Journal of King Saud University – Computer and Information Sciences 35 (2023) 101634

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Automatic explanation of the classification of Spanish legal judgments in jurisdiction-dependent law categories with tree estimators

Jaime González-González, Francisco de Arriba-Pérez, Silvia García-Méndez *, Andrea Busto-Castiñeira, Francisco J. González-Castaño

Information Technologies Group, atlanTTic, University of Vigo, Vigo, Spain

A R T I C L E   I N F O

Article history:
Received 10 May 2023
Revised 3 June 2023
Accepted 20 June 2023
Available online 3 July 2023

Keywords:
Natural language processing
Machine learning
Interpretable and transparent models
Natural language generation
Legal analysis
Human-in-the-loop

A B S T R A C T

Automatic legal text classification systems have been proposed in the literature to address knowledge extraction from judgments and detect their aspects. However, most of these systems are black boxes even when their models are interpretable. This may raise concerns about their trustworthiness. Accordingly, this work contributes with a system combining Natural Language Processing (NLP) with Machine Learning (ML) to classify legal texts in an explainable manner. We analyze the features involved in the decision and the threshold bifurcation values of the decision paths of tree structures and present this information to the users in natural language. This is the first work on automatic analysis of legal texts combining NLP and ML along with Explainable Artificial Intelligence techniques to automatically make the models' decisions understandable to end users. Furthermore, legal experts have validated our solution, and this knowledge has also been incorporated into the explanation process as "expert-in-the-loop" dictionaries. Experimental results on an annotated data set in law categories by jurisdiction demonstrate that our system yields competitive classification performance, with accuracy values well above 90%, and that its automatic explanations are easily understandable even to non-expert users.

1. Introduction

Recent improvements in Natural Language Processing (NLP) technologies and Machine Learning (ML) algorithms have allowed solving an ample variety of problems such as summarization (Trappey et al., 2020; Gambhir and Gupta, 2017), user profiling (Tellez et al., 2018; Flores et al., 2022) and decision making (Trappey et al., 2020; Rana and Varshney, 2021). They have jointly contributed to intelligent conversational assistants (Rustamov et al., 2021; Hasal et al., 2021) and sentiment and emotion analysis systems (Kastrati et al., 2021; Tao et al., 2019). More closely related to our work are text classification problems (Kowsari et al., 2019; Hettiarachchi et al., 2022; Škrlj et al., 2021), and particularly, those in the legal field (Medvedeva et al., 2020; Dyevre, 2021b; Dyevre, 2021a). This work focuses on them.

We contribute to the state of the art by combining NLP and ML to classify judgment texts and automatically explain this classification. As in other court systems, Spanish judgments follow a hierarchy of thematic areas named jurisdictions. Seven main areas exist: administrative, common, commerce, constitutional, criminal, social, and tax law. Different jurisdictions may share some law categories.

Previous authors have addressed the classification of legal texts from different perspectives. Consideration should be given to data analysis techniques that combine lexical, morphological, and syntactic feature inference (Thomas and Sangeetha, 2021) and those that focus on discourse organization patterns (Medvedeva et al., 2020). Others seek to classify legal texts into predefined categories and infer the court decision, *e.g.*, violation of the law or not (Medvedeva et al., 2020). Previous research has also considered the automatic detection of ideological bias in legal judgments (Hausladen et al., 2020). These approaches range from a semi-supervised pattern-based learning approach using annotated data sets (Thomas and Sangeetha, 2021) to unsupervised text-scaling (political dimension classification of parties and politicians (Dyevre, 2021b)).

* Corresponding author.
E-mail addresses: jaimegonzalez@gti.uvigo.es (J. González-González), farriba@gti.uvigo.es (F. de Arriba-Pérez), sgarcia@gti.uvigo.es (S. García-Méndez), abusto@gti.uvigo.es (A. Busto-Castiñeira), javier@det.uvigo.es (F.J. González-Castaño).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2023.101634
1319-1578/

However, most ML systems (either supervised or unsupervised) in these and other fields continue to build black-box solvers even when their models are interpretable, such as Decision Trees (DT) (Lee and Kim, 2016; Le and Moore, 2021) and Random Forests (RF) (Guidotti et al., 2019; Hatwell et al., 2020). Ignoring this aspect of the models may raise concerns about their trustworthiness. In this regard, safety and interpretability (Carvalho et al., 2019; Linardatos et al., 2020) pursue the extraction of knowledge from the trained classifiers by traversing the decision tree paths to better understand the performance of the ML model. In this context, the direct involvement of human evaluators in the ML models themselves, or *human-in-the-loop interpretability* (HITL), is attracting much attention (Drobnic et al., 2020).

We analyze the features that contribute to the decision path of a tree structure and check their threshold bifurcation values and convert this into understandable information, that is, natural language, by using templates. This is the first work on the automatic analysis and explanation of Spanish legal texts by combining NLP techniques and ML algorithms. We are unaware of prior work on the automatic explanation of legal texts' classification in natural language.

The rest of this paper is organized as follows. Section 2 reviews related work in legal text classification and states our contributions. Section 3 presents our proposal for legal text classification and explanation based on NLP and ML techniques. Section 4 described the results obtained on a real annotated experimental data set by paying particular attention to explaining the decisions. Finally, Section 5 concludes the paper and suggests future work.

2. Related work

Automated legal text classification has been an active research field in recent decades. Some first solutions included rule-based classifiers and ontologies, which Bartolini et al. (2004) applied to Italian law paragraphs; simple Neural Networks like self-organized maps as in Schweighofer et al. (2001), with which European laws in English, German, and French were classified; and other ML approaches such as those by Thompson (2001), who compared nearest neighbor likeness, C4.5 rules, and Ripper on diverse WESTLAW documents.

ML techniques are still used nowadays for the same purpose. Coltrinari et al. (2020) proposed a noise-proof classifier based on Linear Discriminant Analysis for Italian legal texts. Moreover, more recent studies follow semi-supervised techniques to extract relevant concepts from judicial texts (Thomas and Sangeetha, 2021). Other works apply Naive Bayes and other supervised methods (Medvedeva et al., 2020; Hausladen et al., 2020) and transformer-based language models and other Deep Neural Network (DNN) architectures, such as that by Tagarelli and Simeri (2021), a fine-tuned version of Italian BERT[1] (Devlin et al., 2019) for law article classification; and the Convolutional Neural Network (CNN) for multi-label Chinese legal text classification by Qiu et al. (2020). Unfortunately, most pre-trained language models' embeddings are not adapted to the legal domain (Beltagy et al., 2019), so they tend to underperform compared to traditional ML algorithms like RF, as noted by Chen et al. (2022). This issue can only be tackled with language model pre-training as in Chalkidis and Kampas (2019); Song et al. (2022), requiring enough computational resources and large amounts of specialized training data.

Even though we are not aware of any prior work on the automatic classification of Spanish legal texts, the contribution of this work also lies in the automatic explanation of the solution. In DARPA's Explainable Artificial Intelligence (XAI) program[2], explicability is defined as the capacity of AI systems to explain their rationale to human users by characterizing their strengths and weaknesses and providing an understanding of their future behavior (Gunning and Aha, 2019). Most explicability techniques fall under some of the following three strategies:

- **Deep Explanation** (DE). A modification of Deep Learning techniques (Mathew et al., 2021) to learn and favor explicable features. Some of these approaches back-propagate the outputs for a given input as in Montavon et al. (2017), which applies Taylor decomposition, and Xu et al. (2022), which infers the transfer function of intermediate layers by taking the conservation property into account. Class Activation Mapping (CAM)-based methods for CNN explicability (Kim et al., 2023b; Kim et al., 2023a) also fall under this category.
- **Interpretable Models** (IM). They use learning techniques that are more structured and easily understandable to humans. Some of this explicable ML techniques are DT (Sagi and Rokach, 2020; Cousins and Riondato, 2019; Özge Sürer et al., 2021), RF, representation techniques (Neto and Paulovich, 2021; Tandra and Manashty, 2021), Gradient Boosting Decision Trees (GBDT) (Delgado-Panadero et al., 2022), and AdaBoost, with interpretability applications such as computer-aided diagnosis (Hatwell et al., 2020). This work belongs to this category.
- **Model Induction** (MI). It is reverse engineering for inferring an explanation from an existing black box model. Some examples of this methodology include models that plot visual representations of features and output correlations (Apley and Zhu, 2020; Forzieri et al., 2021); and the model-agnostic methodology by Wachter et al. (2017), which explores counterfactual exhaustively to detect minimal feature changes that lead to a different output.

