ISSN: 3049-7159

Research Article

Journal of Business and Econometrics Studies

The FAO Experience in Implementing an Automatic Tool to Classify Legal Documents

Marco Scarno[1*] and Andres Vatter Rubio[2]

[1]*Senior Statistician and Data Scientist, Italy*
[2]*FAO Legal Office, Italy*

[*]**Corresponding author**
Marco Scarnò, Senior Statistician and Data Scientist, Italy.

Received: May 19, 2025; **Accepted:** May 27, 2025; **Published:** June 02, 2025

ABSTRACT

This article describes the development and implementation of an automatic tagging tool by the Food and Agriculture Organization (FAO) to extract subject matter metadata from legal documents in the FAOLEX database. The implemented approach started by testing various learning models/methods, comparing their accuracies and, then, selected the Decision Tree as the most performing candidate for the production tool, testing and discussing its performance in a real environment, where new documents are added regularly and where classes may change or be introduced. The article shows how the system can improve over time and provide valuable support to the FAO experts in their document classification and quality check tasks, concluding by highlighting the significance of leveraging AI technologies for efficient document classification and retrieval in the legal domain, and by addressing some of the methodological and managerial implications of adopting AI solutions within organizations.

Keywords: AI, legal Document Classification, NLP Methods, Multilabel Classification

Introduction

Adopting AI solutions within private or public organizations involves three key roles: domain experts, data scientists, and managers. Domain experts submit their problems to data scientists, who develop methods that can facilitate the production process. Managers must allocate money to implement the identified solutions and allocate the time freed up by these solutions. Automatic text processing has the potential to save human resources time, as machines can discern patterns and main concepts from documents, thus eliminating the need for manual reading. However, machines could supply irrelevant information, making it necessary the assessment of the extracted patterns. These considerations suggest that the time humans save in routine activities should be redirected to monitor AI efficacy, demanding additional training for the workforce, who must become skilled at interfacing with the machines and enhancing their operational capabilities.

It is therefore important that managers implement appropriate decisions to encourage the use of new methodologies but, also, supporting the changing responsibilities and job specifications. In this context, the work done by the Food and Agriculture Organization (FAO) on developing a method for the automatic extraction of patterns from a set of legislative documents is noteworthy. This paper presents a recently developed automatic tagging tool to extract FAOLEX subject matter metadata from documents in the FAOLEX database. The first two paragraphs of this paper will illustrate the background and the methodological context on which this work is based. Subsequently, the materials, methods, and the implemented approach will be introduced, followed by the paragraph in which the results will be presented. These results will then be further discussed, leading to the final considerations and remarks.

The Background

With over 200,000 documents, FAOLEX is the largest open access database of regional, national and sub-national laws, regulations and policies on food, agriculture and natural resources management. Users of FAOLEX have direct access to the full text of the documents, as well as summaries and extensive metadata relating to bibliographic, temporal, geopolitical, and subject matter facets. From its inception in 1995 until 2014, the FAOLEX database encompassed two subject

Citation: Marco Scarno, Andres Vatter Rubio. The FAO Experience in Implementing an Automatic Tool to Classify Legal Documents. J Bus Econo Stud. 2025. 2(3): 1-8. DOI: doi.org/10.61440/JBES.2025.v2.66

matter classes: domains and keywords. Domains, numbering 16, aligned with historical thematic areas of FAO, UNEP, and IUCN's collaborative efforts, as reflected in the common portal ECOLEX. Instruments often cover overlapping themes that span multiple domains. Keywords, exceeding 440 in number, form a controlled vocabulary used to describe document content more precisely. At the data entry stage, keywords are associated with specific domains, resulting in a two-dimensional subject matter classification.

This initial setup posed challenges in promptly organizing and filtering the dataset from sectoral perspectives. Laws often adopt a sectoral legislative structure. Different sectors have unique characteristics, challenges, and goals. Sectoral organization permits specialized focus on distinct areas, fostering expertise and informed decision-making. As such, in 2014, a third subject matter class, Primary Subject, was introduced, allowing for the filtering documents across sectoral clusters. This change was accompanied by changes in scope of the collection as well. Primary Subject incorporated 16 subject classes from the Domains list and introduced 9 new subjects, particularly emphasizing social development and legislation that reflected the wider enabling environment. This expansion aligned with the Sustainable Development Goals and the FAO strategic framework of the 4 Betters: Better Production, Better Environment, Better Nutrition, and Better Livelihoods. Concomitantly, official policies were incorporated into the collection in recognition of the important synergies and co-dependence between policy and law.

