Multi-label Classification of Legislative Text into EuroVoc

Guido BOELLA a, 1, Luigi DI CARO a, Leonardo LESMO a, Daniele RISPOLI a, Livio ROBALDO a

a Dipartimento di Informatica, Università di Torino

Abstract. In this paper we present a novel method for the automatic classification of multi-label text documents. In principle, automatic classification of text is usually tackled by supervised Machine Learning techniques like Support Vector Machines (SVM), that typically achieve state-of-the-art accuracy in several domains. Nevertheless, SVM cannot handle multi-labeled documents, thus a specific pre-processing of the data is needed. In this paper we present a novel technique for the transformation of multi-label data into mono-label that is able to maintain all the information, allowing the use of standard approaches like SVM. We then evaluate our system using JRC-Acquis-it, a large dataset of Italian legislation that has been manually annotated according to EuroVoc, demonstrating the potential of our approach compared to the current state of the art.

Keywords. Multi-label Text Classification, EuroVoc.

Introduction

Classification of legal text is an important task given the large amount of documents to deal with to retrieve information and the risks due to not finding all relevant documents. It usually involves knowledge intensive manual work which is slow and costly. Knowledge engineers specify the domain to which each norm belongs, selecting key terms within a domain-specific ontology. In particular, many parliaments and other institutions are increasingly using the EuroVoc descriptors to classify documents. EuroVoc is a large multidisciplinary multilingual hierarchical thesaurus of more than 6,700 classes covering the activities of EU institutions. Given the amount of legal documents produced everyday and the huge mass of pre-existing documents to be classified high quality automated or semi-automated classification methods are most welcome in this domain.

In general, automated text classification is a well-studied task that most of the times works on flat organizations of the categories. Support Vector Machines is known to achieve state-of-the-art levels of accuracy in several domains in that sense. Despite this, sometimes categories are structured in hierarchical organizations. While this situation can be generally faced by flattening out the hierarchy, an accurate exploitation of such a knowledge can lead to significant improvements of the used approach. Moreover, some labels may have very few positive examples, leading to highly skewed data that can be

1 Corresponding Author.
problematic for classifiers. Another increase of complexity in text classification tasks is represented by multi-labeled datasets where each piece of text can belong to more than one category. Finally, the large amount of data poses further complications.

Thus, the research question of the paper is: how to have high quality automated classification for large amounts of documents into large number of classes?

Our methodology is to use supervised Machine Learning techniques like Support Vector Machines (SVM), that typically achieve state-of-the-art accuracy in several domains. Nevertheless, SVM can not handle multi-labeled documents, thus in this paper we introduced a specific pre-processing of the data that preserves all the information while being suitable for SVM.

Due to space reason we do not take into account in this paper the information coming from the hierarchical organization of the categories and the issue of multi-language.

We then evaluate our system using a large dataset of Italian legislation named EuroVoc, demonstrating that our approach outclasses all the existing approaches.

In Section 1 we present the related work about classification of laws, management of category-based and multi-label cases. In the next sections we then illustrate our methodology and the obtained results. Finally, Section 4 concludes the paper.

1. Background and Related Work

In this section we present the main works that have been done in automatic classification of legislative text and classification methods in case of multi-label texts. For a more detailed survey on the techniques we refer to, [1] represents one of the most complete and up-to-date reference.

1.1. Classification of legislative text

In [12], the authors proposed JEX, a system for the classification of texts into Eurovoc categories. The core of their approach is based on the construction of class profiles, that is sets of terms that define the related categories. The classification step is done by computing the cosine similarity of the documents with the class profiles.

1.2. Multi-label Classification

The problem of classifying text documents associated to more than a single category is currently met in several domains and applications. In the case of legislative documents, it is common to think about some legal text that can be associated to different concepts belonging to bureaucracy (e.g., “European contract”), economy (e.g., “inflation rate”), geography (e.g., “Italy”), and so forth.