All these explicability techniques may be refined by human–computer interaction by applying (HITL) (Zanzotto, 2019) training. HITL interactions can be prior knowledge (Lage et al., 2018), or human feedback as in Reinforcement Learning applications (Wells and Bednarz, 2021; Lin et al., 2020).

XAI has gained increased attention in recent years, not only from a technological but also from a social sciences point of view (Miller, 2019). It is especially relevant for legal contexts (Hacker et al., 2020; Górski and Ramakrishna, 2021), where decision-making standards should be rigorous and (once explained) supervised by human experts to preserve transparency and fairness (Shook et al., 2018). This new research trend has led to several recent XAI proposals in law-related NLP applications, most involving predictors. For example, Park and Chai (2021) have predicted online privacy invasion cases in US judgments using different ML algorithms in an explicable manner, and Branting et al. (2021) have developed a proof-of-concept in which a training subset is annotated to improve explicability.

However, to our knowledge, no previous research has applied XAI techniques to legal judgment classification approaches, that is, to automatically explain their decisions in natural language so that they are understandable to a person, the only exception being our previous research on multi-label legal text classification (Arriba-Pérez et al., 2022). From these promising previous results, this paper focuses on the variation of law categories depending on legal jurisdictions. We apply tree-based ML models along with a more sophisticated model based on Gradient Boosting (GB). The

[1] Available at https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased, June 2023.

[2] Available at https://www.darpa.mil/program/explainable-artificial-intelligence, June 2023.

classification stage has been improved by including char-gram features, and thus, the explainability functionality has also been enhanced to manage the reconstruction of those terms when they are not directly translatable to natural language representation. Moreover, due to the complexity this entails, the XAI in this paper has been validated by legal experts. This knowledge has also been incorporated into the explanation process as "expert-in-the-loop" dictionaries.

3. System architecture

Fig. 1 shows the general architecture of the solution. First, the data preprocessing module transforms the original data source into a proper input format for the ML classifiers. Then, the main module performs feature engineering and classifies judgments by law categories independently within each jurisdiction. Its output is evaluated with standard metrics. The explicability module explains the decisions of the classifiers in natural language. In the following subsections, we detail these components.

3.1. Data preprocessing module

This module adapts the source data to the expected input format for the classifier. The data source is a collection of judgments of the Spanish legal system. Specifically, the module performs:

- **Stop words removal**. All words with low semantic load, such as prepositions, determiners, and connectors, are removed.
- **Text lemmatization**. First, all the accents, diaeresis, and diacritical marks are removed from the text. Then the text is split into word tokens and converted to lemmas.
- **Jurisdiction selection**. Finally, the judgments are sorted by the jurisdictions they belong to. This is straightforward since all judgments are identified that way in the data source.

3.2. Main module

3.2.1. Feature engineering

This first stage of the main module generates the features of the input texts. We employ two types of n-grams, char-grams, and word-grams. A char-gram is any sequence of n characters in the text, including blank spaces. A word-gram is any sequence of n words in the text.

3.2.2. Classification stage

The law judgments are classified by a single layer of parallel classifiers, with as many instances as jurisdictions in the data source. In our case, these are *Administrativo* (Administrative law), *Civil/Mercantil* (Common/Commerce law), *Civil* (Common law), *Constitucional* (Constitutional law), *Mercantil* (Commercial law), *Penal* (Criminal law), *Tributario* (Tax law) and *Social* (Social law). Since we are interested in explicability, we have selected RF as the target algorithm. However, we decided to also employ the much simpler DT basic classifier as a baseline. As fourth high-accuracy references, we considered GB, despite its high computational cost, and a Support Vector Machine (SVM) model, which is much more difficult to interpret than RF.

Each jurisdiction classifier is independently trained to maximize classification accuracy for its corresponding law categories. Table 1 shows the law categories in each jurisdiction[3].

[3] Spanish–English translation available at https://bit.ly/3IpxRMV, June 2023.

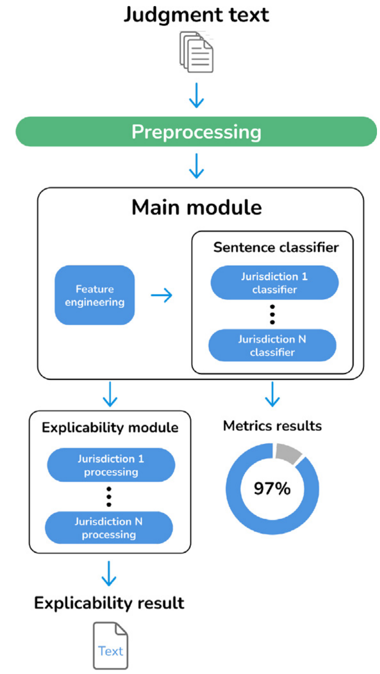

Fig. 1. Architecture of the solution composed of data preprocessing module, the main module with feature engineering and parallel classification stages, and explicability module.

3.3. Explicability module

This last stage performs all explicability subtasks. To support the criterion, this module needs all the structures of the trained trees of the classifiers in the main module. For each jurisdiction, the following subtasks are carried out:

- **Decision path extraction**. In case RF classifiers are used, for each possible decision, there is a set of decision paths (different tree leaves may correspond to the same category). Fig. 2 highlights in red an example of an estimator path in a tree leading to a decision that classifies a judgment into the *Derecho de Obligaciones y Contratos* (Contract law) category.
- **Extraction of relevant features for explicability**. We consider that a feature in a decision path is relevant if the logical routing condition ">" is fulfilled for that feature at some node in the path. The document has more occurrences of that feature than a threshold (which is determined by the training process). We keep a list of these features and their frequency.
- **Term reconstruction**. Many relevant features of the list in the previous subtask will be char-grams. Since char-grams are difficult to interpret, for each char-gram feature, the more noteworthy feature it was extracted from is estimated with the following substeps, which were empirically derived from experimental tests:
  1. If the char-gram is less than four characters long, it is discarded.
  2. The longest word-gram/biword-gram from the list containing the char-gram is identified.
  3. If substep 2 fails, the longest char-gram from the list containing the char-gram is identified.
  4. The frequency of the more noteworthy feature, if detected, is incremented with the frequency of the reconstructed char-gram.

As an example, features like "*hipotec*" or "*ecario*" will be reconstructed as word-gram "*hipotecario*" (mortgage related).

Table 1
Distributions of judgments by jurisdictions and law categories.

Jurisdiction	Law category	Samples
Administrative	Administrative Law	10,475
	Administrative Offence Law	2,071
	Patrimonial Liability of the Administration	1,847
	Civil Service Law	1,030
	Personal Rights	647
	Contract Law	457
	Rights in Rem	399
	Social Security Law	245
	Penitentiary Rights	39
	Collective Labor Law	14
Common/ Commerce	Contract Law	3,103
	Rights in Rem	1,777
Common	Family Law	11,489
	Rights in Rem	4,069
	Contract Law	1,366
	Inheritance Law	1,069
	Personal Rights	559
	Registry and Notary Law	282
Constitutional	Personal Rights	653
Commerce	Contract Law	11,522
	Rights in Rem	1,168
	Personal Rights	50
	Registry and Notary Law	6
Criminal	Crimes against Persons	7,452
	Crimes of Misrepresentation of Net Wealth and against Socioeconomic Order	6,024
	Crimes against the Constitution and the State	3,187
	Crimes against Public Safety	2,983
	Offences against Persons	682
	Offences against Territory Planning	226
	Offences against Public Trust	180
	Offences of Misrepresentation of Net Wealth and against Socioeconomic Order	175
	Penitentiary Rights	75
	Offences against the Constitution and the State	44
	Crimes Committed by Minors	16
	Special Crimes	2
	Personal Rights	1
Tax	Financial and Taxation Law	4,082
Social	Labor Law	8,483
	Social Security Law	7,103
	Collective Labor Law	992
	Labor Administration Law	114
	Personal Rights	5

- **Bagging**. A bag is generated for each law category in each jurisdiction with all expansions of the relevant features that explain classification decisions for that category, ordered by the updated frequencies in the previous substep. In the example in Fig. 2, "*hipotecario*" (mortgage related) would be added to bag *Derecho de Obligaciones y Contratos* (Contract law) of jurisdiction *Civil/Mercantil* (Common/Commerce) as an expansion of relevant feature ("hipotec").
- **Explanation**. The decision about the classification of a particular judgment is explained with the template in listing 1, where < term$_1$>… < term$_{m+p}$> are all the expansions of relevant features in the bag of law category < classifier_output> of jurisdiction < jurisdiction> that can lead to the classification of the text of judgment < judgment_id> into category < classifier_output>, of which the first *m* expanded features of the list are contained in an "expert-in-the loop" dictionary of relevant terms for law category < classifier_output> of jurisdiction < jurisdiction>.

