Automating the classification of legal questions
A preliminary assessment

Elio Amicarelli

doc ref: COBR01016

Overview

In this pilot study a supervised Machine Learning approach is used in order to predict the expertise class (Jurisdiction) of legal questions asked by customers via the legal web service jurofoon.nl. Using a small subset of the data and a simple approach we achieve a rate of 78% correct predictions. The results are very encouraging and clearly highlight the possibility to automate the categorization of legal questions submitted by ARAG's clients via jurofoon.nl.

Introduction

Modern legal companies can be easily reached by customers via phone or the internet. Customers ask questions, the questions are processed and eventually these questions are followed by the formulation of legal advices. At the moment of writing, the website jurofoon.nl adopts manual procedures in order to processes and sort the questions submitted by its cutomers. Implementing an automated system for the elaboration and classification of questions would significantly speed up the functioning of the legal service. In this report we provide evidences supporting the feasibility of this system.

Explaining the approach to a general audience

How do machines learn to recognize patterns in human texts and to associate these texts to correct categories? Broadly speaking there are two main approaches that can be used in order to let machine perform this complicated task, namely supervised and unsupervised learning. In supervised learning, we provide many examples of pre-categorized texts and train the machine to identify the existing relationships between texts and categories. After these relationships have been learned, the machine is able to independently assigns categories to new texts. On the contrary, when we adopt an unsupervised learning approach, we do not provide past examples of text-category associations; instead we directly ask the machine to identify potential categories in a set of texts with the need for the analyst to constantly and carefully check whether the identified categories make sense or not. In this exercise we use a supervised approach relying on an algorithm called extreme gradient boosting machines.

Data

Step 1 - Extracting language features

The raw data we use for this exercise is a set of legal questions. Each question is associated with further details about the related domain of legal expertise and its source. Table 1 shows an example of the typical observation contained in the raw data.

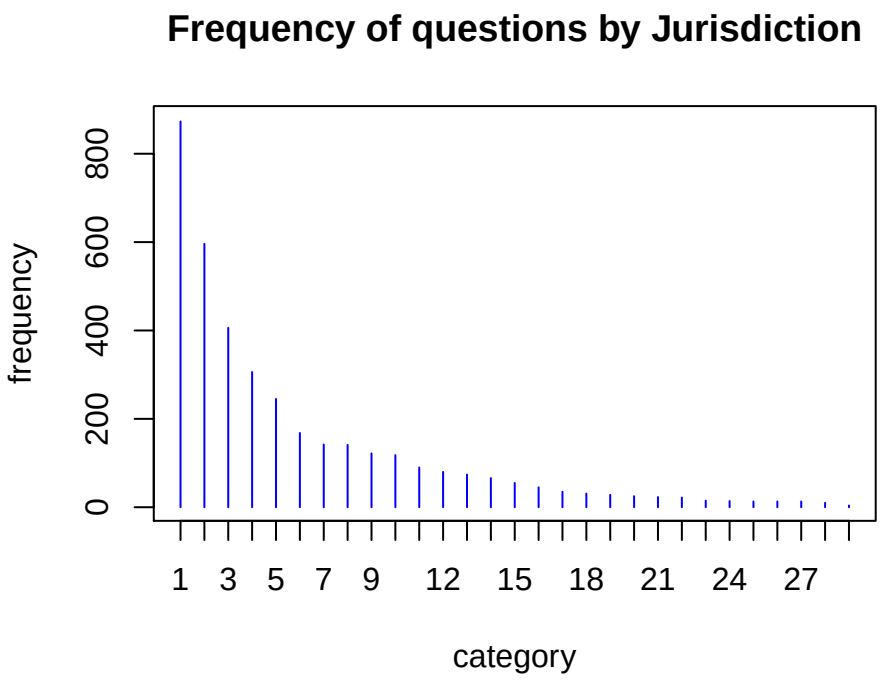
Table 1: Table 1. Example of an observation from the data

Jurisdiction (legal expertise)	Source	Summary
Verbintenissenrecht (algemeen) - (overig)	Reactie via formulier op jurofoon.nl	Wij hebben onenigheid met de buren over gezamelijke schutting. Hij communiceerd direkt via zijn rechtsbijstand. Hoe kan ik het best reageren / wat zijn mijn rechten ?

