Organizing the Unorganized: A Novel Approach for Transferring a Taxonomy of Labels into Flat-Labeled Document Collections

Michele Colombino*,†, Laurentiu Jr Marius Zaharia*,†, Giorgia Iacobellis*,†, Rachele Mignone, Ivan Spada, Chiara Bonfanti, Emilio Sulis, Luigi Di Caro and Guido Boella

Computer Science Department - University of Turin, Via Pessinetto 12, 10149, Torino, Italy

Abstract

This paper presents a novel pipeline for transforming flat-labeled text collections into a hierarchical structure, which involves leveraging simple yet effective similarity methods that account for both lexical and semantic criteria to associate labels from disparate sources. Our approach employs a custom similarity measure, the Reinforced Edit Similarity, to identify probable correspondences based on lexical similarities. A subsequent semantic alignment and validation phase is then performed using an automatic classification mechanism. Preliminary results attest to the effectiveness of our proposal. These results are obtained from the research group of the University of Torino in the NGUPP project.

Keywords

Legal informatics, Legal document classification, Legal taxonomies, Taxonomy alignment, Text embeddings

1. Introduction

Legal informatics concerns the automatic processing of information to support legal activities. One of the most relevant issues encountered in the legal informatics field, and contextually in the judiciary, is the absence of standard criteria for the classification and analysis of legal documents [1, 2, 3]. In a juridical system, courts typically specialize in issuing specific types of judgments based on the cases they handle most frequently. Consequently, courts focus their attention on a set of specific subjects (e.g. civil criminal, family rights, labour, immigration...), for which a deeper and more granular organization of the judgements' labels is noted. This implies, on the one hand, the difficulty of having a national structure of such labels. On the other hand, it appears to be a precise layering of labels on more in-depth topics. As a consequence, such structures used by courts that are close to each other, both spatially and thematically, are often designed and organized by the competent authorities in different ways. This difficulty also affects the identification of judgments concerning similar topics but from different courts, as the judgments are categorized differently, with no explicit correspondence between the label organizations of different courts. A change to the organizational structure can lead to an enormous optimization, as smaller courts, which are used to handle a small number of cases each year, can take advantage of a more comprehensive hierarchical organization of labels from the larger courts. In addition, the elements available to assess the affiliation of a judgment in a certain category, or sub-category, applied by experts in the legal domain, derive partly from their direct experience, and partly by recognising them within the content of the judgements themselves.

This paper investigates the classification and analysis of legal documents by describing a taxonomy alignment pipeline, focusing on judgements and classification headings from two different legal sources. Moreover, we aim at defining a single, shared hierarchy of subject classification labels. This alignment will then be used to classify the available judgements with machine learning models. Exploiting a hierarchy from a digital archive of public judgements, the alignment work will be followed by the transfer of a set of unstructured labels into a well-defined taxonomy, producing an alignment, at first purely lexical, then semantic, by applying a simple criterion of label approximation that will be discussed in more detail in the following paragraphs. Techniques for extracting information from judgments will also be analyzed in order to identify patterns and keywords for implementing more accurate content-based classification. We focus

Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), June 23, 2023, Braga, Portugal.

*Corresponding author.

†These authors contributed equally.

michele.colombino@unito.it (M. Colombino); laurentiu.zaharia@edu.unito.it (L. J. M. Zaharia); giorgia.iacobellis@edu.unito.it (G. Iacobellis); rachele.mignone@unito.it (R. Mignone); ivan.spada@unito.it (I. Spada); chiara.bonfanti@edu.unito.it (C. Bonfanti); emilio.sulis@unito.it (E. Sulis); luigi.dicaro@unito.it (L. D. Caro); guido.boella@unito.it (G. Boella)

0009-0007-3248-1661 (M. Colombino); 0009-0002-3559-8367 (L. J. M. Zaharia); 0009-0003-1730-7711 (G. Iacobellis); 0009-0009-2699-8730 (R. Mignone); 0009-0002-0459-1189 (I. Spada); 0009-0007-8015-7786 (C. Bonfanti); 0000-0003-1746-3733 (E. Sulis); 0000-0002-7570-637X (L. D. Caro); 0000-0001-8804-3379 (G. Boella)

on an Italian case study, describing in detail the type of data available, the technologies used, and the models for automatic classification.

Our contribution focuses on two fundamental parts: the search for a criterion for transferring the labels of a non-hierarchical structure within the labels of an existing hierarchical structure, and the enrichment of the data contained in the labels of this hierarchical structure. In the following of the paper, Section 2 introduces the background with related works, the definitions and the data used to perform the classification and the alignment tasks. Section 3 describes the method, while early results are detailed in Section 4. Section 5 concludes the paper.

2. Background and Related Work

Related work. The task of taxonomy alignment is typically concerned with aligning multiple taxonomies that share similar or related concepts. Although our context differs from this scenario, as we have a flat set of labels on one side and a taxonomy on the other, we present relevant research on taxonomy alignment as it represents the most analogous context in the literature.