4. Results and discussion

4.1. Data set

The data set is composed of 96,163 judgments from the Spanish legal system (average length of 3,103 words/ 19,217 characters each) provided by E4Legal Analytics SL[4] within the framework of a joint research project with the authors. It is larger than the data sets in related works such as Coltrinari et al. (2020) (2,030 documents, 800 words each on average); Medvedeva et al. (2020) (14,000 judicial decision documents from the European Court of Human Rights); Thomas and Sangeetha (2021) and Chen et al. (2022) (30,000 legal documents from India and the United States, respectively); and the POSTURE50K data set of 50,000 judgments in the pre-trained language model by Song et al. (2022). Moreover, the data set has 42 different output classes or law categories (see Table 1). By jurisdiction, there are 10 output classes for Administrative, 2 for Common/Commerce, 6 for Common, 1 for Constitutional, 4 for Commerce, 13 for Criminal, 1 for Tax, and 5 for Social. Every judgment in the data set is formally divided into four sections:

- Header: contextual information about the legal process (*i.e.*, process id, participants involved, court type, etc.)
- Case precedents: description of every fact, assumption, and testimony related to the legal process.
- Law fundamentals: an exposition of every law or legal regulation applied to the legal process considered in the judgment.
- Decision: the court decision in the legal text.

We decided to use only the header and the law fundamentals, as these sections carry more information about the law category. Jointly, they have an average length of 2,405 words/ 14,864 characters. Every judgment has a jurisdiction identifier and has been labeled by a legal expert with a first law category and two optional alternative law categories the judgment could also belong to[5], so that 59.78% of the judgments (that is, most of them) have a single law category label, 32.68% have one alternative law category label, and 7.54% have two alternative law category labels. Table 1 shows the distribution of law categories by jurisdictions in the data set. We remark that this labeling was not produced for this paper; therefore, there is no labeling bias. Judgment labels were produced for filing purposes and not for Machine Learning classification.

Next, we defined the following split into training and testing sets for each jurisdiction:

- **Train and test #1 combined subset**. We reduced the imbalance in the representatives of the different law categories in each jurisdiction of the data set when only considering the main law category label. For example, in jurisdiction Commercial law, there are 11,522 judgments of *Derecho de Obligaciones y Contratos* (Contract law) and only 1,168 judgments of *Derechos Reales* (Rights in Rem/Property law). Specifically:
  1. Every law category of the jurisdiction with more than 5,000 samples was randomly downsampled to the next thousand above the size of the largest law category of that same jurisdiction with less than 5,000 samples.
  2. Every law category of the jurisdiction with less than 50 samples was considered irrelevant and discarded from the train and test #1 combined subset.

  Finally, the train and test #1 combined subset was randomly split into independent train and test #1 subsets with 80% and 20% samples, respectively.
- **Test #2 subset**. Its purpose is to evaluate the classifier on groups of judgments with the real proportions of the law categories of the jurisdiction in the original data set by randomly

[4] Available at https://www.emerita.legal, June 2023.
[5] Note that both first and alternative law categories are within the same jurisdiction.

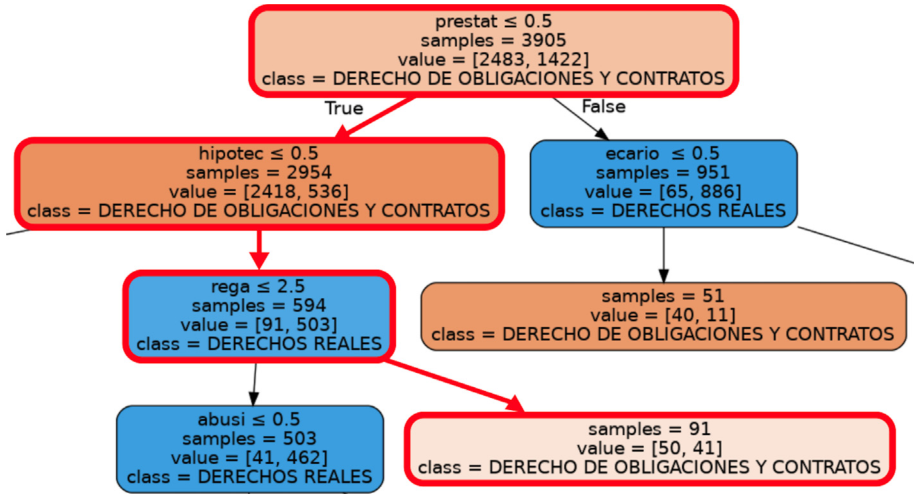

Fig. 2. Example of decision path in a decision tree for Common/Commerce law.

```
La clasificación de la sentencia sentenceID de la jurisdicción jurisdictionName en el
    derecho classifierName puede explicarse por los términos relevantes: term 1, ...,
    term m. Otros términos tenidos en cuenta son term m+1, ..., term m+p.

The classification of the sentence sentenceID of jurisdiction jurisdictionName in the
    law classifierName can be explained by the relevant terms terms 1, ..., and m. Other
     terms taken into account are: term m+1, ..., term m+p.
```

Listing 1. Original template in Spanish and translated into English.

sampling the judgments in that jurisdiction except those in the train set until producing a subset with the same size as test #1 subset. Therefore, test subsets #1 and #2 may share some judgments, but test #2 subset may also include, for example, judgments of law categories that were discarded due to their irrelevance to the jurisdiction.

4.2. Implementations

We made the following development choices for each module in the architecture:

- **Preprocessing** (Section 3.1). For removing stop words, we use the Spanish stop word list from the NLTK library[6]. This list includes prepositions, determiners, and frequent verbs such as *ser* (to be), *estar* (to be), and *tener* (to have) and their most common conjugations. To perform the lemmatization task, the text is split using the word tokenizer also from the NLTK library[6] and then converted to lemmas with the spaCy library[7] using the es_core_news_sm model[8].
- **Feature engineering** (Section 3.2.1). This module uses the CountVectorizer[9] function from the Scikit-Learn Python library to generate the *n*-grams.
- **Classification algorithms** (Section 3.2.2). We employed the SVM[10], DT[11], RF[12] and GB[13] implementations of the Scikit-learn Python library.
- **Relevant features' extraction** (Section 3.3). Every node (excepting its leaves) of a decision tree has two child nodes (this also holds for the individual trees forming a random forest). Each node is defined by feature X and threshold N, where the left child is reached by the inputs with N units of X at most ($\leqslant$condition) and the right child by the inputs with more than N units of X ($>$ condition holds). The Scikit-Learn implementations of DT and RF store and represent trees as numeric arrays of nodes, where the precedence in the array determines if a node is a left or right child of its preceding node. Our explanation methodology consists of a custom function that navigates those arrays by extracting every feature associated with a child that satisfies the $>$ condition until a leaf is reached.

4.3. Tuning and training

- **Char-grams and word-grams**. CountVectorizer instances are configured with the frequency of the *n*-grams in the document and the number of elements they must cover *n*. We tried

6 Available at https://www.nltk.org, June 2023.
7 Available at https://spacy.io, June 2023.
8 Available at https://spacy.io/models/es#es_core_news_sm, June 2023.
9 Available at https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html, June 2023.
10 Available at https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html, June 2023.
11 Available at https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html, June 2023.
12 Available at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, June 2023.
13 Available at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html, June 2023.

Table 2
Input features for the classifiers by jurisdiction.

