*	*51%*	*31%*
(of positively categorised documents)	–	*100%*	*87%*	*82%*	*50%*

Table A3
Number of documents analyzed per country, and their sources.

Country	Number of documents	Issuing governmental authorities/sources
Mexico	15	SECRETARIA DE MEDIO AMBIENTE Y RECURSOS NATURALES
		COMISION NACIONAL FORESTAL
		SECRETARIA DE BIENESTAR
		CAMARA DE DIPUTADOS
Guatemala	11	MINISTERIO DE AMBIENTE Y RECURSOS NATURALES
		MINISTERIO DE EDUCACION
		CONGRESO DE LA REPUBLICA
		INSTITUTO NACIONAL DE BOSQUES
Peru	12	MINISTERIO DE AGRICULTURA Y RIEGO
		MINISTERIO DEL AMBIENTE
		CONGRESO DE LA REPUBLICA
		PRESIDENCIA DEL CONSEJO DE MINISTROS
		ORGANISMO DE EVALUACION Y FISCALIZACION AMBIENTAL
		ORGANISMO SUPERVISOR DE LA INVERSION EN ENERGIA Y MINERIA
		COMITE PERMANENTE DE NORMALIZACION
Chile	872	GOBIERNOS REGIONALES
		MINISTERIO DE AGRICULTURA
		MINISTERIO DE BIENES NACIONALES
		MINISTERIO DE DEFENSA NACIONAL
		MINISTERIO DE DESARROLLO SOCIAL
		MINISTERIO DE ECONOMIA, FOMENTO Y TURISMO
		MINISTERIO DE EDUCACION
		MINISTERIO DE ENERGIA
		MINISTERIO DE HACIENDA
		MINISTERIO DE JUSTICIA
		MINISTERIO DE LAS CULTURAS, LAS ARTES Y EL PATRIMONIO
		MINISTERIO DE MINERIA
		MINISTERIO DE OBRAS PUBLICAS
		MINISTERIO DE RELACIONES EXTERIORES
		MINISTERIO DE SALUD
		MINISTERIO DE TRANSPORTES Y TELECOMUNICACIONES;
		MINISTERIO DE VIVIENDA Y URBANISMO
		MINISTERIO DEL INTERIOR Y SEGURIDAD PUBLICA
		MINISTERIO DEL MEDIO AMBIENTE
		MINISTERIO DEL TRABAJO Y PREVISION SOCIAL
		MINISTERIO SECRETARIA GENERAL DE GOBIERNO
		MINISTERIO SECRETARIA GENERAL DE LA PRESIDENCIA
		MUNICIPALIDADES (various)
El Salvador	465	INSTITUCIONES AUTONOMAS (ALCADIAS MUNICIPALES)
		INSTITUCIONES AUTÓNOMAS (Consejo de Alcaldes del Área Metropolitana de San Salvador)
		INSTITUCIONES AUTÓNOMAS (Corte de Cuentas de la República)

(*continued on next page*)

Table A3 (*continued*)

Country	Number of documents	Issuing governmental authorities/sources
		INSTITUCIONES AUTÓNOMAS (Superintendencia General de Electricidad y Telecomunicaciones)
		MINISTERIO DE AGRICULTURA Y GANADERIA
		MINISTERIO DE ECONOMIA
		MINISTERIO DEL MEDIO AMBIENTE Y RECURSOS NATURALES
		MINISTERIO DE HACIENDA
		MINISTERIO DE RELACIONES EXTERIORES
		MINISTERIO DE SALUD
		MINISTERIO DE TRABAJO Y PREVISION SOCIAL
		PRESIDENCIA DE LA REPUBLICA
		ORGANO LEGISLATIVO

From: Accelerating Incentives: Identifying economic and financial incentives for forest and landscape restoration in Latin American policy using Machine Learning.

References

Affolder, N., 2019. Contagious environmental lawmaking. J. Environ. Law 31 (2), 187–212. https://doi.org/10.1093/jel/eqz011.

Bedi, G., Carrillo, F., Cecchi, G., et al., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 1, 15030. https://doi.org/10.1038/npjschz.2015.30.

Bowman, S.R., Angeli, G., Potts, C., Manning, C.D., 2015. A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 632–642. https://doi.org/10.18653/v1/D15-1075.

Brandt, J., 2019. Text mining policy: Classifying forest and landscape restoration policy agenda with neural information retrieval. In: FEED 19 Workshop at KDD 2019 (Anchorage, AK, USA).