This task has gained significant attention in recent years due to its applications in knowledge integration, data integration, and semantic interoperability. Several approaches have been proposed to tackle this problem, including ontology matching, hierarchical clustering, and rule-based methods. Ontology matching is a popular approach that leverages semantic similarity measures to align taxonomies [4]. Hierarchical clustering methods group similar nodes from different taxonomies [5], while rule-based methods use expert knowledge to map concepts across taxonomies [6]. Recent studies have explored the use of machine learning techniques, such as deep learning, to improve the accuracy of taxonomy alignment [7].

As we previously mentioned, our proposed method is novel in that it addresses a slightly different scenario. Specifically, we consider two collections of documents that are labeled with distinct sets of labels, where only one of the sets is organized in a taxonomy. This scenario is particularly noteworthy for several reasons, such as providing more structure to documents with flat labels or augmenting a coherent text collection with additional documents that lack extensive labeling.

Definitions. This section introduces some terms related to the legal domain, as well as keywords that require careful definition and disambiguation to access the meaning of the technical parts of this article.

- **Judgement**: (i.e. Sentenza, in Italian) is the judicial decision given by a judge or court, in relation to a case, and can be identified by several different parameters. One identifier is obtained by combining the judgement code and the year of publication. The former indicates a sequential code given by the Court when it is published, the latter indicates the year of publication.
- **Subject**: (i.e. Materia) indicates the macro-category to which a given judgement belongs, as well as the section of the court that issued it. The judgements that are dealt with in this paper are related to the subject area of Labour Law. Other examples of subject areas are Civil Law (i.e. Civil Law), Tax (i.e. Tributaria), etc.
- **Label**: (i.e. Voce) indicates a categorisation label of a judgement. These labels respond to the individual court's way of conceiving and categorising judgements. Specifically, labels can be presented in a taxonomic form, i.e. organised into labels and sub-labels, or they can be unstructured. Examples of labels are "risarcimento danni", "invalidità civile", "retribuzione", etc.
- **NGR**: an acronym standing for 'Numero Ruolo Generale', it is an identifier corresponding to a numerical sequence specific to a particular court and assigned by that court to a specific case. It is used to link all the acts and documents relating to a specific case in a single folder.

3. Method

Following the recovery of judgements from two different sources (Turin Court, and a public online archive), we applied different preprocessing techniques to model the data, rendering them useful to the study. This section describes the results of a classification test used to compare the non-aligned labels results with the ones obtained in this study, to demonstrate the presence of improvements following the alignment between labels. Finally, we define the pipeline of actions executed to obtain the alignment of two taxonomies of labels coming from two different sources. Figures 1 and 2 show two of the main steps of the alignment pipeline with the corresponding benefit in term of enrichment the Leggi d'Italia taxonomy with the judgments of Turin corpus.

3.1. Data and sources

Sources. The judgements used in this study derive from two different sources: a set of 27,477 judgements issued by the Court of Turin, relating to the labour section and a set of 21,562 judgements extracted from Leggi d'Italia [8], an online archive which is a point of reference in legal matters in Italy. These led to a comparison

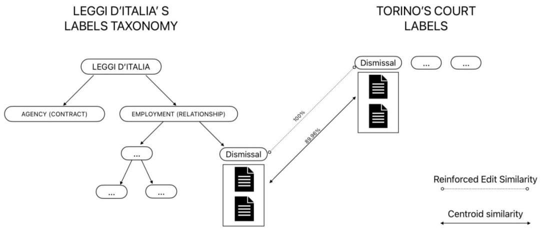


Figure 1: Two of the main steps of the transferring taxonomy pipeline. Firstly we compute a reinforced edit similarity between the labels of the two hierarchy. Secondly, we calculate the centroid cosine similarity between the candidate clusters of judgments.

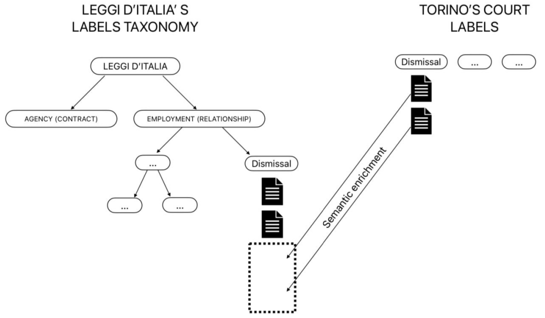


Figure 2: The contribution of our alignment criterion to the enrichment of the Leggi d'Italia taxonomy with the judgments of Turin corpus.

of documents from different courts on the same semantic level in order to identify similar patterns concerning the way judgements are drafted. The data have been standardized, as the judgements obtained have different digital formats, such as 'pdf', 'docx', 'doc', 'docm' and 'html'. Of all the Turin judgments, only a subset of 4,804 are labelled. This finding is very significant, as subsequent work on classification will be influenced by the reduced volume and will form the basis for a first attempt at transferring taxonomies, as will be discussed below.