Jurisdiction	Number of features
Administrative	32237
Common/Commerce	32412
Common	28674
Commerce	36440
Criminal	31816
Social	27923
Total	189502

different ranges of values for these parameters and checked the resulting model accuracies for different combinations of values. The final frequency choice was minimum and maximum n-gram frequencies of 5% and 50%. For char-grams sizes, the choice was $n \in [3, 7]$. Regarding word-grams, only word-grams and biword-grams (any two words separated by a blank space) were accepted.

- **Selection of relevant features**. We applied the SelectPercentile[14] function from the Scikit-Learn Python library to select relevant n-grams. This function evaluates the relationship between features and variables to predict a class. It is configured with two parameters, a scoring function, and a selection percentile value. We chose Chi-squared as the score function for its good performance (Arriba-Pérez et al., 2020). Then, seeking a compromise between accuracy and execution time, the selection percentile was set to 20%. Table 2 details the number of input features for the classifiers by jurisdiction.
- **Classifier hyperparameters' optimization**. Hyperparameters define a classifier's structural/numerical behavior at the training stage. To find an efficient combination of hyperparameters, we independently applied the GridSearchCV[15] function from the Scikit-Learn Python library to the different jurisdictions in our problem. GridSearchCV needs arrays of ranges for every hyperparameter. In the case of SVM, the hyperparameters that are considered are the inverse of the regularization parameter (`C`), the loss function, the maximum number of iterations, the threshold for stopping criteria and the class balancing method. In the case of DT, they are the maximum depth of the tree, the minimum number of samples to split a node or for a node to become a leaf, the criterion and the strategy for splits, and the maximum number of features to be considered. In the case of RF, the total number of estimators that form the forest also exists. In the case of GB, two extra parameters are the learning rate across the iterations and the proportion of samples to train every individual estimator.
Tables 3–6 show some explored parameter ranges and the resulting values for each algorithm. In addition, for the SVM classifier, `C` was evaluated for 1e-4, 1e-3, 1e-2, 1e-1 and 1, and in all cases 1e-4 was selected. For the DT classifier, `min_samples_leaf` was evaluated for 0.005, 0.01, and 0.0015, and in all cases, 0.005 was selected. The `splitter` processes tested were `best` and `random`, and the former was chosen. Regarding RF, `min_samples_leaf` was evaluated for the same values as in DT, and 0.005 was selected in all cases except for Common/Commerce, for which 0.001 was picked; for `max_features`, `auto` was selected instead of `sqrt`; `criterion` was tested with `gini` and `entrophy` and the former was preferred. Finally, for GB, `max_features` was the same as for RF; `max_depth` values of 2, 4, and 6 were tested, and this last value was chosen; and, for `n_estimators`, 100, 200 and 500 were tested, and 200 was selected.

[14] Available at https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectPercentile.html, June 2023.
[15] Available at https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html, June 2023.

4.4. Classification results

The experiments were executed on a computer with the following specifications:

- Operating System: Ubuntu 20.04.3 LTS 64 bits
- Processor: IntelXeon Platinum 8375C 2.9 GHz
- RAM: 64 GB DDR4
- Disk: 500 GB SSD

To evaluate the performance of the algorithms, we chose three different macro metrics averaging the individual metrics for each of the law categories within each jurisdiction: accuracy, recall, and F1. Due to the high imbalance of samples of the different law categories in some jurisdictions, F1 and recall metrics are weighted averages assigning proportionally larger weights to the categories with more samples. We used two different methodologies of evaluation:

- 1 to 1: only considering a classification is successful if it correctly guesses the main law category of a judgment.
- 1 to 3: if it correctly guesses the main law category or any of the two secondary law categories, if available.

Tables 7 and 8 show the results for the SVM algorithm, Tables 9 and 10 show the results for the DT algorithm, Tables 11 and 12 show the results for the RF algorithm, and Tables 13 and 14 show the results for the GB algorithm. Column "Time" represents overall computing time, including processing, training, and execution times. As expected, RF outperforms the DT baseline in all the jurisdictions for all the metrics. The accuracy gap was as large as 13% for Criminal law. However, it was not that large in many cases, indicating that the problem can be treated with tree methodologies. The difficulty of classifying some jurisdictions with a basic tree methodology can be explained by the different sizes of law categories per jurisdiction (*e.g.*, Common/Commerce law has only two possible law categories, and Criminal law has nine). SVM performed better than DT. Compared with the RF model, the much less interpretable SVM model had lower accuracy for the administrative, commerce, and social jurisdictions. If we compare the results with test subset #1 with those with test subset #2, the latter is consistently better since it respects the real proportions of judgments, and some categories are highly represented. When comparing methodologies 1 to 1 and 1 to 3, as we could expect, the results improve noticeably with the second. This is because double and triple labeling reduces both the effect of labeling errors by human experts and the confusion between law categories that are very similar from the perspective of some judgments.

Note that the 1 to 3 metrics of RF are outstanding, well above 90% in all cases but for Criminal law, although they are also close to 90% in that case. The metrics are similar if we compare RF with GB. The winner depends on the jurisdiction, although all 1 to 3 metrics exceed 90% with GB. This proximity backs the selection of RF for the explicability module, owing to its suitability. Besides, regarding computational cost, GB was much harder to train. DT and GB were the fastest and slowest methods, respectively. Therefore, SVM and RF achieved the best trade-off between performance and computing time.

4.5. Explicability results

As introduced in Section 3.3, the explicability module follows an "expert-in-the-loop" approach. Specifically, we asked an expert lawyer of E4Legal Analytics SL to inspect the expansions of the 50 most frequent relevant features of each law category within each

Table 3
SVM hyperparameters.

	loss {hinge, squared_hinge}	**max_iter** {250, 500, 1000}	**tol** {0.000001,0.00001, 0.0001, 0.001}	**class_weight** {None, Balanced}
Administrative	hinge	250	0.0001	None
Common	squared_hinge	250	0.00001	None
Common/Commerce	squared_hinge	250	0.00001	Balanced
Commerce	hinge	250	0.00001	None
Criminal	hinge	1000	0.0001	None
Social	squared_hinge	250	0.00001	None

Table 4
Decision Tree hyperparameters.

	criterion {gini, entropy}	**max_depth** [2,8]	**max_features** {auto, sqrt}	**min_samples** {0.0005, 0.001, 0.0015}
Administrative	entropy	8	auto	0.005
Common	entropy	8	sqrt	0.005
Common/Commerce	entropy	7	auto	0.100
Commerce	entropy	8	auto	0.005
Criminal	gini	7	sqrt	0.100
Social	entropy	8	sqrt	0.005

Table 5
Random Forest hyperparameters.

	max_depth {5, 10, 15, 25, 50, 100}	**min_samples** {0.0005, 0.001, 0.0015}	**n_estimators** {200, 500, 1000, 2000}
Administrative	100	0.0010	2,000
Common	15	0.0005	200
Common/Commerce	15	0.0005	2,000
Commerce	100	0.0005	1,000
Criminal	50	0.0010	200
Social	25	0.0010	1,000

Table 6
Gradient Boosting hyperparameters.

	learning_rate {0.05, 0.1, 0.15}	**subsample** {0.5, 0.6, 0.7, 0.8}	**min_samples_leaf** {0.01, 0.05, 0.1, 0.25}	**min_samples** {0.005, 0.05, 0.1}
Administrative	0.05	0.80	0.01	0.05
Common	0.10	0.80	0.01	0.10
Common/Commerce	0.10	0.80	0.10	0.05
Commerce	0.10	0.50	0.01	0.05
Criminal	0.10	0.80	0.01	0.05
Social	0.10	0.80	0.01	0.10

Table 7
SVM results by jurisdiction (1 to 1).

	Accuracy (%)		**F1 (%)**		**Recall (%)**		**Time (s)**
Jurisdiction	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	
Administrative	85.34	85.08	85.05	85.28	85.16	84.90	19.75
Common	87.19	90.39	86.97	90.29	87.12	90.32	24.02
Common/Commerce	96.72	96.72	96.71	96.71	96.72	96.72	1.51
Commerce	89.42	91.14	89.03	91.51	89.15	90.85	11.98
Criminal	76.50	77.06	75.98	76.67	76.26	76.82	119.08
Social	79.19	82.58	78.81	83.45	78.94	82.32	11.79

jurisdiction, taken from the decision paths of the trained models, and pick those that seemed relevant to her. This produced an "expert-in-the-loop" dictionary[3] per (jurisdiction, law category) pair with up to 50 expanded terms. Table 15 shows the complete list of jurisdictions, the law categories within each jurisdiction, and their translations to English. The expert layer was presented

Table 8
SVM results by jurisdiction (1 to 3).