In 2022, FAOLEX introduced a fourth subject matter metadata class: primary keyword(s); these are a subset of the original keyword's vocabulary list; they number over a 100, representing the most relevant terms used to classify content. The selection criterion for choosing the primary keywords is evidence-based, drawing from FAOLEX the concepts that are the main topic of legislative instruments and policies. Acknowledging the inherent diversity within each sector or primary subject, primary keywords offer easy accessibility across jurisdictions, aiding users in navigating similar instruments. Over 11,000 major records, comprising legislation and policies governing specific areas, have been manually assigned with primary keywords. However, this metadata field is not yet available to end-users, as the remaining 190,000 documents need retrospective enrichment.

The process of collecting, analysing and entering new records in FAOLEX is manual and time-consuming. A rough estimate evidenced that the entry of a new record can take between 30 minutes to three hours, depending on the length and relevance of the text. This is a significant investment of human resources in an activity that could be performed partially or in full by a machine. This time constitutes a significant investment of human resources in an activity that could be performed by a machine, provided it can deliver accurate results. To reach this objective, there should be a willingness to invest in additional human resources that will study and develop the automated process, covering the costs of a proper IT infrastructure. Moreover, due to continuous technological advancements and the specificity of the problem, these costs would not only be faced in the initial phase but would be more or less constant over time. Furthermore, even considering the complete automation of a similar process, it would still be unreasonable to assume the total replacement of human judgment by a machine, as this would then lack the ability to evaluate the effectiveness of the method over the long term.

Due to these premises, the approach that was followed in developing an automatic classifier tried to minimize costs by reducing R&D time, infrastructure costs, and ongoing maintenance. This was done while ensuring that the resulting performance could remain on par with the most recent studies on similar subjects found in academic literature.

The Methodological Context

From a methodological perspective, the task here described involves creating a classification system for textual data based on a supervised approach, due to the possibility of taking advantages of prior knowledge of the classes to which texts were assigned. Delving deeply into the methodological aspect, the system must deal with multi-label classifications, because classes (domain, primary subject, primary and simple keywords) assume nonexclusive labels. For instance, almost nine are the average keywords that could be associated to a document. Lastly, it must be noted that the labels are not balanced (many of them are rare) and, as for the keywords, these are selected from a list of more than 400 available options.

Text classification is based on learning and extending relationships between terms in the texts and classes that are present or expected. These relationships can be identified through a machine learning approach, which results in a mathematical function that provides the classes' scores as output given an input (for instance, the text of a document). In recent years these relationships are estimated with more sophisticated and complex algorithms named as Deep Learning, in which are intermediating other layers in the flow between the input and the output ones. These layers have the scope to identify logical structures that could improve the resulting classification. As an example, these intermediate layers could automatically recognize the main topic of a document. The actual evolution of Deep Learning brought nowadays to Language Models, which consist of enriching the relationships between texts and classes with the most appropriate sequence of words derived from having studied a considerable number of documents.

Nowadays there are numerous studies on automatic classifiers for textual data; Kowasari provides a detailed list of these approaches [1]. The Food and Agriculture Organization (FAO) has already conducted several studies on the potential application of automatic document classification methods, starting from Lauser, in which it is presented an approach to use binary support vector machines (SVM) for automatic subject indexing of full-text documents with multiple labels [2].

The first study on a classifier for legal documents was conducted in 1997 [3]. In recent years, the number of such studies has increased due to the availability of many open legal data sets and more advanced hardware architectures that allow for testing sophisticated methods. Martins and Silva provide a review of the trends in classifying legal documents for Brazilian Portuguese studies [4]. They found that Deep Learning is referred to in more than 50% of cases, followed by standard learning approaches and then Language Models. It is worth noting that their work

is dated 2021, two years before the emergence of ChatGPT (a Large Language Model) as a new paradigm for treating texts and extracting their latent characteristics.