From a technical point of view, dealing with this kind of information leads to more complex classification systems. For instance, Support Vector Machine is one of the most used Machine Learning algorithm to learn from text corpora, but it needs the documents to be mono-label. There are mainly two approaches to face such complexity: by transforming multi-label data into mono-label, or by adapting existing algorithms to work with mono-label data. In the next sections we will briefly go over them.

In mono-label scenarios, the classification task is only concerned with learning from a set of examples that are associated with a single label. The easiest case is when the
possible labels are just two, so that the learning problem is called a binary classification problem. In presence of more than two labels then it is called a multi-class classification problem. Our approach is meant to solve both multi-class and multi-label classification cases.

1.2.1. Problem transformation Methods

The first approach to tackle multi-label classification problems regards the transformation to standard mono-label data, suitable for any supervised learning algorithms.

The easiest way consists in removing all the documents that have more than one label from the dataset [2]. This is obviously ineffective, especially when all (or almost all) the documents are multi-label. Another method randomly selects one of the multiple labels for each documents, discarding the rest [2]. Even if this allows the system to learn from multi-label data, it actually does not represent a real treatment of the problem. Moreover, the accuracy of such solution is likely to be very low. A third approach called power set considers each different set of labels that is present in the multi-label data as a single label [5]. For instance, if the labels of a document are ‘A’, ‘B’, and ‘C’, the system transforms the labels of the document in a single label ‘ABC’. One of the negative aspects of this method is that it may lead to data sets with a large number of classes and few examples per class. Then, another transformation method learns one binary classifier for each label in the data. To classify a new document, it needs to pass over all the classifiers to determine its associated set of labels. Of course, in case of thousands of categories (as in the data that we will use), this strategy becomes unsustainable.

The idea that initially inspired our approach has been firstly described in [14]. However, to the best of our knowledge, it has not been considered for further refinements as well as it was not included in any of the evaluations. The intuition is to consider each n-labelled document as a collection of n minor documents each one associated to only one label. The main issue to deal with is about how to segment the original document, that is how to choose the features to maintain for each of the new mono-label documents. In [14] there is no mention to this problem, so they only considered the creation of n identical mono-label documents from one multi-label document. This is likely to carry to extremely overlapping classes from which it becomes difficult to learn any model. In Section 2 we present how we managed to do such transformation efficaciously.

1.2.2. Algorithm Adaptation Methods

The second approach to cope with multi-label data is to adapt standard learning algorithms that usually work with mono-label data. Our technique, however, is instead completely different in purpose, since data transformation carry to datasets that can be used for training any supervised learning algorithm.

In [3], the authors adapted the C4.5 algorithm for multi-label data by modifying the formula of the entropy calculation. [11] proposed two extensions of AdaBoost [7] for multi-label classification. The authors of [15] presented a technique called ML-kNN which adapts the simple k-NN learning algorithm. Similarly to our approach, [8] presented two improvements for the SVM classifier by using a different problem transformation method that creates one binary classifier for each category and it thus has strong limitations with respect to our system, as already mentioned in previous sections. Finally, [13] presented an algorithm that treats the multi-label classification task as an association rules mining approach.
2. Method

In this section we present our approach for the automatic classification of multi-label legal documents into the EuroVoc scheme. In particular, our contribution is twofold in the sense that we both try to solve the problem of transforming the multi-label dataset into a mono-label version that keeps all the information, and then we propose a slight variation that largely improves the current state-of-the-art systems, demonstrating the potential of our proposal.

2.1. Pre-processing of the Data

The process of transforming text into vectors requires selection of suitable terms, and use of a weighting function as part of the frequency calculations. Generally speaking, the accuracy of the classification methods is highly dependent on the quality of these procedures.