	Accuracy (%)		F1 (%)		Recall (%)		Time (s)
Jurisdiction	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	
Administrative	92.82	94.84	92.69	95.10	92.63	94.64	19.75
Common	94.97	96.19	95.09	96.27	94.89	96.11	24.02
Common/Commerce	98.56	98.56	98.57	98.57	98.56	98.56	1.51
Commerce	96.89	94.40	96.79	94.59	96.59	94.11	11.98
Criminal	88.75	87.87	88.52	87.33	88.47	87.59	119.08
Social	90.81	92.74	90.80	94.21	90.51	92.44	11.79

Table 9
Decision Tree results by jurisdiction (1 to 1).

	Accuracy (%)		F1 (%)		Recall (%)		Time (s)
Jurisdiction	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	
Administrative	74.19	75.79	73.36	75.89	74.03	75.79	1.13
Common	76.52	80.41	76.15	81.03	76.46	80.35	1.26
Common/Commerce	95.28	95.28	95.27	95.27	95.28	95.28	0.69
Commerce	87.71	95.18	86.94	94.76	87.44	94.88	0.69
Criminal	56.77	59.81	54.21	57.14	56.58	59.62	1.17
Social	72.74	82.74	72.31	83.35	72.51	82.48	0.49

Table 10
Decision Tree results by jurisdiction (1 to 3).

	Accuracy (%)		F1 (%)		Recall (%)		Time (s)
Jurisdiction	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	
Administrative	81.57	86.37	80.55	86.34	81.40	86.19	1.13
Common	85.20	86.33	85.17	86.91	85.13	86.26	1.26
Common/Commerce	97.85	97.85	97.86	97.86	97.85	97.85	0.69
Commerce	94.87	97.67	94.97	97.11	94.57	97.36	0.69
Criminal	70.30	72.18	69.49	71.15	70.07	71.95	1.17
Social	85.32	89.68	85.16	90.54	85.05	89.39	0.49

Table 11
Random Forest results by jurisdiction (1 to 1).

	Accuracy (%)		F1 (%)		Recall (%)		Time (s)
Jurisdiction	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	
Administrative	86.84	85.91	86.51	85.87	86.66	85.73	235.19
Common	85.28	89.38	84.47	88.91	85.21	89.30	87.80
Common/Commerce	97.33	97.33	97.34	97.34	97.33	97.33	130.11
Commerce	94.25	96.89	93.73	96.49	93.95	96.59	95.58
Criminal	78.10	79.30	77.15	78.46	77.85	78.46	106.84
Social	85.00	84.03	84.63	85.18	84.73	83.76	93.49

Table 12
Random Forest results by jurisdiction (1 to 3).

	Accuracy (%)		F1 (%)		Recall (%)		Time (s)
Jurisdiction	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	
Administrative	93.19	95.92	92.93	96.00	92.99	95.72	235.19
Common	93.39	95.05	93.37	95.01	93.31	94.98	87.80
Common/Commerce	98.46	98.46	98.46	98.46	98.46	98.46	130.11
Commerce	99.22	98.91	99.07	98.39	98.91	98.60	95.58
Criminal	87.83	88.79	87.60	87.64	87.55	88.51	106.84
Social	96.94	94.68	97.21	97.03	96.62	94.37	93.49

Table 13
Gradient Boosting results by jurisdiction (1 to 1).

	Accuracy (%)		F1 (%)		Recall (%)		Time (s)
Jurisdiction	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	
Administrative	89.42	88.23	89.21	88.23	89.23	88.05	6758.11
Common	89.25	92.50	89.02	92.42	89.18	92.42	6582.51
Common/Commerce	96.62	96.62	96.62	96.62	96.62	96.62	354.93
Commerce	93.78	95.49	93.44	95.32	93.49	95.19	925.03
Criminal	82.31	83.39	81.87	83.00	82.04	83.12	11414.21
Social	86.45	87.42	86.05	87.93	86.17	87.14	1216.75

Table 14
Gradient Boosting results by jurisdiction (1 to 3).

	Accuracy (%)		F1 (%)		Recall (%)		Time (s)
Jurisdiction	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2	
Administrative	95.15	96.90	94.96	96.97	94.95	96.70	6758.11
Common	94.97	96.43	94.86	96.39	94.89	96.35	6582.51
Common/Commerce	97.85	97.85	97.84	97.84	97.85	97.85	354.93
Commerce	98.29	97.51	98.07	97.09	97.98	97.21	925.03
Criminal	91.43	91.99	91.19	91.52	91.14	91.70	11414.21
Social	96.94	95.65	97.04	96.66	96.62	95.34	1216.75

Table 15
Law categories. English translation.

Administrative (*Administrativo*)	
Derecho Administrativo	Administrative Law
Derecho Administrativo-Sancionador	Administrative Offence Law
Responsabilidad Patrimonial de la Administración	Patrimonial Liability of the Administration
Derecho de la Función Pública	Civil Service Law
Derecho de Persona	Personal Rights
Derecho de la Contratación Pública	Contract Law
Derechos Reales Administrativos	Rights in Rem
Derecho de la Seguridad Social	Social Security Law
Common/Commerce (*Civil/Comercial*)	
Derecho de Obligaciones y Contratos	Contract Law
Derechos Reales	Rights in Rem
Common (*Civil*)	
Derecho de Familia	Family Law
Derechos Reales	Rights in Rem
Derecho de Obligaciones y Contratos	Contract Law
Derecho Sucesorio	Inheritance Law
Derecho de Persona	Personal Rights
Derecho Registral y Notarial	Registry and Notary Law
Constitutional (*Constitucional*)	
Derecho de Persona	Personal Rights
Commerce	
Derechos de Obligaciones y Contratos	Contract Law
Derechos Reales	Rights in Rem
Derecho de Persona	Personal Rights
Criminal (*Penal*)	
Delitos contra las Personas	Crimes against the Persons
Delitos contra el Patrimonio y el Orden Socioeconómico	Crimes of Misrepresentation of Net Wealth and against Socioeconomic Order
Delitos contra la Constitución y el Estado	Crimes against the Constitution and the State
Delitos contra la Seguridad Colectiva	Crimes against Public Safety
Faltas contra las Personas	Offences against the Persons
Delitos contra la Ordenación del Territorio	Offences against Territory Planning
Delitos contra la Fe Pública	Offences against Public Trust
Faltas contra el Patrimonio y el Orden Socioeconómico	Offences of Misrepresentation of Net Wealth and against Socioeconomic Order
Derecho Penitenciario	Penitentiary Rights
Tax (*Tributario*)	
Derecho Financiero y Tributario	Financial and Taxation Law
Social (*Social*)	
Derecho del Trabajo	Labor Law
Derecho de la Seguridad Social	Social Security Law
Derecho Colectivo del Trabajo	Collective Labor Law
Derecho de la Administración Laboral	Labor Administration Law
Derecho de Persona	Personal Rights

Table 16
Validation of dictionaries by a human expert (in percentage).

Jurisdiction	Law category	Question #1	Question #2
Administrative	Administrative Law	96.00	98.00
	Administrative Offence Law	94.00	98.00
	Patrimonial Liability of the Administration	94.00	100.00
	Civil Service Law	96.00	96.00
	Personal Rights	86.00	92.00
	Contract Law/ Public Contracting Law	100.00	100.00
	Rights in Rem	100.00	100.00
	Social Security Law	90.00	92.00
Common/ Commerce	Contract Law	100.00	98.00
	Rights in Rem	100.00	100.00
Common	Family Law	100.00	100.00
	Rights in Rem	98.00	100.00
	Contract Law	100.00	100.00
	Inheritance Law	96.00	96.00
	Personal Rights	100.00	100.00
	Registry and Notary Law	100.00	100.00
Commerce	Contract Law	96.00	96.00
	Rights in Rem	98.00	98.00
	Personal Rights	84.00	84.00
Criminal	Crimes against the Persons	96.00	96.00
	Crimes of Misrepresentation of Net Wealth and against Socioeconomic Order	100.00	100.00
	Crimes against the Constitution and the State	96.00	96.00
	Crimes against Public Safety	94.00	96.00
	Offences against the Persons	88.00	82.00
	Offences against Territory Planning	76.00	76.00
	Offences against Public Trust	84.00	84.00
	Offences of Misrepresentation of Net Wealth and against Socioeconomic Order	78.00	78.00
	Penitentiary Rights	74.00	74.00
Social	Labor Law	98.00	98.00
	Social Security Law	92.00	98.00
	Collective Labor Law	92.00	96.00
	Labor Administration Law	94.00	78.00

with two questions on each expansion of a selected relevant feature for law category *A* and jurisdiction *B*.