Despite the trends, there is no full convergence between the numerous studies in identifying the most effective approach; Chen found that Random Forest, a specific machine learning method, outperformed Deep Learning [5]. On the other hand, Chalkidis found more accurate this last, and in their results the increase in accuracy is 12% when passing from a logistic regression to a DL method [6]. However, logistic regression is not often considered the most performing method and, moreover, the metric used to evaluate the various classification results was not the same between different studies.

Indeed, the greater accessibility to complex computational resources or to pretrained language models constitutes the main trending factor that are behind the use of Deep Learning based methods; on the other hand, it has to be noted that their costs for realizing systems to be used in a production environment should be carefully evaluated, due to need of powerful machines to apply classification methods to new documents.

Materials and Methods

FAOLEX is a database of national legislation, policies and bilateral agreements on food, agriculture, and natural resources management. It is constantly being updated, with an average of 8,000 new entries per year. It contains legal and policy documents drawn from more than 200 countries, territories and regional economic integration organizations and originating in over 40 languages (almost 30% English, 24% Spanish, 14% French and others). Before inserting a new document in the database, several metadata are manually assigned to it, like the Domain, the Primary subject, the Keywords, from which the primary one is extracted. All these metadata have the scope to facilitate the users searches.

The primary goal of this study is to assist users in determining the values of the metadata, which can have multiple, non-exclusive classes. Table 1 outlines their key features, based on 201,203 documents as of November 2023, with the exception of Primary Keywords, which have only been assigned to 11,192 documents.

Table 1: figures on metadata values as in the FAOLEX database

Object	Max number of classes	Average number of distinct classes in a document (and standard deviation)
Domain	16	1.4 (0.99)
Primary subject	24	1.1 (0.34)
Keywords	446	8.5 (7.9)
Main Classifying Keywords	108	1.1

Metadata classes are not balanced; for instance, “Food and nutrition” is the most common Domain (18%), while “Air & atmosphere” is the rarest one (3%). Concerning the Keywords, the most common is “Institution” (22%), the rarest “Novel food” (0.07%). The distribution of these classes (%) is in figure 1, 2, 3 and 4.