Figures 3 and 4 show clearly that the judgements' labels from the two sources are structured differently. In fact, while Turin's labels are organized in a linear structure without a precise hierarchy, Leggi d'Italia's labels have a taxonomic relationship, structured in concepts and sub-concepts. The "/" character shows the end of a sub-tree and the start of a new sub-tree in the hierarchy. Secondly, looking at the distributions, it is immediately apparent that the two sets of judgments are highly unbalanced, with inevitable consequences in terms of automatic classification. A small portion of the Leggi d'Italia's labels tree is shown in figure 5. It can be observed how the labels are layered in sub-trees.

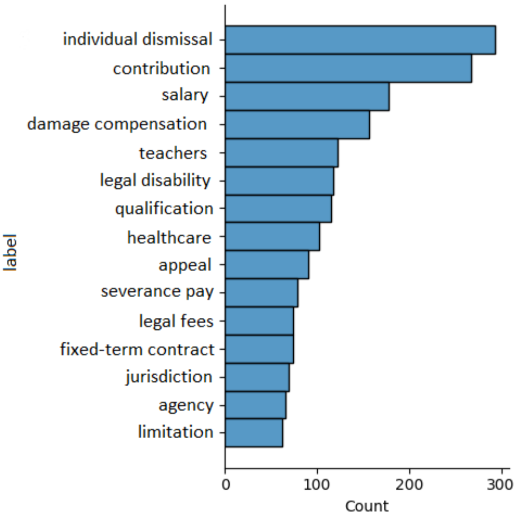


Figure 3: Distribution of the 15 most frequent labels of Turin court's dataset. The labels shown in the figure are a subset of the 309 labels of the Turin dataset used for classifying the judgements.

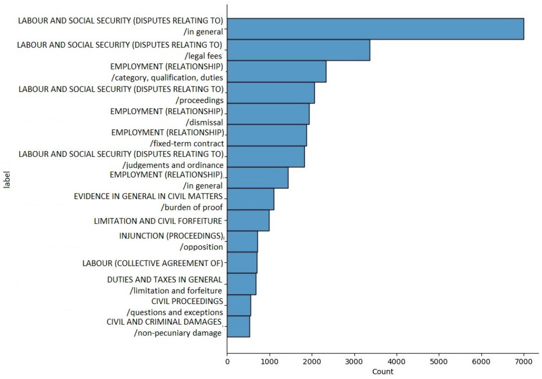


Figure 4: Distribution of the 15 most frequent labels of the 300 ones of the Leggi d'Italia dataset.

Data. An important phase preceding the work of alignment and automatic classification of the judgements concerns the retrieval of the data, as well as the segmentation and organization of the textual content of the processed documents. As anticipated in the previous paragraph, the set of judgements of Leggi d'Italia, henceforth called *"corpus-LI"*, was retrieved as a result of a scraping work conducted on the Leggi d'Italia web platform [8] using the python library *scrapse* [9]. The suite allows both retrievals of digital documents of judgements and extraction of the content in JSON format. A similar work, for uniformity, was conducted

```
LEGGI D'ITALIA
└── AGENCY (CONTRACT)

LEGGI D'ITALIA
└──  DUTIES AND TAXES IN GENERAL
    ├── Inspection
    ├── Commission’s competence
    ├── Service of documents
    ├── Constitutional legitimacy issues
    ├── Taxable subjects
    ├── . . .

LEGGI D'ITALIA
└── INJUNCTION (PROCEEDINGS)
    ├── Injunction
    ├── In general
    ├── Opposition
    └── Constitutional legitimacy issues

LEGGI D'ITALIA
└── EMPLOYMENT (RELATIONSHIP)
    ├── Health checks
    ├── Project contract
    ├── Dismissal
    ├── Sickness and invalidity
    ├── Mobbing
    ├── . . .

LEGGI D'ITALIA
└── SOCIAL WELFARE
    ├── Sales agents and representatives
    ├── Combative benefits
    ├── Dealers
    ├── Social security organizations and institutions
    ├── Pharmacists
    ├── . . .
```

Figure 5: A small portion of the the Leggi d'Italia's labels tree.

on the set of judgements of the court of Turin, which for simplicity we will call *"corpus Turin"*. The textual content was extracted and segmented tracing the same representation obtained on the corpus-LI. Finally, we obtain the following two JSON representations: JSON metadata and JSON corpus.

JSON metadata: This first representation in JSON format collects all the metadata found among the textual content of a judgment. Such information can be useful not only for data visualization purposes, such as knowing how many judgements were issued by a certain court rather than another but also for automatic classification and alignment purposes. Among the most significant pieces of information, the following metadata was collected:

- **tribunale**: (i.e court) the indication of the specific legal body that issued a certain judgment, e.g.: court of Cuneo.
- **sezione**: (i.e. Materia) the indication of the section to which the court that issued the judgment belongs.
- **voce**: (i.e label) Indication of the classification heading of the judgment.
- **sent code**: identification code of the judgment
- **sent year**: year of publication
- **nrg code**: general role code
- **nrg year**: year associated with the general role code. The nrg code and nrg year pair identifies a specific case within a court.