Before moving to the analysis we first transformed the description contained in the Summary field in quantifiable data. For this quick exercise a simple transformation has been implemented: after removing useless information (e.g. stop words), we stored information about the frequency of appearance for each word contained in each Summary.

Step 2 – Subsetting the data

For this quick exercise we decided to focus only on those questions collected via the jurofoon.nl website and that have a valid specification for the Jurisdiction (legal expertise) field. The graph in *Figure 1* shows the frequency for the 29 different legal expertise categories available. For clarity of visualization the categories have been numbered from 1 to 29, *Table 3* in the Appendix shows the correspondence between each number and the respective category.



Since the algorithm needs enough data in order to learn the relationships between text and categories, we are going to focus only on those categories having more than 200 questions in the dataset, which are the following:

- “Verbintenissenrecht (algemeen) - (overig)”
- “Arbeidsrecht - (overig)”
- “Personen- en familierecht - (overig)”
- “Alimentatiezaken - Alimentatie”

- "Huurrecht - Huurzaken particulieren"

To sum up, we are going to build a system that given a question coming from one of the abovementioned categories, is able to predict to which category the question actually belongs to. As already said we have chosen these categories based on the amount of data available and with the aim of providing the client with a quick example of what can be achieved by harnessing the power of predictive modelling. However, this approach can be easily extended to all other categories by using more data and more resources.

Analysis and results

The dataset we are going to use for this analysis contains 2426 observations. From these 2426 observations, 400 randomly selected observations are not going to be fed into the algorithm during the learning phase. This test set of 400 observations will be used only to test the predictive performance of the algorithm on new questions. Extreme Gradient Boosting Machines have been trained on the data described in the previous section containing 2026 observations and the best model has been selected via 10-fold stratified cross-validation.

Table 2 shows the actual and predicted classes for the observations contained in the test set. Overall, the model correctly predicted 78% of the 400 observations contained in the test set. As can be seen, the model does a fairly good job in predicting all the 5 classes.

Table 2: Columns: observed class, Rows: predicted class.

	0	1	2	3	4
0	136	22	8	9	18
1	4	80	0	0	0
2	0	0	39	0	9
3	5	0	0	23	0
4	5	0	7	1	34

Based on these results we expect that even this very preliminary model if implemented would be able to correctly predict approximately 78% of the future questions that will be received by jurofoon.nl if they belong to one of the five categories used in this exercise. With additional data and resources this predictive approach can be generalized to all the categories currently available in the database and the accuracy can be dramatically improved as well.

Appendix

Table 3: Table 3. Complementary information for Figure 1

Category id	Category full description
1	Verbintenissenrecht (algemeen) - (overig)
2	Arbeidsrecht - (overig)
3	Personen- en familierecht - (overig)
4	Alimentatiezaken - Alimentatie
5	Huurrecht - Huurzaken particulieren
6	Personen- en familierecht - Echtscheiding
7	Arbeidsrecht - Ontslag
8	Huurrecht - (overig)
9	Verbintenissenrecht (algemeen) - Incasso
10	Strafrecht - (overig)
11	Bestuursrecht - (overig)
12	Consumentenrecht - (overig)
13	Personen- en familierecht - Erfrecht
14	Aansprakelijkheidsrecht - (overig)
15	Sociaal Verzekeringsrecht - (overig)
16	Insolventierecht - (overig)
17	Huurrecht - Huurzaken bedrijven
18	Algemene praktijk MKB en zakelijk - (overig)
19	Fiscale zaken - (overig)
20	Insolventierecht - Faillissementen
21	Ambtenarenrecht - (overig)
22	Vreemdelingenrecht - (overig)
23	Intellectueel eigendomsrecht - Merkenrecht
24	Europees en internationaal recht - (overig)
25	Aansprakelijkheidsrecht - Letselschade
26	Auteursrecht - (overig)
27	Banken en effecten - (overig)
28	Gezondheidsrecht - (overig)
29	Bouwrecht - (overig)