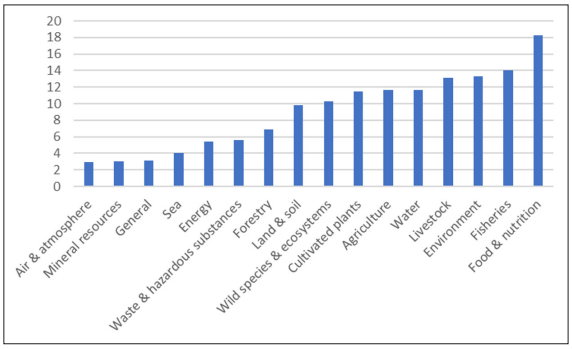

Figure 1: distribution (%) of classes for Domains in FAOLEX documents

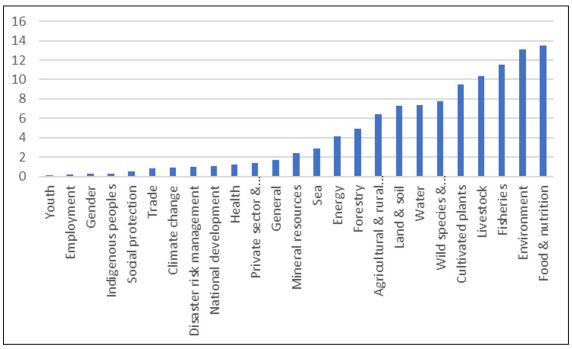

Figure 2: distribution (%) of classes for Primary Subjects in FAOLEX documents

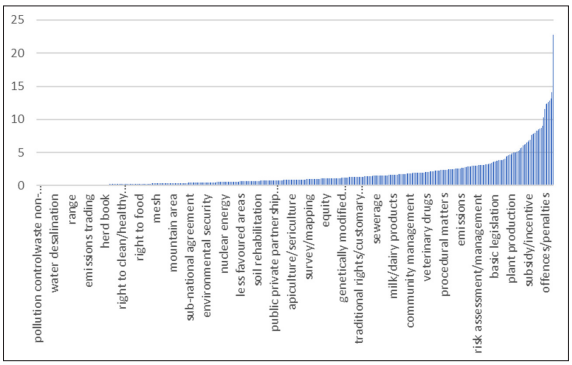

Figure 3: distribution (%) of classes for Keywords in FAOLEX documents

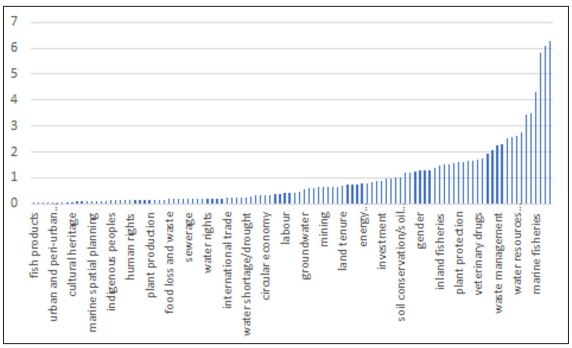

Figure 4: distribution (%) of classes for Primary Keywords in FAOLEX documents

Unbalanced classes can be challenging from a methodological perspective since an automatic system may be biased towards those more frequent, while the rarer ones could require greater effort to be identified by FAOLEX experts. An appropriate evaluation measure is necessary to assess the performance of a classification system, allowing for the testing of its effectiveness on various subsets of documents, other than the one used for training the model itself.

Evaluation measures for binary cases are the most common, because of the possibility to rely on the confusion matrix (introduced since 1904 by K. Pearson), in which concordances or discordances between real and predicted classes are counted [7]. From such representation it is easy to derive many indexes, from the simpler ones, like the success or the error rates, to the most articulated, like the True positive rate, the Chi Square, etc. See Fawcett or, Sokolova for a review of the available indicators) [8,9].

Multilabel classification results are more difficult to be reviewed; this because the number of possible and nonexclusive classes and the difference between the number of real (sometime defined as "gold labels") and of predicted classes. As an example, a document can be associated to three classes, but the automatic system found for it just two, on which only one is corrected. In this case an evaluation measure will consider the concordances (number of correctly identified classes, i.e. 1), but that could be referred alternatively to the number of real or of predicted (or their mixture) classes.

In this study we will privilege the average rate of concordances, or accuracy, defined as:

$$R = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{gl_i}$$

Where n is the number of documents (for instance those in a test set) and, for the *i-th* document:

- c_i is the number of concordances
- gl_i is the number of gold labels

Chalkidis et al. in their works (2019, cited) question on how R could penalize those classification systems in which the number of gold labels differs significantly from those ranked as top p by the automatic models. For this reason, they propose a different measure:

$$R' = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{\min(gl_i, p_i)}$$

In this case the *min* is referred to the minimum number of classes between the real (gl_i) and those predicted (p_i). To make our results comparable, and where it makes sense, we will also report this measure.

Before delving into the details of the classification methods that were tested and implemented in this study, it is pertinent to mention the document processing flow:

- Starting from a PDF file, its text was extracted by means of the Apache Tika library;
- The resulting text was treated, deleting multiple spaces, new lines, punctuation marks
- The language of the document was detected and not English texts were translated in English with the Argos Translate Python Library (covering more than 30 not EN languages)
- The text passed through a lemmatization step (a linguistic process able to group together the inflected forms of a word so they can be analysed as a single item), to retain the lemmas, if more than 2 characters, of the following part of speech: Adjectives, Adverbs, Nouns, Proper nouns, Verbs

Original texts, selected lemmas, basic metadata (titles, original file names, file dates, etc.) were then stored in a NoSQL database (Apache Solr).

On top of this flow a specific web application was realized, called Essence, that permits users to interact with documents, metadata, by easily uploading or retrieving them [10].

All the document processing flow, the web application and the automatic classification system were implemented or installed on a Virtual Machine hosted by Google; in particular, on a first-generation Compute Engine, with 4 CPU and 15GB of RAM. A similar machine, but with an additional external disk of 500GB, is the one on which it is installed Solr, working as text repository for this and for other projects. Considering that users evaluate daily only several dozens of documents, the IT configuration dedicated to the project can be deemed quite economical, particularly in scenarios where frequent usage is required.

