The challenges of legal analysis between text mining and machine learning.

Maria Francesca Romano[1], Giovanni Comandè[1], Denise Amram[1], Pasquale Pavone[1]

[1] Scuola Superiore Sant'Anna di Pisa & EMbeDS Project – Italy

m.romano@santannapisa.it

Abstract

Textual analysis is useful to better understand social phenomena from different angles. In previous research (Rey 2017), we imported judgements into TaLTaC, evaluating the results obtained as a possible source of data for economic analyses of estimating the criminal economy (Romano 2018, Romano et al 2020).

We realized that a more precise analysis is possible through the interplay of different expertise: for instance, by way of developing context-based automatic queries that distinguish among texts discovered in the context of facts description or motivations in judgements.

Thus, our paper will focus on the potential combinations of techniques made possible by the rigid structure of judicial decisions, clearly contextualizing texts in different sections.

In addition, the significant amount of personal (often sensitive) data contained in rulings creates a series of technical and legal challenges linked to data protection and ethics. The analysis of judicial decisions to identify patterns of harmonization of practices, faces problems when encounters results seen from a statistical point of view: legal systems evolve thanks to the technical discretion of judges in interpreting legal rules and departing from prevalent interpretations, so a statistical outlier can be the spark of innovation and change in case law. In this paper, we illustrate the example of case-law on separation maintenance and divorce alimony in Italy; our goal is to compare the contribution of non-supervised classification and text mining approach in a legal framework.

Keywords: case-law; text mining; unsupervised classification; GDPR; pseudonymization.

Résumé

L'analyse textuelle est utile pour analyser les phénomènes sociaux sous différents angles. Dans des recherches antérieures (Rey 2017), nous avons analysé les textes avec TaLTaC, en considération des résultats obtenus comme une source possible de données pour des analyses économiques de l'estimation de l'économie criminelle (Romano 2018, Romano et al 2020).

Avec des requêtes automatiques contextuelles qui distinguent les textes découverts dans le contexte de la description des faits ou les motivations des décisions, l'article explorera les avantages et les inconvénients de l'utilisation des techniques statistiques pour l'extraction et l'annotation des données.

Ainsi, notre article se concentrera sur les combinaisons potentielles de techniques rendues possibles par la structure rigide des décisions judiciaires, avec une conceptualisation des textes dans différentes sections.

En outre, la quantité importante de données personnelles (souvent sensibles) contenues dans les jugements crée une série de défis techniques et juridiques liés à la protection des données et éthiques. L'analyse des jugements pour identifier les *patterns* d'harmonisation adresse des questions en comparaison avec les résultats statistiques: les systèmes juridiques évoluent grâce à la discrétion technique des juges pour interpréter les règles de droit, qui s'écarte des interprétations courantes, donc une valeur statistique aberrante peut être l'étincelle d'innovation et de changement de la jurisprudence.

Dans cet article, nous illustrons l'exemple de la jurisprudence sur le maintien de la séparation et la pension alimentaire en Italie; notre objectif est de comparer la contribution de la classification non supervisée et de l'approche text mining dans un cadre juridique.

Mots clés : jugements; text mining; classification non supervisée; RGDP ; pseudonymization

1. Introduction

Textual analysis is a useful tool to learn about social phenomena that can be analysed from different angles (legal, economic, social). Most text analysis applications concern newspaper texts, interview contents, social media texts, etc. or books and scientific articles (for linguistic analysis).

Any text in natural language can be treated with the well-established techniques of textual analysis, but it is less widespread the automatic processing of texts of a legal nature, such as judgments of the judiciary. This although a legal scholarship is emerging that considers the textual and statistical analysis of legal texts as a source of guidance (Livermore et Rockmore 2019). These studies are all oriented towards the attribution of a meaning to the relevance of words, for example based on their recurrence (Carlson et al 2016) or the importance of the composition of the judging body (Ash et Chen 2018). From here to tickle fantasies of predicting the outcome of a controversy, the step is short and in line with the old dream of the automated judge (Ciampi 1992).

In previous research (Rey 2017) we have addressed the phenomenon of ”corruption” using judgments containing this word, importing them into TaLTaC, evaluating the results obtained as a possible source of data for economic analyses of estimating the criminal economy (Romano 2018, Romano et al 2020).

Along this line of investigation, we realized that a more precise strategy of analysis is possible through the interplay of different expertise: for instance by way of developing context-based automatic queries that distinguishes among texts discovered in the context of facts description or motivations in judgements (Boukchina et al 2018). In this dimension, our paper will explore pros and cons of the use of statistical techniques for data extraction and annotation, comparing them to the possibilities offered by machine learning tools.

This paper presents the guidelines of our interdisciplinary research group, focusing on judgments regarding maintenance orders in case of spouses’ separation and alimony in case of divorce.