- **Question #1**. Is this term relevant in legal texts?
- **Question #2**. Is this term relevant to the law category *A* pertaining to the jurisdiction *B*?

Table 16 shows the percentages of terms per law category that were picked. Note that most of them were so. This is an exciting result as a double check of training quality by a human expert. Almost every classifier obtained expert validation scores above 90%. Table 18 shows an example of the production of the "expert-in-the-loop" dictionary of law category *Derecho de la Seguridad Social* (Social Security Law) of jurisdiction *Administrativo* (Administrative law).

Note that the expert discarded "*eral se*" and "*l segur*" as meaningless to a human -although the latter is likely related to "*segur*"*idad Social* (Social Security), and thus, it is relevant-; "*refundido*" (consolidated), which is a meaningful but rather general adjective; and "*españa*" (Spain) a term that is likely to appear in many Spanish judgments about nationals.

Once the "expert-in-the-loop" dictionaries were validated, it was possible to complete the explicability module described in Section 3.3. Table 17 shows some examples of judgment explanations produced by the system.

Even for a non-expert, we observe that the expanded terms that explain the judgments are meaningful and highly related to their respective jurisdictions. There are some irrelevant errors, such as extra characters owing to the expansion of char-grams that included blank spaces (*e.g.*, "*ABUSIVO.*", abusive). It seems that some relevant expanded terms were not included in the dictionaries simply because they were not frequent enough (*e.g.*, "*ESTATUTO TRABAJADORES*", worker regulations). However, once added, they contribute enormously to explaining the judgment. In very few cases, the terms offer little information (*e.g.*, "*CORRESPONDIENTE*", corresponding) or are hard to interpret ("*EY 1/20*", possibly from "*LEY 1/20*", law 1/20). These explanations give a legal expert a clear first impression about the contents of a judgment.

Table 17
Examples of judgment explanations produced by the system.

Jurisdiction	Is this term relevant to legal texts?
Administrative Law.	*La clasificación de la sentencia 90483 de la jurisdicción ADMINISTRATIVO en la categoría DERECHO ADMINISTRATIVO puede explicarse por los términos legales relevantes ["TECNICA", "SOLICITUD", "CONTRATO", "CONVOCATORIA", "TRABAJO", "CONCESION"]. Otros términos tenidos en cuenta son ["RECURSO REPOSICION", "HABER JUSTIFICAR", "REPRESENTACION", "DETERMINACION", "ARTICULO 22.4", "CERTIFICACION", "CORRESPONDER", "TERRITORIAL", "DECLARACION", "EXPLOTACION"].*
Common/Commerce Law.	*La clasificación de la sentencia 18639 de la jurisdicción CIVIL/MERCANTIL en la categoría DERECHOS REALES puede explicarse por los términos legales relevantes ["AMORTIZACION", "ESCRITURA", "GARANTIA", "CONSUMIDOR", "CLAUSULA", "PRESTAMO", "EJECUCION", "HIPOTECARIA", "PRESTAMO,", "CONSTITUCION", "GARANTIA HIPOTECARIO", "DEUDA", "CLAUSULA ABUSIVO", "ABUSIVO", "HIPOTECARIO"]. Otros términos tenidos en cuenta son ["EXIGENCIA BUEN", "INTERÉS DEMORA", "CONTROVERTIDO", "CONTRACTUAL", "DEVOLUCION", "EQUILIBRIO", "ABUSIVO.", "CLARIDAD", "RELATIVO", "EY 1/20"].*
Common Law.	*La clasificación de la sentencia 56362 de la jurisdicción CIVIL en la categoría DERECHO DE FAMILIA puede explicarse por los términos legales relevantes ["TITULAR", "PENSION ALIMENTICIO", "FAMILIA", "PROTECCION", "ECONOMICO", "INGRESO", "PROGENITOR", "ALIMENTO", "MENSUAL", "PENSION", "MENOR", "PADRE", "ALIMENTICIO", "MEDIDA", "HIJO", "MATRIMONIAL"]. Otros términos tenidos en cuenta son ["AMBOS PROGENITOR", "GUARDA CUSTODIA", "CORRESPONDIENTE", "REGIMEN VISITA", "INTER MENOR", "VALORACION", "HIJO MENOR", "MENOR EDAD", "EXISTENCIA", "MENSUAL"].*
Commerce Law.	*La clasificación de la sentencia 7972 de la jurisdicción MERCANTIL en la categoría DERECHO DE OBLIGACIONES Y CONTRATOS puede explicarse por los términos legales relevantes ["AUTORIZACION", "CREDITO", "VIVIENDA", "EJECUCION", "INDUSTRIAL", "RIESGO", "CELEBRADO", "PROPIEDAD", "OBRA"]. Otros términos tenidos en cuenta son ["FALTA LEGITIMACION", "PRUEBA PERICIAL", "INCUMPLIMIENTO", "CERTIFICACION", "CONTRATACION", "LEGITIMACION", "FINANCIACION", "DISPOSITIVO", "TRANSMISION", "ADQUISICION"].*
Criminal Law.	*La clasificación de la sentencia 74469 de la jurisdicción PENAL en la categoría DELITOS CONTRA EL PATRIMONIO Y EL ORDEN SOCIOECONOMICO puede explicarse por los términos legales relevantes ["MERCANTIL", "DELITO ESTAFA", "RESPONSABILIDAD CIVIL", "DOCUMENTO", "PATRIMONIAL", "FALSEDAD", "ESTAFA", "TRAFICO", "ESTAFA,"]. Otros términos tenidos en cuenta son ["PROCEDIMIENTO ABREVIADO", "JUZGADO INSTRUCCION", "JUZGADO PENAL", "CALIFICACION", "NATURALEZA", "ACREDITAR", "EJECUCION", "APARTADO", "NEGOCIO", "LABORAL"].*
Social Law.	*La clasificación de la sentencia 44343 de la jurisdicción SOCIAL en la categoría DERECHO DEL TRABAJO puede explicarse por los términos legales relevantes ["ORDINARIO", "RECLAMACION", "MEDIDA", "ESTATUTO"]. Otros términos tenidos en cuenta son ["PROCEDIMIENTO ORDINARIO", "ESTATUTO TRABAJADORES", "PRESTACION SERVICIO", "EMPRESA DEMANDADO", "PRESTAR SERVICIO", "RESPONSABILIDAD", "INTERPRETACION", "CENTRO TRABAJO", "CIRCUNSTANCIA", "MANTENIMIENTO"].*

Table 18
Questionnaire to generate the "expert-in-the-loop" dictionary of law category *Derecho de la Seguridad Social* of jurisdiction *Administrativo*, with the answers by the legal expert.