JSON corpus:. This second representation includes all the content information of a certain judgment. The most relevant ones are:

- **oggetto**: (i.e object) in the form of a short sentence, it represents the topic addressed in the case from which the judgment is issued. Typically it is very informative about whether a judgment belongs to a certain category, but it is not sufficient.
- **conclusioni**: (i.e conclusions) Some indication of the conclusions of the trial for the parties in the case.
- **fatti**: (i.e facts) represents the central body of the judgment in which the facts of the case are discussed.
- **decisione**: (i.e decision) the decision made by the judge. In some cases, fact and decision are merged together.
- **P.Q.M**: (i.e. Per Questi Motivi) the final verdict.

A third representation, for convenience, in **unified** format was derived by merging the previous two.

3.2. Preprocessing

Considering the importance of the data processed and respecting the privacy of the parties involved in the cases, the judgements from the court of Turin were subjected to a process of pseudo-anonymization. This operation allowed us to manipulate the judgements without involving the personal data of the litigants. Sensitive data, such as proper names, tax codes, were obscured by symbolic labels so as to preserve the semantics and relationships between the entities involved within the text. The judgments of the *corpus-LI*, on the other hand, since they are made public, are already presented anonymized. Importantly, complete anonymization was not conducted; in fact, not all sensitive entities were retrieved. Tools such as NER for the Italian language are not very reliable and are prone to error, which is why we refer to a pseudo-anonymization. The preprocessing used to manipulate the judgments in the classification and alignment tasks consists of a pipeline of operations, listed below:

- conversion in lower case
- removal of numerical quantities
- removing punctuation and special characters
- removal of people's first and last names
- removal of stopwords

- lemmatization

Sensitive data such as first and last names were removed at the preprocessing stage to ensure the least possible dirty data to be given as input to machine learning models. To identify these entities, we retrieved a dataset of proper names found on the Agenzia per l'Italia Digitale web portal [10]. All preprocessing was done using the python Spacy library [11], however, the list of stopwords for the Italian language was enlarged using external resources [12].

3.3. Classification

Datasets. Preceding the taxonomy transfer phase, our work focused on a preliminary classification task. This preliminary task was exploratory in order to better understand the most appropriate vector space modeling patterns and representations on the data at our disposal. Considering the imbalance of the data, classification tests were conducted on a limited number of judgments. Specifically, two corpora were created which in the following we will call *"corpus_11_labels_torino"*, constructed using some the 15 labels in figure 3 for a total of 318 judgments and a *"corpus_10_labels_LI"* for a total of 7,308 judgments from Leggi d'Italia, whose details will be discussed in the section 4. The reason why these 2 datasets have a different sizes is to be found both from the unbalancy of the distribution of the labels, we can see in figures 4 and 3, and from the results of the alignment process. Various vector space modeling techniques were used to create the datasets. Starting from these representations, several classification tests were conducted employing some machine learning models. From the extracted judgments in JSON format, textual contents related to the following fields (references in 2) were retrieved for the creation of the datasets: *"subject"*, *"fact"*, *"decision"*, *"conclusion"*. The information contained in the *"P.Q.M"* was discarded, as these are very recurring phrases, frequently used formulas in all subjects, as such, negatively affect the classification. Similar considerations will be taken up in section 5. Starting from these fields, 4 different datasets were created for the *"corpus_11_labels_torino"*. At the end of the preprocessing pipeline on the *"corpus_11_labels_torino"*, the use of TF [13] and TF-IDF [14] led us to define two sparse matrices of shape 10,955 x 318. To have a recent comparison regarding the state of the art on the embeddings representation, the remaining 4 datasets were created using the following resources:

- **Doc2Vec:** Doc2Vec [15] is an unsupervised neural network model that learns fixed-length feature vectors for representing textual data. The network architecture, like for word2vec [16], provides two different algorithms for the embeddings generation: "Continuous Bag of Words" (CBOW) and "Skip-Gram'(SG)" [16]. For the learning process, we considered the first one, CBOW, which implementation is visible in the python library: gensim.models.Doc2Vec [17]. The model, after a preprocessing step, specifically required for this implementation of the algorithm, was trained for 30 epochs with the following hyperparameters: vector_size = 300, negative=5, hs=0,min_count=2,sample=0, alpha=0.025, min_alpha=0.001.
- **Italian-Legal_bert**: Italian-Legal_bert [18] is a version of a pretrained BERT-BASED [19] model (ITALIAN XXL BERT [20]) trained on italian legal texts. The embeddings of this model are obtained running an additional round of training for 4 epochs on a 3,7GB preprocessed text from the National Jurisprudential Archive using the Huggingface PyTorch-Transformers library [21].