2. Background

From a technical standpoint, case-law materials offer an interesting field of analysis under several perspectives.

First of all, each judgment follows a recurrent structure which is functional to the judicial reasoning development: at the beginning of each file there are general data of the judgment (ID record, Date, Court, Court composition, plaintiff, and defendant names), case description, and the proceedings description with references to plaintiff’s requests, the burden of the proof, and then the interpretation of the legal issues at stake and their interplay with the evolution of a legal system.

Secondly, automated annotation will help overcoming the hurdles of pseudonymization / anonymization in order to consider case-law materials as a dataset to be processed and analysed with ML techniques.

Thirdly, legal systems improve through the human interpretation of given provisions by judges: to analyse trends, analogies, and differences in the judicial activities is functional to better understand a legal system and its evolution.

From this perspective, in this paper, we apply the data extraction and annotation techniques to judgments regarding maintenance orders in case of spouses' separation and alimony in case of divorce. In fact, even if these two measures are based on pre-determined legislative parameters, parameters are supposed to be interpreted and applied in the given context of the decision by the judge, while exercising her technical discretionary power. In particular, according to the article 156 of the Italian Civil Code, the maintenance order to pay a sum in the context of spouses' separation is provided whereas one of the two spouses has not an "autonomous source of income". In order to ascertain the meaning of "absence of autonomous sources of income", case law has developed several criteria related, for instance, to the need to preserve a standard of living similar to the one experienced during the marriage. However, the most recent the High Court opinions stated that this criterion could be aplied only if the plaintiff demonstrates an economic disparity. Thus, the absence of an autonomous source of income refers to the concept of state of economic hardship, which is the main requirement to gain access to the other measure, the alimony, in case of divorce under article 5 of the Italian Divorce Act n. 898/1970.

Pending separation, the maintenance order is also subjected to a comparative analysis of the other spouse's financial position and it is excluded whereas liability for separation has been attributed (i.e. *addebitata*) to the applicant.

At the occasion of divorce, which is the declaration dissolving a marriage (while separation only suspends the spouses' reciprocal obligations), alimony consists of a solidarity measure aimed at avoiding that after divorce one of the spouses remains in a state of financial distress: *i.e.* she/he cannot provide for a decent living standard.

It is evident that each criterion applied by courts could be grounded under a plurality of reasons, each of them affecting both the *an debeatur* or the *quantum debeatur (*if and how much*)*.

To allow access to alimony in a divorce proceeding, the judge has to assess and consider a series of elements. These include the contribution provided by the requesting spouse to the family life and to the common, as well as individual, economic well-being. These elements aim at preserving the principle of solidarity within the family regardless of the actual role as breadwinner among spouses: being both employed, a housewife and a worker, holding both only part-time jobs, etc. The amount of alimony will also depend on the duration of the marriage and the age of the applicant, for instance in relation to the potential ability to engage in a new working life. In both cases (the maintenance and the alimony), the *quantum debeatur* is the result of the mentioned criteria assessed on the evidence offered on trial by the parties and their appreciation by the court.

The system shows many biases. A well-known one is the gender pay gap, whose effects do not completely emerge in the described legal framework. Similarly, the role of a health impairment suffered by one of the spouses is not legally a requirement to access this measure based on solidarity. There is not a clear and nationally relevant set of guidelines. Yet, some trial courts have developed from their case law protocols which maintain a local and not a national relevance.

3. Goals

From this perspective, an automated analysis of decisions may address known bias from a very different perspective helping to uncover yet undiscovered ones. Automated textual analysis can also help identifying judicial leading precedents capable to improve the whole

system. The aim is not to predict the *quantum* but to identify and compare adjudicating models and to extract trends based on correlations of elements that normally a human analysis is not able to recognise or it is able to detect just in a limited dataset. Applying Machine Learning (ML) techniques, we aim at developing new avenues to interpret judicial materials and to propose innovative solutions to solve practical matters in and out of court. From this perspective, the challenge is to extract all the determinants of the decisions enabling the ability to anticipate possible future outcomes.

Against this backdrop, we endorse an ethical approach to developing the described predictors. The application of ML techniques to case-law must be ethically acceptable and in line with shared values (*i.e.* the EU Charter of Fundamental Rights, Nice 2000 and the recent High-Level Expert Group Ethics Guidelines for Trustworthy AI, 2019). Here we concentrate on two main ones. The first has to do with the intrinsic way in which deep legal systems evolve. Each case can be a legal mistake or the spark of innovation (Comandé 2011): law evolves also thanks to the technical discretion of judges in interpreting legal rules and departing from prevalent interpretations. Since a statistical outlier can be the spark of innovation and change in case law, clear tools must be developed to avoid the so-called automation bias, that phenomenon in which individuals rely on automation results even when they know or should know that automation can be mistaken (Comandé 2019 and 2018).