The following steps summarizes the methodological approach that has been followed to implement the classification system object of this study:

- different classification models/methods were tested and the best performing one was selected, checking its accuracy gain when adding new documents; such task considered only those documents having at least one Primary Keyword.
- The selected model/method identified above was trained on the Domains, Primary Subjects and Keywords

Moreover, a further aspect that this study addresses refers to the periodically retrain of the models/methods, considering those new documents inserted by the users and their final decisions on the gold labels. In fact, it may be possible to identify a model that can provide satisfactory accuracies when trained, but it may exhibit degrading performance when used daily on new documents. The reasons behind this behaviour can be many, as a partial representative training set not able to manage the trends in the classes belonging to new documents. For instance, it is possible to have trained a model in which the keyword "antimicrobial resistance" was initially present in just a few cases, but it will be more frequent as new laws will be issued. In this case, this class will be better identified by adding more related documents in the training set.

Results

An automatic classification system requires an IT infrastructure to handle the estimation of models/methods, as well as their daily application to new documents. The cost of this infrastructure can vary widely, depending on factors such as the amount of RAM, the number and speed of CPUs, the disk space. The use of GPUs,

which are processors used to implement Deep Learning or Language Models, can also affect the cost. In this study, classic machine learning approaches were used to avoid the need for complex and expensive architectures that would require GPUs.

These approaches learn the relationships between terms in the documents (the n-grams, composed of single or pairs of chosen lemmas) and the classes in a training set of documents. These relationships are estimated using the term frequency-inverse document frequency (TF-IDF) matrix, which is a mathematical representation of the significance of a term in a collection or corpus of documents.

In particular, the following learning approaches were considered (as in the SKLEARN Python library):

- Parametric: Linear Logistic Regression, Ridge Classifier, both applied in a multi label context, i.e. with a model for each class.
- Non-parametric: Decision Trees (normal or randomizing the decision trees as in an extra-tree approach or in a Random Forest Classifier), K-nearest Neighbours, Multi-Layer Perceptron.

Non-parametric methods have the potential to fit a wider range of possible relationships, but their complexity should deal with the overfitting issue, which indicates the perfect capability of a method to reproduce the training data, but that fails to fit new documents in a reliable manner. Vapnik and Chervonenkis examined the relationships between the size of the training set, the complexity of a classification method and the generalization error [11].

They introduced the concept of VC dimension to represent the complexity of a method, and identified the mathematical function that links it to the size of the training set to reduce the generalization error. This function could help avoid the risk of overfitting, but it is only valid when the complexity is strictly lower than the dimension of the training set. Instead, in case of a neural network (as for the others non-parametric models that were considered), the VC dimension is in the order of the number of parameters that need to be estimated. Since the number of parameters is likely higher than the number of n-grams extracted from the documents (these could be millions), it is easy to see that such a task is potentially affected by overfitting.

To determine the most effective approach, all documents with at least one Primary Keyword were divided into two sets. The division was done randomly, while ensuring that the representativeness criterion for all classes to be estimated was respected, without oversampling the more frequent ones. As a result, the training set consisted of 40% of the initial documents.

The accuracy of the different approaches, as evaluated on the test set, is presented in Table 2, which also includes the time required to complete the estimation and generalization steps.

Out of the three most accurate methods, K-nearest Neighbors (KNN), Ridge Classifier (RG), and Decision Tree (DT), the DT method was chosen. This is because KNN took too long for training and generalization, and RG did not provide probabilities for each predicted class (at least in the release of the library that was considered, which are useful for users to better revise the estimated classes.

Table 2: accuracies on the test set (60% of documents with at least one Primary Keyword) and time required to complete the estimation and generalization steps

Method	Accuracy (R)	Time (hh:mm)
Linear logistic regression	0.33	01:19
Ridge classifier	0.59	00:27
Decision tree	0.61	00:46
Extra trees classifier	0.29	02:50
K-nearest Neighbors	0.62	27:14
Random Forest Classifier	0.35	02:52
Multi-Layer Perceptron	0.41	08:00

Each method tested has its own set of hyperparameters that could be adjusted to improve the fit to the training data. However, we chosen to not adjust them to avoid the need for constant fine-tuning when the training set or the objects being considered change. Cross-validation was also avoided due to the high computational resources required, which is not compatible with a streamlined approach. The focus was on identifying the most reliable method, which required using the same training and test sets, a condition that would not be met in the case of cross-validation.