Models. Our classification work focused more on data representation than on the use of neural models and fine-tuning of networks. A first experiment has seen the use of a multiclass SVM [22] as a baseline model. Assuming nonlinearly separable data, we trained the SVM model using an "rbf" kernel-trick [23]. In the second order, considering the dimensions of the datasets, we conducted some tests using a Logistic Regression [24] model with a "lbfgs" solver. In presence of sparse and poor data, these models tend to show the same behaviour. Furthermore, we considered a Random Forest classifier [25] with max 2,000 trees, which, instead, results more efficiently on datasets with a limited number of features. Finally, the same tests were repeated running an Ensemble Learning task with a simple Voting classifier [26] using all the previous models.

3.4. Pipeline

Considering the structure of the data and the small volume of judgments, we initially attempted taxonomic alignment between Leggi d'Italia's labels and those of the Turin court. To proceed critically, we defined a pipeline that considers the transfer process in steps, articulated in: *label comparison*, *semantic similarity* and *validation with classification.*

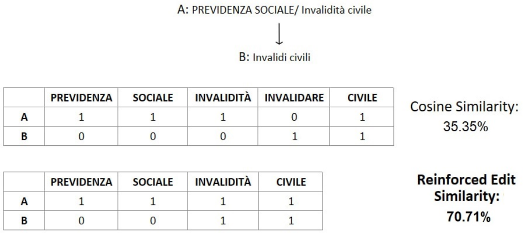
3.4.1. Labels comparison

In this first stage we considered the labels' alignment exclusively from a lexical point of view. From a first superficial reading, it is easy to find some similarities, by looking at the 2 lists of labels in figure 3 and 4. Taking the following *labels* as examples: *"Labour and social security (Disputes relating to) Legal fees"* and *"Legal fees"*, respectively from Leggi d'Italia and from Turin. Without looking at the content of the judgments, it would seem that the two voices speak about very close topics, however, only through a deeper analysis can this observation be confirmed or refuted. The label comparison phase consisted of searching for criteria of approximation between the labels of the two sets of judgments, hence leading us to define the Reinforced Edit Similarity.

Reinforced Edit Similarity. The alignment criterion used is a combination of *edit distance* [27] and *cosine similarity* [28]. Since, Leggi d'Italia presents a taxonomy articulated in a tree structure, we decided to distribute it in N levels of labels, with N=3 the maximum depth. Starting from the leaves, and going up to the root, we calculated the score for each entry pair of Turin and Leggi d'Italia. It should be considered that the labels were preprocessed, not only to facilitate better approximation but also because they had punctuation symbols, special characters, and many. In a first step, approximation was performed by tokenizing the labels, then applying cosine similarity on the vectorized representation created with counter vectorizer [29]. Later, we abandoned this criterion as it did not take into account the differences in lemmatization of the words with respect to their POS tag. Figure 6 shows how a similarity score of 35.35% was derived from the two labels "social security/civil disability" and "civil invalids" (i.e., social security/Legal disability, legal invalids), against reinforced edit similarity score of 70.71%.To facilitate a better approximation, once we switched to the vector representation using the *CountVectorizer* module, we calculated the edit distance between each pair of words, with a threshold ≤ 2. Pairs that do not have a distance greater than the threshold were transformed in such a way as to unify them (make them identical). In this way, a subsequent application of cosine similarity will present a higher score, rewarding in fact, those labels that are lexicographically close. At the end of the alignment process, for each pair of Turin and Leggi d'Italia labels, we considered the one with the highest cosine similarity score on the various levels. Table 1 shows some results of the lexical similarity scores of the Turin labels, evaluated on the various levels of labels in the Leggi d'Italia taxonomy.

3.4.2. Semantic comparison

Embeddings representation. In this step of the pipeline we focused on the semantic aspect. Our first goal has been to choose how to converge to a single vector representation. Looking at the table 3, in section results 4, we chose to transform the judgements into embeddings Doc2vec. This choice has been motivated by the facts that on the available data we had, Doc2vec represented the model that had the better results.



A: PREVIDENZA SOCIALE/ Invalidità civile

↓

B: Invalidi civili

	PREVIDENZA	SOCIALE	INVALIDITÀ	INVALIDARE	CIVILE
A	1	1	1	0	1
B	0	0	0	1	1

Cosine Similarity:
35.35%

	PREVIDENZA	SOCIALE	INVALIDITÀ	CIVILE
A	1	1	1	1
B	0	0	1	1

Reinforced Edit Similarity:
70.71%

Figure 6: Comparison of the similarity scores between the counter vectorizer's cosine similarity and the reinforced edit similarity. We perform a matching between two labels: "social security/civil disability" and "civil invalids". Before the vectorization process we lemmatize the words, so the word "invalidi" (i.e invalids) is trasformed into "invalidare" (i.e invalidate) and "civili" (i.e civils) into "civile" (i.e civil). If we apply the reiforced edit similarity before the lemmatization phase, we obtain a cosine similarity score of 70,71%, because the words "invalidi" and "invalidità" (i.e disability) are trasformed into the same word, i.e "invalidità".