About the selection and transformation of terms, instead of simply using a list stop-words (e.g., “the”, “a”, etc.) to remove uninformative terms, we morphologically transform the text using lexical roots (i.e., the lemmas) using a dependency parser called TULE that performs a deep analysis over the syntactic structure of the sentences. The aim of this step is to eliminate noise while reducing redundant linguistic variability. Typically only nouns are then considered as informative features. The use of a syntactic parser is the basis for a fine-grained pre-processing: 1) it permits to disambiguate terms at syntactic level; 2) it allows for direct selection of the informative units (i.e., lemmatized nouns, verbs and so on); 3) it can be used for the generation of semantic chunks, i.e., syntactic subtrees that express high-level concepts. In future work we aim at studying this type of patterns to outperform our classification tool. Our approach may increase the system complexity as a whole, but it is a better solution than WordNet-like methods which only have top-domain ontologies and thus are unable to recognise and lemmatize many legal domain-specific terms.

2.2. From Multi-Label to Mono-label

As seen in the related work section, the transformation of a multi-label text corpus into a mono-label can be done through several strategies. In this work we used an idea that comes from an approach mentioned in [14], but never studied and evaluated anyhow. In this paper, we resumed such idea extending it with an algorithm based on Information theory.

The transformation of a multi-label dataset into a monolabelled one enables the use of a standard Support Vector Machine classifier, that is known to be best choice when dealing with textual databases. The general idea is that one \(n\)-labelled document can be seen as a collection of \(n\) different documents. Since a document is represented by a numerical vector (according to the Vector Space Model [10]), it can also be viewed as a merging of single-labelled vectors. This, however, is based on the assumption that one feature only belongs to one label, which is clearly an approximation of the possible state of affairs.

Given the vectorial representation of a text \(\vec{d}\) and its set of associated labels \(S_d\), the system splits the document into \(|S_d|\) virtual documents, each one belonging to one
label. The problem of splitting and assigning vectors and categories is solved making use of the Pointwise Mutual Information (PMI) measure calculated over the co-occurrence matrix computed on the features (or terms) and the labels. In detail, the system builds a matrix $M$ of $r = |C|$ rows and $c = |F|$ columns, where $C$ and $F$ are respectively the set of categories and the set of features, and where each value $M_{i,j}$ contains the co-occurrence of the feature $f_i$ with the category $c_j$ (with $0 < i < |F|$, and $0 < j < |C|$).

Once the matrix $M$ is calculated, it is then transformed in a PMI-based matrix where each value $M'_{i,j}$ is replaced with:

$$M'_{i,j} = \frac{P_{i,j}}{P_i * P_j}$$

where $P_{i,j}$ is the probability of having a non-zero co-occurrence value for the feature $f_i$ and the category $c_j$ (that is, $M_{i,j} > 0$) in the entire corpus, while $P_i$ and $P_j$ are the individual probability to find in the corpus the feature $f_i$ and the category $c_j$ respectively. The utility of $M'$ is to capture the strength of the associations between features and categories.

At this point, for each original document vector $\vec{d}$ to be segmented, given the set of categories $S_d$ to which it belongs, the system creates $n = |S_d|$ new document vectors $\vec{d}_k$ (each one associated to exactly one class) in the following way:

$$\vec{d}_k = <\text{sel}(f_1), \text{sel}(f_2), ..., \text{sel}(f_{|F|})>$$

where $k \in S_d$ (it represents the category associated to the new vector), and where $\text{sel}(f_i)$ is a selection function that can assume the following values:

$$\text{sel}(f_i) = \begin{cases} f_i, & \text{if } \exists S'_d \subseteq S_d \text{ and } |S'_d| = Q \text{ and } \forall x \in (S_d \setminus S'_d) \Rightarrow (M_{i,x} \geq \min M'_{i,k} \forall k \in S_d) \\ 0, & \text{otherwise} \end{cases}$$

that is, $\text{sel}(f_i)$ assumes the actual value $f_i$ coming from the original vector for the new vector $\vec{d}_k$ belonging to category $k$, if there is no PMI-value $M'_{i,k}$ that is greater than $M'_{i,k}$. In other words, $\text{sel}(f_i)$ will be equal to $f_i$ only if the strength of the association between such feature $f_i$ and category $k$ is higher than any other association between the same feature $f_i$ and the other categories in $S_d$.