The second step of our approach was to verify the gain in accuracies that could derive when increasing the size of the training set. As introduced, this is important to test the feasibility of the method in incorporating changes in classes, new trends in legislation documents, etc. In figure 5 are the accuracies (R) evaluated by considering the DT on the Primary Keyword by varying the size of the training set with a fixed test set.

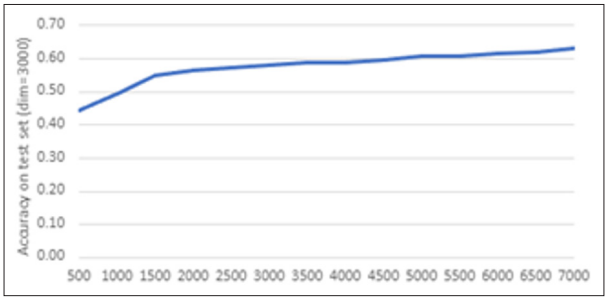

Figure 5: accuracies (R) on a random test set of documents (3000) for Primary Keyword for different sizes of traiRning sets

As the size of the training set increases, the gain in accuracies becomes more evident, which confirms the reliability of the Decision Tree in dealing with the complexity of the task without overfitting.

This result is valid in a theoretical situation where the classification method is tested with documents that have already been evaluated. However, the aim of this project is to develop a tool that can support users in a real environment where new documents are added daily and where new classes could be introduced (as for the "Coronavirus disease") or revised. In such a context, the selected method needs to be constantly evaluated

and retrained when its accuracy decreases or when substantial changes occur.

In this optic, the approach proceeded with the training of the selected method (DT) for the different objects, i.e. for the Primary subjects, Domains, Keywords. Different strategies of extracting a training set from the more two hundred thousand of documents were followed, considering that the long-tailed distribution and the complex linkage (co-occurrence) of Keywords' classes, where a small subset of these (namely head classes) have many instances, while a majority (namely tail labels) have only a few instances.

To deal with these conditions we concentrated on the Primary subject, because characterized by few classes (24), but for which the keywords could be also linked. In a first step we started by randomly selecting half of the documents for each distinct class, with a partial oversampling of the rarest cases. We then executed the learning/test steps, verifying the results in terms of the predicting accuracy for the Keywords, adding progressively new documents for those classes having a high error rate (more than 90%) and few documents in the training set.

After having iterated the above steps for few times, and with the objective to have almost one hundred thousand of documents in the learning set, we proceeded to the final train of the method. Figure 6 shows the resulting classes' compositions for Primary subjects, considering their original distribution and their relative frequencies in the training set. So, for instance, almost all documents of the rarest class, Employment, were considered in the training set.

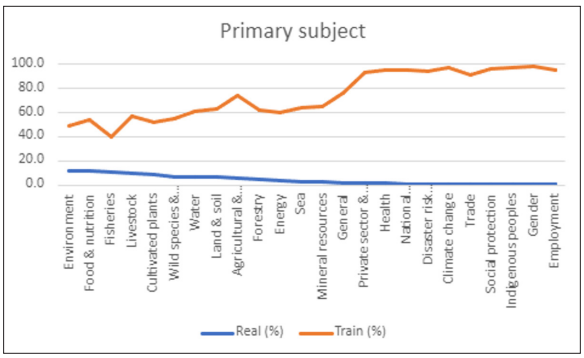

Figure 11: resulting composition of Primary subject' classes in the training set

The resulting accuracies (R and R') referred to the documents not used for the training are in table 3.

Table 3: accuracies (R and R') on the test set (almost 100 thousand documents)

Model for:	R	R'
Domain	0.59	0.73
Primary subject	0.60	0.66
Keywords	0.36	0.38

Considering these results as satisfactory, we made available the tool to the FAOLEX users, letting them add new documents and using the estimated classes as a support in identifying the metadata.

Obviously, the final values of the classes that they could have selected were different from those provided by the tool, so that it was possible to further monitor the real accuracies and to periodically retrain the method by adding to the original training set the new documents inserted. The details on the accuracies before and after these retrains are in table 4, referring to almost 3300 new documents inserted in 6 months.

Table 4: accuracies (R and R') as observed on the new inserted documents, that were further added in the training set previously used.