Label Torino	Label Leggi d'Italia	similarity level 1	similarity level 2	similarity level 3
Individual dismissal	EMPLOYMENT (RELATIONSHIP)/ Dismissal	0.000	0.707	0.000
damage compensation	CIVIL AND CRIMINAL DAMAGES/ Non-pecuniary damage	0.353	0.499	0.000
jurisdiction	ADMINISTRATIVE JUSTICE/ Jurisdiction/division of jurisdiction between ordinary and administrative courts	0.000	1.00	0.353
agency	AGENCY (CONTRACT)	0.707	0.000	0.000
limitation	RETIREMENT/ Limitation	0.000	1.000	0.000

Table 1
Reinforced edit similarity scores on a subset of the turin's labels. The last three column show the scores on the three levels of labels of the Leggi d'Italia's hierarchy.

Semantic similarity. In this phase we recovered an even number of judgements from both the labels sources. Some of the labels we used at this point of the pipeline, resulting from the labels comparison phase, are visible in table 1. We then proceeded to transform them in a Doc2vec form, obtaining then two clusters composed by judgements of the same cardinality. To value the closeness between these clusters, we applied metrics such as cosine similarity between the centroids vectors. The semantic similarity score in combination with the Reinforced Edit Similarity score contributed to an overall score that allowed us to evaluate the alignment of the labels. Specifically, all matching of labels that returned a semantic similarity score $\geq 70\%$ were retrieved. Overall, considering that some matches concerns labels that in both sources have limitations due to the fewer number of judgements, all of those that have a cardinality of less than 10 judgements have been discarded, as we deemed them of less importance. Table 2 shows the results of semantic similarity on a subset of the Turin labels, chosen from those most populated on both sources.

Label Torino	Label Leggi d'Italia	Centroid similarity
agency	agency (contract of)	92.820
subordinate work	Subordinate work (Relationship of)/ Category, qualification, duties	87.110
social allowance	Social welfare/ Legal disabled person	75.489
dismissal	Subordinate work (Relationship of)/ dismissal	89.960
notification	Duties and taxes in general/ Service of documents	71.880
Injunction	Injunction proceedings/ Injunction	70.719

Table 2
Centroid similarity on a subset of the labels with the highest reinforced edit similarity score.

3.4.3. Validation with classification

The last step of this pipeline implements an *a posteriori* validation of the quality of our alignment by performing a classification of the judgments on the labels in the table 1. In this phase we demonstrate, as shown in section 4, how the alignment did not have significantly negative impacts on the classification of judgments. From an initial set of 309 labels on the Turin corpus, only a subset of 11 labels returned a centroid similarity score ≥ 70%. Here, the list of the candidates labels: "agency", "social allowance", "subordinate work", "dismissal", "individual dismissal", "injunction", "notification", "proof", "severance pay", "sickness allowance" and "assistance". At last we can confidently say that the results we obtained validate the alignment process. In particular, "individual dismissal" and "dismissal" are associated to the same label of the Leggi d'Italia hierarchy: "Subordinate work (Relationship of)/dismissal", so during the classification process, these labels are considered as the same label. At the end of this final step of the pipeline we train some machine learning models using the "corpus_10_labels_LI" as training set and the "corpus_11_labels_torino" as testing set, for all these 11 labels.

4. Results

In this section, we will show in more detail all the results of our experiments, after and before the alignment pipeline. All data visualized in the following tables are derived by applying a 10-fold cross-validation method on the datasets and models defined in the section 3.3.

4.1. Pre alignment classification

Table 3 shows the results of the main evaluation metrics we considered: accuracy, precision, recall and f1 score on the "corpus_11_labels_torino". All the results obtained from the different models, except for the dataset created by Doc2Vec embeddings, reflect our expectations about the decreasing of the performances. Italian-legal-BERT reported the worst results, due to the excessive sparseness of the data, while doc2vec appears to guarantee excellent performance even with the baseline models. For that reason, all the post alignment classification tests are conducted using only the doc2vec representation.

Classification test on "corpus_11_labels_torino"				
Dataset	Random forest	SVM	Logistic Regression	Enseble Voting
		Average accuracy		
TF	0.784	0.776	0.802	0.816
TF-IDF	0.784	0.805	0.794	0.808
Ita-legal BERT	0.722	0.714	0.786	0.741
Doc2Vec	0.914	0.954	0.962	0.957
		Average precision		
TF	0.859	0.829	0.791	0.765
TF-IDF	0.865	0.859	0.837	0.853
Ita-legal BERT	0.773	0.835	0.766	0.853
Doc2Vec	0.943	0.966	0.972	0.965
		Average recall		
TF	0.730	0.723	0.785	0.788
TF-IDF	0.726	0.744	0.737	0.750
Ita-legal BERT	0.640	0.595	0.748	0.750
Doc2Vec	0.878	0.945	0.955	0.955
		Average f1 score		
TF	0.752	0.756	0.782	0.788
TF-IDF	0.745	0.773	0.751	0.768
Ita-legal BERT	0.660	0.602	0.752	0.768
Doc2Vec	0.898	0.954	0.962	0.955

Table 3
Evaluation of the performances of the four datasets derived by the *"corpus_11_labels_torino"*

4.2. Post alignment classification

Table 7 shows the accuracy scores evaluated on the "corpus_11_labels_torino" testing set, using all four models introduced in the section 3. As we noted, the performances of the models decreases significantly, as the number of items increases. Looking at the SVM curve, for the first 6 labels, the accuracy has a score of 95%, which decreases to a value of 80%, for a total of 11 labels. If we compare these results with the previous ones on the pre-alignment classification tests, we note that, for the first 8 labels, the performance of the SVM model does not suffer a significant decrease, as instead it happens for the Logistic Regression and the Random Forest classifiers.