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. arXiv Preprint. https://arxiv.org/abs/2005.14165.

Centro de Documentación Judicial, 2020. https://www.jurisprudencia.gob.sv/busqueda/busquedaLeg.php?id=2 (accessed 15 November 2020).

Cheng, S., Augustin, C., Bethel, A., Gill, D., Anzaroot, S., Brun, J., DeWilde, B., Minnich, R., Garside, R., Masuda, Y., Miller, D., Wilkie, D., Wongbusarakum, S., McKinnon, M., 2018. Using machine learning to advance synthesis and use of conservation and environmental evidence. Conserv. Biol. 32, 762–764. https://doi.org/10.1111/cobi.13117.

Chile, Ley, 2020. https://www.bcn.cl/leychile/ (accessed 15 November 2020).

Cockli, C., Moon, K., 2020. Environmental Policy. International Encyclopedia of Human Geography. Elsevier, pp. 227–233 (ISBN: 9780081022962).

Comisión Nacional Forestal, 2020. Reglas de Operación del Programa Apoyos para el Desarrollo Forestal Sustentable 2021. Diario Oficial de la Federación. https://www.conafor.gob.mx/apoyos/index.php/inicio/app_apoyos#/detalle/2021/92.

Congreso de la República de Guatemala, 2010. Resolución N° 4.28.2015: Reglamento de la Ley de Incentivos Forestales para Poseedores de Pequeñas Extensiones de Tierra de Vocación Forestal o Agroforestal – PINPEP. Diario de Centro América 90.

Congreso de la República de Guatemala, 2015. Decreto N° 2-2015: Ley de fomento al establecimiento, recuperación, restauración, manejo, producción y protección de bosques en Guatemala – PROBOSQUE.

Davidson, J., 2020. Zero-Shot Learning in Modern NLP: State-of-the-art NLP Models for Text Classification without Annotated Data. https://joeddav.github.io/blog/2020/05/29/ZSL.html (accessed 30 January 2021).

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019, pp. 4171–4186.

Diario Oficial de la Federación, 2020. Acuerdo por el que se emiten las Reglas de Operación del Programa Sembrando Vida, para el ejercicio fiscal 2020.

Elomina, J., Pülzl, H., 2021. How are forests framed? An analysis of EU forest policy. In: Forest Policy and Economics, vol. 127, p. 102448.

FitzHenry, F., Murff, H., Matheny, M., Gentry, N., Fielstein, E., Brown, S., Reeves, R.M., Aronsky, D., Elkin, P.L., Messina, V.P., Speroff, T., 2013. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. Med. Care 51 (6), 509–516. http://www.jstor.org/stable/23434332 (accessed 13 November 2020).

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Garcia-Gonzalo, J., Borges, J., 2019. Models and tools for integrated forest management and policy analysis: an editorial. Forest Policy Econ. 103 (2019), 1–3.

Green, A., 1997. Kappa statistics for multiple raters using categorical classifications. In: SAS Conference Proceedings: SAS Users Group International, p. 22.

Greene, D., Cross, J.P., 2015. Unveiling the political agenda of the European Parliament plenary: a topical analysis. In: Proceedings of the ACM Web Science Conference (WebSci '15). ACM, New York, NY, USA. https://doi.org/10.1145/2786451.2786464. Article 2, 10 pages.

Grimmer, J., Stewart, B.M., 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. 21, 267–297. https://doi.org/10.1093/pan/mps028.

Gunn, S.R., 1998. Support vector machines for classification and regression. ISIS Tech. Rep. 14, 5–16. http://www.isis.ecs.soton.ac.uk/isystems/kernel/.

Hagen, L., Uzuner, A., Kotfila, C., Harrison, T.M., Lamanna, D., 2015. Understanding Citizens' direct policy suggestions to the Federal Government: a natural language processing and topic modeling approach. In: 2015 48th Hawaii International Conference on System Sciences, pp. 2134–2143. https://doi.org/10.1109/HICSS.2015.257.

Haney, B.S., 2020. Applied natural language processing for law practice. In: B.C. Intell. Prop. & Tech. F.. https://doi.org/10.2139/ssrn.3476351.

Howlett, M., Ramesh, R., 1993. Patterns of Policy Instrument Choice: Policy Styles, Policy Learning and the Privatization Experience; Review of Policy Research 12; 1–2, 3–24. https://doi.org/10.1111/j.1541-1338.1993.tb00505.x.