Model for:	Accuracies on 1586 new documents		Accuracies on 1075 new documents (after having added the 1586 to the original train set)		Accuracies on 726 new documents (after having added the 1075 to the revised train set)	
	R	R'	R	R'	R	R'
Domain	0.56	0.7	0.51	0.65	0.6	0.74
Primary Subject	0.58	0.6	0.57	0.59	0.61	0.68
Keywords	0.35	0.37	0.35	0.38	0.39	0.43

In table 5 are, instead, the accuracies for the Main Classifying Keywords, considering that for them the number of documents that had at least one of these associated, in the 6 months, were less than those considered for the other objects.

Table 5: accuracies (R and R') for the Main Classifying Keywords as observed on the new inserted documents, that were further added in the training set previously used.

Model for:	Accuracies on 513 new documents		Accuracies on 752 new documents (after having added the 513 to the original train set)		Accuracies on 726 new documents (after having added the 752 to the revised train set)	
	R	R'	R	R'	R	R'
Main Classifying Keywords	0.39	0.42	0.34	0.37	0.47	0.51

It must be noted that, between the first and the second period here considered, there was a revision of the classes (Primary subjects, Keywords and Main Classifying Keywords); for instance, the value “youth” was added to these.

Discussion of the Results

Before to introduce specific considerations on the performances of the proposed tool, it is worth to note that the general accuracies of the methods that we considered are aligned with results obtained by other authors. For instance, in their recent work, Haihua Chen et al. found that Random Forest performs better than Logistic regression, with an accuracy of 0.5 when classifying legal documents in 50 classes (in a not multi-label case) [12]. The poor performances of Logistic Regression in respect of other methods were noted also by Undavia et al. [13].

Few are the authors that considered the Decision Trees in their research; according to Kowsari [2019, cited] such method is fast for both learning and prediction, despite it could be sensitive for small perturbations in the data (as stated by Giovanelli) [14]. Decision tree was found to be accurate in both preliminary trainings and their subsequent refinements, even when new documents were added. The method showed the capability to improve in a dynamic process, where classes could not be well represented in the original training data but became more frequent, or when new ones are introduced.

Obviously, accuracies depend on the number of possible classes; Domain and Primary Subject are characterized by relatively few distinct values (16 Domains and 24 Primary Subjects) and results showed that in more than the half of cases the estimated classes are corrected. Moreover, the resulting accuracies after few retrains are higher than those evaluated on the test set for the first train. In particular, the initial accuracies for Domains were 0.59 (R) and 0.73 (R'), that passed to 0.6 (R) and 0.74 (R'). For Primary Subjects, they were 0.6 (R) and 0.66 (R’), passing to 0.61 (R) and 0.68 (R'). The difference between the two accuracy measures, R and R’, indicates that real classes could have usually more values than the estimated ones (table 6).

Table 6: statistics on the occurrences of the main metadata associated to the FAOLEX documents

Object	Number of possible classes	Average number of distinct classes in a document	Average number of predicted classes
Domain	16	1.4	1.3
Primary subject	24	1.1	1.1
Keywords	446	8.5	7.8
Main Classifying Keywords	108	1.1	1

Keywords, instead, are still not well recognized by the method; this is strictly related to the not representativeness of the training set but, also, to the poor discriminant capabilities of the classes themselves. In fact, there are classes like “civil code", “model law” whose concepts are wider and not easy to be correctly identified.

Concerning the Main Keywords, the method still needs to be further trained, but this again depends on the poor representativeness of the training set.

Finally, it must be noted that the resulting models are complex from the point of view of the dimension of their input data. It must be considered that the different n-grams extracted by all the documents in the training sets are more than three million, thus leading to a complex representation of the different trees. We are studying the possibility to introduce an automatic pruning of the resulting nodes by following, for instance, what in Quinlan [15].

Final Consideration and Remarks

There are different levels of interpretation that derive from having developed and implemented a tool like the one here described.