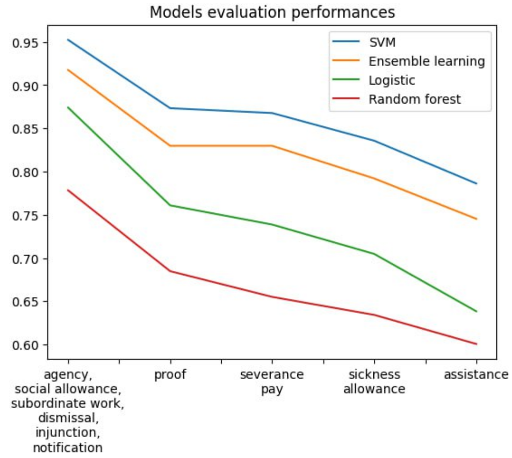


Figure 7: This graph shows the evaluation of the performances of the four models we considered in the classification experiments. As the number of the labels increases, the accuracy of the models decreases significantly for all the models except for SVM. On the X axis there are all the labels with the highest centroid similarity score. On the Y axis we show the accuracy percentage.

```
LEGGI D'ITALIA
└── AGENCY (CONTRACT) <==> AGENCY

LEGGI D'ITALIA
└──  DUTIES AND TAXES IN GENERAL
    ├── Inspection
    ├── Commission's competence
    ├── Service of documents <==> NOTIFICATION
    ├── Constitutional legitimacy issues
    ├── Taxable subjects
    ├── . . .

LEGGI D'ITALIA
└── INJUNCTION (PROCEEDINGS)
    ├── Injunction <==> INJUNCTION
    ├── In general
    ├── Opposition
    └── Constitutional legitimacy issues

LEGGI D'ITALIA
└── EMPLOYMENT (RELATIONSHIP) <==> EMPLOYMENT
    ├── Health checks
    ├── Project contract
    ├── Dismissal <==> DISMISSAL
    ├── Sickness and invalidity
    ├── Mobbing
    ├── . . .

LEGGI D'ITALIA
└── SOCIAL WELFARE <==> SOCIAL ALLOWANCE
    ├── Sales agents and representatives
    ├── Combative benefits
    ├── Dealers
    ├── Social security organizations and institutions
    ├── Pharmacists
    ├── . . .
```

Figure 8: A portion of the Leggi d'Italia's hierarchy with the Turin's labels (blue) aligned.

5. Conclusions and Future work

In this paper we explored a first approach of a transposition and alignment of a not-hierarchic structure in a well defined taxonomy, using a pipeline of different approaches. With the judgements that we obtained from two different sources and their labels, we realized a first step in which we defined lexical similarity on the labels, while testing a new metric of *lexical proximity* resulting from the combination of existing techniques. Hence, going down to the semantic level, we applied *cosine similarity* by calculating the similarity of the centroids in the groups of judgments we identified as similar in the first step. After these two steps, as a check on the validity of our new found method, we trained some machine learning models, then evaluated the performance on the data before and after the alignment. As the final check on performance did not change negatively for some models, we were assured that the alignment did not lead to a loss of information in the newly constructed groups of judgments. Indeed, the processing of the data and the various phases of the pipeline we therefore described can be in the future further analyzed with new metrics and calculus approaches or with a more targeted study on how to preprocess the judgements' content to obtain more useful information. The latter can be added to the computation of a semantic similarity between judgements.

5.1. Improve preprocessing

Given a preliminary analysis of the dataset, integrate to the processing pipeline a possible correction and/or elimination of words that contain spelling errors. Given an initial analysis of word occurrences, it was found that those with minimal frequency contained spelling errors. By extracting ten words at random, at least half have spelling errors.

5.2. Keywords extraction

Further improve the preprocessing with an expectation of increase the accuracy of the transferring taxonomy pipeline described in section 3.4, by extracting the most significant keywords of each labels. The goal is to remove the most frequent words that have an even distribution across all labels, thus having a low significant impact, and identify those that best identify each label. As can be see in figure 9, the word "operator" has a higher frequency under the label “agency” than under other labels, nominating itself as a potential keyword.