With this technique, all the multi-label original vectors are separated in mono-label vectors that can be used in a standard SVM-based classification environment.

### 2.3. Selection Parameter

Up to now, each feature of an original multi-label document vector becomes a feature in only one of the created segmented vectors. This means that one feature will only serve as semantic unit for just one category at time. This can be a limitation of the method, since one feature can be of importance for more than one category. For this reason, we introduced the selection parameter $Q$, that modifies the original selection function in the following way:

$$\text{sel}_Q(f_i) = \begin{cases} f_i, & \text{if } \exists S'_d \subseteq S_d \text{ and } |S'_d| = Q\text{ and } \forall x \in (S_d \setminus S'_d) \Rightarrow (M_{i,x} \geq \min M'_{i,k}) \\ 0, & \text{otherwise} \end{cases}$$
that is, \( \text{sel}_Q(f_i) \) is equal to \( f_i \) if there exists a subset of \( S_d \) named \( S_d' \) of cardinality \( Q \) such that each one of its element has a PMI-value with feature \( f_i \) greater than (or equal to) all the elements outside \( S_d' \) (but in \( S_d \)). This way, the system allows the use of feature \( f_i \) for exactly \( Q \) segmented vectors.

### 2.4. Support Vector Machines

In our system we used the well-known Machine Learning supervised approach called Support Vector Machines (SVM), since it usually achieves state-of-the-art accuracy levels for textual data [4]. Our method first builds a mono-label version of the original multi-label dataset, so it can be used with SVM without any adaptation mechanism.

SVM makes use of the vectorial representation of the texts [10] and works by calculating the hyperplane having the maximum distance with respect to the nearest data examples. More in detail, we used Liblinear [6], a library for linear classification that is suited for fast text classification tasks on large datasets. In fact, SVM-based classifiers usually have limitations on the size of the input data, while Liblinear can work on data with millions of instances and features.

Our SVM-based system can output a probability distribution of the classes that can belong to a document, so it is possible to multi-label the documents. Now, the (common) problem is to understand how many categories to associate to a document. Usually, the system is evaluated by measuring the accuracy with different fixed cutoff over the probability distribution.

### 3. Evaluation

In this section we present the strategy for evaluating our proposed approach. First, we give an overview of the data that we used, then we define the metrics, and in the end we present the results that we achieved.

![Figure 1](image_url) **Figure 1.** The distribution of the documents with respect to the number of associated categories within the corpus. The average number of categories for each document is 5.6.
3.1. Data

For our experiments, we used JRC-Acquis-it\(^2\), a freely-available corpus of around 23K Italian legislative text documents written between the 1950s and now. Most of these documents were manually labelled according to Eurovoc, a multilingual, multidisciplinary thesaurus with about 7K categories (also called classes, labels, or descriptors from now on) covering the activities of the EU, the European Parliament in particular. It contains terms in several languages and it is managed by the Publications Office of the European Union, an interinstitutional office whose task is to publish the publications of the institutions of the European Union. Eurovoc is an ontology-based information collector that groups and links concepts through different types of relationships. The top level of the scheme is defined by 21 general concepts. Then, a second level defines 127 specifications (called micro-thesauri). Then, a set of other relationships like *related-term* relates all the nodes.

The dataset JRC-Acquis has been already used in [12], and it is known to contain very skewed data, from where it is usually difficult to learn models. Figures 1 and 2 show the distributions of the documents and the categories in the data.

3.2. Evaluation Metrics

In flat classification scenarios, it is common practice to evaluate classification systems by means of *Precision* and *Recall* (and *F-Measure*). While Precision is the fraction of retrieved instances that are relevant, Recall is the fraction of relevant instances that are retrieved. F-Measure is the harmonic mean of Precision and Recall.