	Is this term relevant to legal texts?	Is this term relevant to the law category Derecho de la Seguridad Social pertaining to jurisdiction Administrativo?
seguridad	Yes	Yes
seguridad social	Yes	Yes
trabajo	Yes	Yes
funcionario	Yes	Yes
reclamacion	Yes	Yes
responsabilidad	Yes	Yes
patrimonial	Yes	Yes
social	Yes	Yes
indemnizacion	Yes	Yes
trabajador	Yes	Yes
desempeñar	Yes	Yes
puesto trabajo	Yes	Yes
extranjero	Yes	Yes
incapacidad	Yes	Yes
cotiza	Yes	Yes
prestación	Yes	Yes
contrato	Yes	Yes
liquidacion	Yes	Yes
laboral	Yes	Yes
carrera	Yes	Yes
regimen	Yes	Yes
profesional	Yes	Yes
enfermedad	Yes	Yes
permanente	Yes	Yes
autorizacion	Yes	Yes
refundido	No	No
personal	Yes	Yes
tributario	Yes	Yes
ley general	Yes	Yes
inspeccion	Yes	Yes
responsabilidad patrimonial	Yes	Yes
ingreso	Yes	Yes
contratacion	Yes	Yes
accidente	Yes	Yes
l segur	No	No
percibir	Yes	Yes
sanitario	Yes	Yes
provincial	Yes	Yes
eral se	No	No
empresa	Yes	Yes
funcion publica	Yes	Yes
retribución	Yes	Yes
españa	No	No
mercantil	Yes	Yes
actividad	Yes	Yes
territorio	Yes	Yes
reglamento	Yes	Yes
baja	Yes	Yes
beneficiario	Yes	Yes
español	No	Yes

5. Conclusions

Motivated by the literature's lack of interpretable legal text classification systems, we propose a solution to automatically explain the classification of Spanish legal judgments with tree estimators. Our work contributes to state-of-the-art with a novel architecture that combines NLP techniques with ML algorithms to classify legal texts and provide explanations in natural language on the outcome of the models, thus increasing the trustworthiness of the results to human operators. The explanations are produced as natural language templates that make sense even to non-experts users.

As far as we know, our work is the first proposal for automatic analysis of Spanish legal texts by combining NLP and ML from a XAI perspective and producing explanations in natural language in this domain. Legal experts have validated the solution, and this knowledge has also been incorporated into the explanation process as "expert in the loop" dictionaries.

Experimental results on a large data set of judgments annotated by jurisdictions and their law categories show a satisfactory performance of the interpretable classifiers. Macro and micro evaluation metrics, averaged and weighted, attained values well above 90% in most experiments.

We plan to extend this analysis to other languages and court systems in future work.

CRediT authorship contribution statement

Jaime González-González: Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft. **Francisco de Arriba-Pérez:** Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Supervision. **Silvia García-Méndez:** Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Supervision. **Andrea Busto-Castiñeira:** Methodology, Investigation, Writing – review & editing. **Francisco J. González-Castaño:** Conceptualization, Methodology, Validation, Formal analysis, Formal analysis, Investigation, Writing – original draft, Supervision.

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.

Acknowledgments

This work was partially supported by (i) Xunta de Galicia grants ED481B-2021-118, ED481B-2022-093, and ED431C 2022/04, Spain; (ii) Ministerio de Ciencia e Innovación grant TED2021-130824B-C21, Spain; and (iii) University of Vigo/CISUG for partial open access charge. Moreover, this work has been supported by E4Legal Analytics sl. The authors are indebted to Mrs.Elen Irazabal for her help.

References

Apley, D.W., Zhu, J., 2020. Visualizing the effects of predictor variables in black box supervised learning models. J. Roy. Stat. Soc.: Series B (Stat. Methodol.) 82, 1059–1086. https://doi.org/10.1111/rssb.12377.

Arriba-Pérez, F.D., García-Méndez, S., González-Castaño, F.J., González-González, J., 2022. Explainable machine learning multi-label classification of Spanish legal judgements. J. King Saud Univ.- Comput. Informat. Sci. 34, 10180–10192. https://doi.org/10.1016/j.jksuci.2022.10.015.

Arriba-Pérez, F.D., García-Méndez, S., Regueiro-Janeiro, J.A., González-Castaño, F.J., 2020. Detection of financial opportunities in micro-blogging data with a stacked classification system. IEEE Access 8, 215679–215690. https://doi.org/10.1109/ACCESS.2020.3041084.

Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V., Soria, C., 2004. Automatic Classification and Analysis of Provisions in Italian Legal Texts: A Case Study. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3292. pp. 593–604, https://doi.org/10.1007/978-3-540-30470-8_72.

Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: a pretrained language model for scientific text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing, Association for Computational Linguistics, pp. 3613–3618, https://doi.org/10.18653/v1/D19-1371.

Branting, L.K., Pfeifer, C., Brown, B., Ferro, L., Aberdeen, J., Weiss, B., Pfaff, M., Liao, B., 2021. Scalable and explainable legal prediction. Artif. Intell. Law 29, 213–238. https://doi.org/10.1007/s10506-020-09273-1.

Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: a survey on methods and metrics. Electronics 8, 1–34. https://doi.org/10.3390/electronics8080832.

Chalkidis, I., Kampas, D., 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artif. Intell. Law 27, 171–198. https://doi.org/10.1007/s10506-018-9238-9.

Chen, H., Wu, L., Chen, J., Lu, W., Ding, J., 2022. A comparative study of automated legal text classification using random forests and deep learning. Informat. Process. Manage. 59, 102798–102812. https://doi.org/10.1016/j.ipm.2021.102798.

Coltrinari, R., Antinori, A., Celli, F., 2020. Surviving the legal jungle: text classification of italian laws in extremely noisy conditions. In: Proceedings of the Italian Conference on Computational Linguistics, vol. 2769, Accademia University Press, pp. 122–127, https://doi.org/10.4000/books.aaccademia.8390.

Cousins, C., Riondato, M., 2019. CaDET: interpretable parametric conditional density estimation with decision trees and forests. Mach. Learn. 108, 1613–1634. https://doi.org/10.1007/s10994-019-05820-3.

Delgado-Panadero, A., Hernandez-Lorca, B., Garcia-Ordas, M., Benitez-Andrades, J., 2022. Implementing local-explainability in Gradient Boosting Trees: Feature Contribution. Inf. Sci. 589, 199–212. https://doi.org/10.1016/j.ins.2021.12.111.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, pp. 4171–4186.

Drobnic, F., Kos, A., Pustišek, M., 2020. On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. Electronics 9, 1–15. https://doi.org/10.3390/electronics9050761.

Dyevre, A., 2021a. Text-mining for lawyers: how machine learning techniques can advance our understanding of legal discourse. Erasmus Law Rev. 14, 7–23. https://doi.org/10.5553/ELR.000191.

Dyevre, A., 2021b. The promise and pitfall of automated text-scaling techniques for the analysis of jurisprudential change. Artif. Intell. Law 29, 239–269. https://doi.org/10.1007/s10506-020-09274-0.

Flores, A.M., Pavan, M.C., Paraboni, I., 2022. User profiling and satisfaction inference in public information access services. J. Intell. Informat. Syst. 58, 67–89. https://doi.org/10.1007/s10844-021-00661-w.

Forzieri, G., Girardello, M., Ceccherini, G., Spinoni, J., Feyen, L., Hartmann, H., Beck, P. S.A., Camps-Valls, G., Chirici, G., Mauri, A., Cescatti, A., 2021. Emergent vulnerability to climate-driven disturbances in European forests. Nat. Commun. 12, 1–12. https://doi.org/10.1038/s41467-021-21399-7.

Gambhir, M., Gupta, V., 2017. Recent automatic text summarization techniques: a survey. Artif. Intell. Rev. 47, 1–66. https://doi.org/10.1007/s10462-016-9475-9.

Lukasz Górski, Ramakrishna, S., 2021. Explainable artificial intelligence, lawyer's perspective. In: Proceedings of the International Conference on Artificial Intelligence and Law, ACM, pp. 60–68, https://doi.org/10.1145/3462757.3466145.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2019. A survey of methods for explaining black box models. ACM Comput. Surv. 51, 1–42. https://doi.org/10.1145/3236009.

Gunning, D., Aha, D., 2019. DARPA's explainable artificial intelligence (XAI) program. AI Mag. 40, 44–58. https://doi.org/10.1609/aimag.v40i2.2850.

Hacker, P., Krestel, R., Grundmann, S., Naumann, F., 2020. Explainable AI under contract and tort law: legal incentives and technical challenges. Artif. Intell. Law 28, 415–439. https://doi.org/10.1007/s10506-020-09260-6.

Hasal, M., Nowaková, J., Saghair, K.A., Abdulla, H., Snášel, V., Ogiela, L., 2021. Chatbots: Security, privacy, data protection, and social aspects. Concurr. Comput.: Practice Exp. 33, 1–13. https://doi.org/10.1002/cpe.6426.