The first one concerns data scientists, aimed to create a simple yet effective process to assist users in their daily tasks. Instead of striving for a highly accurate solution that would require significant investment in infrastructure, testing, and calibration, their goal was to develop a multi-layered system that could be easily improved and fine-tuned. This included a NoSQL database to store documents, metadata, and predicted classes, a PDF parser, translators, and a predictive step that could easily access information about models and terms from external files. Additionally, the system allowed for easy retraining when new documents were added and included constant performance checks to ensure that the current method could be easily replaced if its accuracy decreased significantly. For example, the Decision Tree could be easily replaced with a Ridge Classifier Concerning the users of the tool (the domain experts, beneficiaries of an AI solution), we noticed a progressively acceptance of what was proposed due to the possibility to verify its improvements, knowing that they are fundamental for it.

While the automatic classifier may not be able to fully replace human experts, it has demonstrated its ability to improve over time. Additionally, its predictions have proven valuable in assisting the FAOLEX team with their monthly quality checks of document classifications. This because, at the end of each month, an expert reviews the classifications made by others and suggests any necessary changes. This tool can reduce the number of documents that need to be reviewed, as any document with matching real and predicted classifications can be considered implicitly correct. With an average accuracy of 50%, it is estimated that half of the documents can be excluded from the quality check.

The final level of interpretation is aimed at the organization's managers, who are in charge for handling the costs and benefits of implementing an AI solution. This is because a specific solution not only requires time to be researched, developed, and implemented, but also needs ongoing maintenance and improvement. As a result, the investment should be linked to a return on investment (ROI) of at least three years, i.e. to the average period in which new developments in research, external services, or infrastructure could alter the projected costs and benefits. For instance, systems based on LLM (such as ChatGPT) have recently emerged, providing opportunities like those presented here. However, such services are associated with

costs that are significantly higher than those used to develop the tool that was presented.

In the future, the reduction of their cost may represent an incentive for their use, despite possible issues in managing sensitive contents or the rights to use the results provided by external companies.

References

1. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, et al. Text Classification Algorithms: A Survey. Information. 2019. 10: 150.
2. Lauser B, Hotho A. Automatic multi-label subject indexing in a multilingual environment, in Proceedings of the 7th European Conference in Research and Advanced Technology for Digital Libraries (ECDL 2003), Trondheim (Norway). 2003
3. Curran T, Thompson P. Automatic Categorization of Statute Documents. 8th ASIS SIG/CR Classification Research Workshop.1997. 19-30.
4. Martins V, Silva C. Text Classification in Law Area: a Systematic Review. In Anais do IX Symposium on Knowledge Discovery, Mining and Learning. 2021. 33-40.
5. Chen H, Wu L, Chen J, Lu W, Ding J. A comparative study of automated legal text classification using random forests and deep learning. Information Processing & Management. 2022. 59
6. Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Ion Androutsopoulos. Large-Scale Multi-Label Text Classification on EU Legislation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence-Italy, Association for Computational Linguistics. 2019. 6314-6322
7. Pearson K. Mathematical Contributions to the Theory of Evolution. XIII. On the Theory of Contingency and Its Relation to Association and Normal Correlation. Dulau and Co., London. 1904. 426
8. Fawcett Tom. An Introduction to ROC Analysis (PDF). Pattern Recognition Letters. 2006. 27: 861-874.
9. Sokolova M, Guy Lapalme. A systematic analysis of performance measures for classification tasks, Information Processing & Management. 2009. 45: 427-437.
10. Fabi C, Scarnò M, Craig Steforn Matadeen. Essence - an integrated framework for documents retrieving and analysis. 2023.
11. Vapnik VN, Chervonenkis, Ya A. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. Theory of Probability & Its Applications. 1971. 16: 264.
12. Haihua Chen, Lei Wu, Jiangping Chen, Wei Lu, Junhua Ding. A comparative study of automated legal text classification using random forests and deep learning, in Information Processing & Management. 2022. 59: 102798.
13. Undavia S, Meyers A, Ortega JE. A Comparative Study of Classifying Legal Documents with Neural Networks, 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), Poznan, Poland, 2018. 515-522.
14. Giovanelli C, Liu X, Sierla S, Vyatkin V, Ichise R. Towards an aggregator that exploits big data to bid on frequency containment reserve market. In Proceedings of the 43rd Annual Conference of the IEEE Industrial Electronics Society (IECON 2017), Beijing, China, 29 October–1 November. 2017. 7514-7519.
15. Quinlan JR. Simplifying decision trees. Int. J. Man-Mach. Stud. 1987. 27: 221-234.