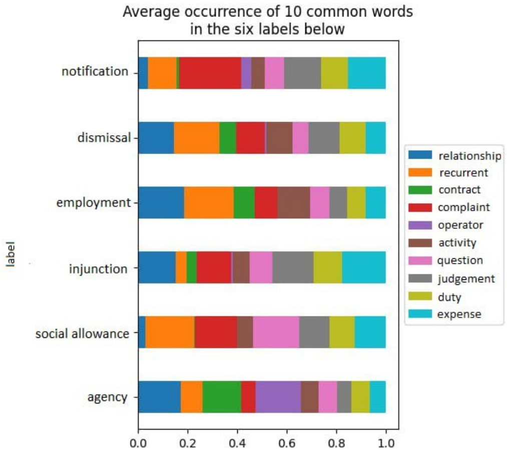


Figure 9: Average of ten common words in the six labels in the graph. From this graph, which offers an initial analysis, it can be seen that the word “duty” is evenly distributed across all labels (to be eliminated), compared to the word ”operator” (potential keyword).

5.3. Other similarity approaches

Use of different labels similarity approaches like Polyfuzz [30]. This package is used to compare similarity between strings using different type of models to create n-grams on a character level. After generating the n-grams and applying the models on the strings' words, it use cosine similarity to compare the generated vectors.

References

[1] D. M. Katz, R. Dolin, M. J. Bommarito, Legal informatics, Cambridge University Press, 2021.

[2] E. Sulis, L. Humphreys, F. Vernero, I. A. Amantea, D. Audrito, L. D. Caro, Exploiting co-occurrence networks for classification of implicit inter-relationships in legal texts, Inf. Syst. 106 (2022) 101821. doi:10.1016/j.is.2021.101821.

[3] D. Audrito, E. Sulis, L. Humphreys, L. Di Caro, Analogical lightweight ontology of eu criminal procedural rights in judicial cooperation, Artificial Intelligence and Law (2022) 1–24.

[4] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, Weta: Automatic taxonomy alignment via word embeddings, Computers in Industry 138 (2022) 103626. doi:https://doi.org/10.1016/j.compind.2022.103626.

[5] P. Clerkin, P. Cunningham, C. Hayes, Ontology discovery for the semantic web using hierarchical clustering, Technical Report, Trinity College Dublin, Department of Computer Science, 2002.

[6] S. Fernández, J. R. Velasco, M. A. López-Carmona, A fuzzy rule-based system for ontology mapping, in: Principles of Practice in Multi-Agent Systems: 12th International Conference, PRIMA 2009, Nagoya, Japan, December 14-16, 2009. Proceedings 12, Springer, 2009, pp. 500–507.

[7] J. Chen, E. Jiménez-Ruiz, I. Horrocks, D. Antonyrajah, A. Hadian, J. Lee, Augmenting ontology alignment by semantic embedding and distant supervision, in: The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18, Springer, 2021, pp. 392–408.

[8] Leggi d'Italia p.a., accessed 27.04.2023. URL: https://pa.leggiditalia.it/#mode=home,__m=site.

[9] scrapse, accessed 27.04.2023. URL: https://pypi.org/project/scrapse/.

[10] Agenzia per l'Italia digitale, accessed 27.04.2023. URL: https://www.dati.gov.it/view-dataset.

[11] Industrial-strength natural language processing, accessed 27.04.2023. URL: https://spacy.io/.

[12] Stopwords italian, accessed 27.04.2023. URL: https://github.com/stopwords-iso/stopwords-it.

[13] H. P. Luhn, The automatic creation of literature abstracts, IBM J. Res. Dev. 2 (1958) 159–165.

[14] K. S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Documentation 60 (2021) 493–502.

[15] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, in: International Conference on Machine Learning, 2014.

[16] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, 2013.

[17] Gensim topic modelling for humans, accessed 27.04.2023. URL: https://radimrehurek.com/gensim/models/doc2vec.html.

[18] D. Licari, G. Comandè, ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law, in: Symeonidou et al. (Ed.), EKAW, volume 3256 of *CEUR Workshop Proceedings*, CEUR, Bozen-Bolzano, Italy, 2022. URL: https://ceur-ws.org/Vol-3256/#km4law3.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: ACL: HLT, Vol. 1, ACL, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[20] Italian bert, accessed 27.04.2023. URL: https://huggingface.co/dbmdz/bert-base-italian-xxl-cased.

[21] hugging face transformers, accessed 27.04.2023.

URL: https://huggingface.co/docs/transformers/index.

[22] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144–152.

[23] Support vector classification, accessed 27.04.2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

[24] Logistic regression classifier, accessed 27.04.2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[25] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[26] Soft voting/majority rule classifier for unfitted estimators, accessed 27.04.2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html.

[27] Levenshtein edit-distance, accessed 27.04.2023. URL: https://www.nltk.org/api/nltk.metrics.distance.html.

[28] Compute cosine similarity between samples in x and y., accessed 27.04.2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html.

[29] Convert a collection of text documents to a matrix of token counts., accessed 27.04.2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

[30] Polyfuzz, accessed 27.04.2023. URL: https://maartengr.github.io/PolyFuzz/.