In multi-label classification contexts, accuracy is often calculated by averaging Precision, Recall, and F-Measure values. There are two conventional methods of calculating these average values: Micro-average gives equal importance to each document and thus it uses a global contingency table to compute the accuracy values. Macro-Average instead calculates Precision and Recall for each category and then takes the average of these.

\(^2\)http://langtech.jrc.it/JRC-Acquis.html
our experiment we evaluate the system by using the Micro-average system, thus giving to each document the same importance.

Figure 3. Precision, Recall, and F-Measure calculated over the entire corpus, with the basic transformation strategy (i.e., with the selection parameter \( Q = 1 \). As can be noticed, the system achieves very high precision only for the first ranked categories, reaching the Break Even Point (BEP) near the fifth descriptors with a F-measure of about 44%. According to the F-Measure value, the system achieves the best accuracy (with a value of 44.61%) using the top 3 ranked categories.

Figure 4. Precision, Recall, and F-Measure calculated over the entire corpus, with the selection parameter \( Q = 2 \). In this case, the system achieves higher Precision and Recall values, reaching an F-Measure value of 61.47% for the first 4 categories, and a value of 58.32% for the top six categories.

3.3. Results

Our preliminary results of the proposed approach are shown in Figures 3 (basic strategy) and 4. As can be seen, the system achieves a high precision on the first ranked categories to the detriment of Recall, which is inevitably low due to the average number of descriptors per document (around 5.6). However, using 6 as fixed number of categories to estimate (because it is the closest value to the average number of categories per document), Precision gets much lower than reasonable values if we think at normally-accepted ac-
Figure 5. Accuracy levels when filtering out categories with few associated documents in the corpus. As can be seen, the system reaches higher accuracy values in presence of well-represented categories, after an initial drop. More in detail, while Precision remains quite stable with a filter of more than 20 necessary documents, Recall grows linearly up to 80%.

accuracy values in classification scenarios. This is in line with current state-of-the-art approaches working on EuroVoc. While these results call for more sophisticated mechanism that can make use of the hierarchical organization of Eurovoc, they demonstrate the potential of our approach compared with the work proposed in [12].

In addition, we tested our approach filtering out those categories that are poorly represented in the corpus, to see how the accuracy of the system is subjected to such data. As can be noticed in Figure 5, after a surprising initial drop, the system reaches higher accuracy values in presence of well-represented categories, as expected. More interestingly, while Precision remains quite stable with a filter of more than 20 necessary documents, Recall grows linearly up to 80%.

Finally, we tested our approach using a selection parameter $Q$ higher than 1 (see Section 2.2), so that one feature can belong to more than one category for each document during the mono-labelling transformation. Figure 6 presents the results with $Q = 2$. As can be seen, the accuracy of the system increases up to 61.47% in terms of F-Measure for the first 4 categories. This result demonstrates the potential of our approach as compared with the accuracy values reached in [9] and [12] with the same data. In future work, we plan to investigate novel refinements of this approach using higher values for the selection parameter $Q$, as well methods to automatically estimate the number of categories to take for each document to classify.

4. Conclusions and Future Work

In this paper we presented a novel approach to treat multi-label text classification problems. First, we introduced how to transform a generic multi-label datasets to mono-label ones. Then, we presented a slight variation which is able to improve current state-of-the-art methods, demonstrating the high potential of our proposal. We evaluated our technique by relying on standard evaluation methods on the legal domain, using a freely-available corpus of 23K documents called JRC-Acquis, and the EuroVoc european classification scheme. In future work we will concentrate on how to make use of the ontolog-
Figure 6. The difference in accuracy (only F-Measure) when using $Q = 1$ (in red) and $Q = 2$, in green. With $x = 4$, F-Measure goes from 44.51% to 61.47% (+16.96%).

The difference in accuracy (only F-Measure) when using $Q = 1$ (in red) and $Q = 2$, in green. With $x = 4$, F-Measure goes from 44.51% to 61.47% (+16.96%).


crical information contained in the hierarchical EuroVoc scheme to boost the effectiveness of the approach.

References