Hatwell, J., Gaber, M.M., Azad, R.M.A., 2020. Ada-WHIPS: explaining AdaBoost classification with applications in the health sciences. BMC Med. Inform. Decis. Mak. 20, 250–274. https://doi.org/10.1186/s12911-020-01201-2.

Hausladen, C.I., Schubert, M.H., Ash, E., 2020. Text classification of ideological direction in judicial opinions. Int. Rev. Law Econ. 62, 105903–105921. https://doi.org/10.1016/j.irle.2020.105903.

Hettiarachchi, H., Adedoyin-Olowe, M., Bhogal, J., Gaber, M.M., 2022. Embed2Detect: temporally clustered embedded words for event detection in social media. Mach. Learn. 111, 49–87. https://doi.org/10.1007/s10994-021-05988-7.

Kastrati, Z., Dalipi, F., Imran, A.S., Nuci, K.P., Wani, M.A., 2021. Sentiment analysis of students' feedback with NLP and deep learning: a systematic mapping study. Appl. Sci. 11, 1–23. https://doi.org/10.3390/app11093986.

Kim, B.J., Koo, G., Choi, H., Kim, S.W., 2023a. Extending class activation mapping using Gaussian receptive field. Comput. Vis. Image Underst. 231, 103663–103669. https://doi.org/10.1016/j.cviu.2023.103663.

Kim, S.H., Park, J.S., Lee, H.S., Yoo, S.H., Oh, K.J., 2023b. Combining CNN and Grad-CAM for profitability and explainability of investment strategy: application to the KOSPI 200 futures. Expert Syst. Appl. 225, 120086–120098. https://doi.org/10.1016/j.eswa.2023.120086.

Kowsari, Meimandi, J., Heidarysafa, Mendu, Barnes, Brown, 2019. Text Classification Algorithms: A Survey. Information, 10, 1–68. https://doi.org/10.3390/info10040150.

Škrlj, B., Martinc, M., Lavrac, N., Pollak, S., 2021. autoBOT: evolving neuro-symbolic representations for explainable low resource text classification. Mach. Learn. 110, 989–1028. https://doi.org/10.1007/s10994-021-05968-x.

Lage, I., Ross, A.S., Kim, B., Gershman, S.J., Doshi-Velez, F., 2018. Human-in-the-loop interpretability prior. In: Advances in Neural Information Processing Systems, vol. 2018-December pp. 1–10.

Le, T.T., Moore, J.H., 2021. treeheatr: an R package for interpretable decision tree visualizations. Bioinformatics 37, 282–284. https://doi.org/10.1093/bioinformatics/btaa662.

Lee, H., Kim, S., 2016. Black-box classifier interpretation using decision tree and fuzzy logic-based classifier implementation. Int. J. Fuzzy Logic Intell. Syst. 16, 27–35. https://doi.org/10.5391/IJFIS.2016.16.1.27.

Lin, J., Ma, Z., Gomez, R., Nakamura, K., He, B., Li, G., 2020. A review on interactive reinforcement learning from human social feedback. IEEE Access 8, 120757–120765. https://doi.org/10.1109/ACCESS.2020.3006254.

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable AI: a review of machine learning interpretability methods. Entropy 23, 1–45. https://doi.org/10.3390/e23010018.

Mathew, A., Amudha, P., Sivakumari, S., 2021. Deep Learning Techniques: An Overview, vol. 1141. Springer. https://doi.org/10.1007/978-981-15-3383-9_54.

Medvedeva, M., Vols, M., Wieling, M., 2020. Using machine learning to predict decisions of the European Court of Human Rights. Artif. Intell. Law 28, 237–266. https://doi.org/10.1007/s10506-019-09255-y.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.-R., 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn. 65, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008.

Neto, M.P., Paulovich, F.V., 2021. Explainable matrix - visualization for global and local interpretability of random forest classification ensembles. IEEE Trans. Visual Comput. Graphics 27, 1427–1437. https://doi.org/10.1109/TVCG.2020.3030354.

Park, M., Chai, S., 2021. AI model for predicting legal judgments to improve accuracy and explainability of online privacy invasion cases. Appl. Sci. 11, 1–16. https://doi.org/10.3390/app112311080.

Qiu, M., Zhang, Y., Ma, T., Wu, Q., Jin, F., 2020. Convolutional-neural-network-based multilabel text classification for automatic discrimination of legal documents. Sensors Mater. 32, 2659–2672. https://doi.org/10.18494/SAM.2020.2794.

Rana, P., Varshney, L.R., 2021. Trustworthy predictive algorithms for complex forest system decision-making. Front. Forests Global Change 3, 1–15. https://doi.org/10.3389/ffgc.2020.587178.

Rustamov, S., Bayramova, A., Alasgarov, E., 2021. Development of dialogue management system for banking services. Appl. Sci. 11, 1–18. https://doi.org/10.3390/app112210995.

Sagi, O., Rokach, L., 2020. Explainable decision forest: Transforming a decision forest into an interpretable tree. Informat. Fus. 61, 124–138. https://doi.org/10.1016/j.inffus.2020.03.013.

Schweighofer, E., Rauber, A., Dittenbach, M., 2001. Automatic text representation, classification and labeling in European law. In: Proceedings of the International Conference on Artificial Intelligence and Law, pp. 78–87. https://doi.org/10.1145/383535.383544.

Shook, J., Smith, R., Antonio, A., 2018. Transparency and fairness in machine learning applications. Texas A&M J. Property Law 4, 443–463. https://doi.org/10.37419/JPL.V4.I5.2.

Song, D., Vold, A., Madan, K., Schilder, F., 2022. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. Informat. Syst. 106, 101718–101729. https://doi.org/10.1016/j.is.2021.101718.

Özge Sürer, Apley, D.W., Malthouse, E.C., 2021. Coefficient tree regression: fast, accurate and interpretable predictive modeling. Mach. Learn. 1–37. https://doi.org/10.1007/s10994-021-06091-7.

Tagarelli, A., Simeri, A., 2021. Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. Artif. Intell. Law, 417–473. https://doi.org/10.1007/s10506-021-09301-8.

Tandra, S., Manashty, A., 2021. Probabilistic Feature Selection for Interpretable Random Forest Model, vol. 1364. Springer. https://doi.org/10.1007/978-3-030-73103-8_50.

Tao, Y., Zhang, F., Shi, C., Chen, Y., 2019. Social media data-based sentiment analysis of tourists' air quality perceptions. Sustainability 11, 1–23. https://doi.org/10.3390/su11185070.

Tellez, E.S., Moctezuma, D., Miranda-Jiménez, S., Graff, M., 2018. An automated text categorization framework based on hyperparameter optimization. Knowl.-Based Syst. 149, 110–123. https://doi.org/10.1016/j.knosys.2018.03.003.

Thomas, A., Sangeetha, S., 2021. Semi-supervised, knowledge-integrated pattern learning approach for fact extraction from judicial text. Expert Syst. 38, 1–20. https://doi.org/10.1111/exsy.12656.

Thompson, P., 2001. Automatic categorization of case law. In: Proceedings of the International Conference on Artificial intelligence and Law, ACM Press. pp. 70–77, https://doi.org/10.1145/383535.383543.

Trappey, A.J., Trappey, C.V., Wu, J.-L., Wang, J.W., 2020. Intelligent compilation of patent summaries using machine learning and natural language processing techniques. Adv. Eng. Inform. 43, 101027–101039. https://doi.org/10.1016/j.aei.2019.101027.

Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual explanations without opening the black box: automated decisions and the GDPR. SSRN Electronic J., 841–887 https://doi.org/10.2139/ssrn.3063289.

Wells, L., Bednarz, T., 2021. Explainable AI and reinforcement learning—A systematic review of current approaches and trends. Front. Artif. Intell. 4, 1–15. https://doi.org/10.3389/frai.2021.550030.

Xu, D., Quan, W., Zhou, H., Sun, D., Baker, J.S., Gu, Y., 2022. Explaining the differences of gait patterns between high and low-mileage runners with machine learning. Sci. Rep. 12, 1–12. https://doi.org/10.1038/s41598-022-07054-1.

Zanzotto, F.M., 2019. Viewpoint: human-in-the-loop artificial intelligence. J. Artif. Intell. Res. 64, 243–252. https://doi.org/10.1613/jair.1.11345.