Initiative 20x20, 2020. Restoring Latin America's Landscapes. https://initiative20x20.org/about (accessed 29 January 2021).

Jaramillo, C.F., Kelly, T., 1999. Deforestation and property rights. In: Keipi, K. (Ed.), Forest Resource Policy in Latin America. Inter-American Development Bank, Washington, D.C., pp. 111–134

Kehl, K.L., Elmarakeby, H., Nishino, M., Van Allen, E.M., Lepisto, E.M., Hassett, M.J., Johnson, B.E., Schrag, D., 2019. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. JAMA Oncol. 5, 1421–1429. https://doi.org/10.1001/jamaoncol.2019.1800.

Kissinger, G., Almeida, M.D., Coello, J., 2015. Fiscal incentives for agricultural commodity production: options to forge compatibility with REDD+. UN-REDD Programme Policy Brief 7.

Klie, J.C., Eckart de Castilho, R., Gurevych, I., 2020. From zero to Hero: human-in-the-loop entity linking in low resource domains. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020), Virtual conference.

Kohn, M.S., Sun, J., Knoop, S., Shabo, A., Carmeli, B., Sow, D., Syed-Mahmood, T., Rapp, W., 2014. IBM's health analytics and clinical decision support. Yearbook Medical Informatics 9, 154–162. https://doi.org/10.15265/IY-2014-0002.

Kouzis-Loukas, D., 2016. Learning Scrapy. Packt Publishing Ltd.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics. 33 (1), 159–174.

Liddy, E.D., 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd ed. Marcel Decker, New York.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J., 2021. Deep learning–based text classification: a comprehensive review. arXiv 54 (3). https://doi.org/10.1145/3439726. Article 62 (April 2021), 40 pages.

Ministerio de Agricultura de Chile, 2008. Ley N°20.283: Ley sobre recuperación del bosque nativo y fomento forestal.

OECD, 2017. Policy Instruments for the Environment Database. https://www.oecd.org/environment/indicators-modelling-outlooks/policy-instrument-database/.

Oficina de Estudios y Políticas Agrarias, 2021. Notas de Políticas Silvoagropecuarias y Rurales. https://www.odepa.gob.cl/notas-de-politicas-silvoagropecuarias-y-rurales.

Rana, P., 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. Environ. Res. Lett. 14, 024008.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3982–3992. https://arxiv.org/abs/1908.10084.

República de El Salvador, 2016. Ordenanza Municipal para la Protección, Conservación y Recuperación del Medioambiente. In: El Concejo Municipal de San José Villeneuve, en el Departamento de la Libertad.

Richards, K.R., 2000. Framing Environmental Instrument Choice. In: Duke Environmental Law and Policy Forum; 10; 221, pp. 221–285. https://scholarship.law.duke.edu/delpf/vol10/iss2/1.

Scoones, I., 2016. The politics of sustainability and development. Annu. Rev. Environ. Resour. 41 (1), 293–319.

Smith, R., 2007. Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02 September 2007, pp. 629–633.

Taylor, R., Davis, C., Brandt, J., Parker, M., Stäuble, T., Said, Z., 2020. The rise of big data and supporting technologies in keeping watch on the world's forests. Int. For. Rev. 22 (Supplement 1), 129–141.

Tricco, A., Langlois, E., Straus, S., 2017. Rapid Reviews to Strengthen Health Policy and Systems: A Practical Guide.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. ArXiv, 1706.03762 [Cs]. http://arxiv.org/abs/1706.03762.

vonHedemann, N., 2020. Transitions in payments for ecosystem Services in Guatemala: embedding forestry incentives into rural development value systems. Dev. Chang. 51, 117–143. https://doi.org/10.1111/dech.12547.

Wiedemann, G., Remus, S., Chawla, A., Biemann, C., 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized Embeddings. arXiv 1909, 10430.

Williams, A., Nangia, N., Bowman, S.R., 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, pp. 1112–1122. https://doi.org/10.18653/v1/N18-1101 (Long Papers).

Zarin, J., 2017. A Comparison of Case Law Results between Bloomberg Law's 'Smart Code' Automated Annotated Statutes and Traditional Curated Annotated Codes. https://doi.org/10.2139/ssrn.2998805.

World Resources Institute. Restoration Policy Accelerator, 2021–. (Accessed 20 October 2021).