To disentangle these issues, our paper will depart from analysing the state of art in annotation and extraction techniques applied to judicial texts and the corresponding applicable ethical and legal framework in a comparative perspective. In turn, we will illustrate the necessary technical and ethical-legal specifications for developing an approach to ML techniques for the textual analysis of case-law.

A second problematic issue is the needed protection of individuals whose data are processed. Under the GDPR (EU Reg. 2016/679 General Data Protection Regulation) data subjects enjoy specific rights that are expanded when their data appear in judicial files and decisions. Here again, our aim is to expand as much as possible the use of personal data in our research fostering the highest level of protection possible for data subjects. To reach this goal we blend techniques, organizational measures and reasoned interpretation of legal rules, considering that the use of ML techniques model must be lawful, ethically acceptable and robust.

In this framework, we are discussing and testing a solid strategy for the automated detection of expression semantically equivalent in order to annotate their presence in each judgement and proceed in obtain results from statistical models. Notice that the strategy we present in the following section is still preliminary: it is a proposal and it needs further testing to define it as a robust and responsible one.

4. Data and Methods

Thanks to an agreement between the Scuola Superiore Sant'Anna and the Tribunale of Genova (the first instance trial court), we obtained the judgements about separation maintenance or divorce alimony orders issued by that court. Our analysis focuses on the ones issued between 2014 and 2019 in order to have a sufficiently large set of data and a controlled time sufficient to monitor eventual changes in operational rules at trial court level leveraged by the Supreme Court's most recent opinions on our research topics.

The 5075 judgements, made available in PDF format, were transformed into text files, in a format compatible for the import into the software TaLTaC[1], used for the analyses.

In summary, our analysis strategy involves a series of (supervised) steps using TaLTaC resources and tools, with the aim of simulating the behavior of an expert who reads and annotates the judgments: in the end, the judgment database is exported and it is possible to apply statistical models to summarize the results and / or estimate the results of the judgments.

The judgements' texts are written with a well-known structure: Introduction, Fatto/Diritto (Facts' description) and PQM (namely *Per Questi Motivi*, For These Reasons, the section that introduces the decision); this structure helps in detecting information hidden in the text (for example, in the Introduction there is the list of the *Parti* (the spouses) and in PQM there are information about the judge's decisions. Inside this structure each writer uses many different words (or multiwords) with the same semantic meaning.

The successive (and recursive) steps of analysis, using different software and different datasets, are described in Figure 1 and reported in detail in the following subsections.

<table>
<tr><td colspan="2">Analysis by steps</td></tr>
<tr><td colspan="2">1. Import the Original Judgements and create the DataWarehouse</td></tr>
<tr><td colspan="2">2. Information Extraction from personal data
Variables obtained: age class, gender, country of birth from taxpayer's code</td></tr>
<tr><td colspan="2">3. Query to detect personal data and annotation to obtain Pseudonymization / Anonymization</td></tr>
<tr><td colspan="2">4. Control for False Negative
An expert reads a sample in order to detect false negative; if errors are detected, correction of the queries and repeat from Step 3 to 4</td></tr>
<tr><td colspan="2">5. New Data Warehouse with Anonymous Judgements</td></tr>
<tr><td>6A. Semantic Tagging (using results of Queries for Information Extraction)</td><td>6 B. Lexical and textual analysis</td></tr>
<tr><td>7A Semantic tagging Test
An expert reads a sample in order to detect false negative and false positive; if errors are detected, correction of the queries and repeat from Step 6A to 7A</td><td>7B Multidimensional analysis on the Matrix words by judgements (PQM section)</td></tr>
<tr><td>8A. Creating Variables for judgements</td><td>8B Cluster analysis</td></tr>
<tr><td>9A. Export of Matrix of judgements by variables</td><td>9B Export cluster's number by judgement</td></tr>
<tr><td colspan="2">10. Merge of the matrices obtained in 9A and 9B</td></tr>
<tr><td colspan="2">11. Multiple Correspondence Analysis on the matrix and Comparison analysis</td></tr>
</table>

Figure 1 Strategy proposed for the judgements' analysis

The steps from 1 to 4 have the aim of pseudonymization / anonymization of personal and sensitive data (see 4.1 below). In the following steps the new datasets are free from any sensitive information (personal names, taxpayer's code, date of birth, residence).

Notice that the iterative steps 6A-8A are applied to create variables for each judgement; in this case the iteration refers not only to the errors' correction cycle but also to the information for all the variables needed.

4.1. Anonymization and pseudonymization of sensitive data

1 Both, version 2 of TaLTaC and version 4 (in the experimental phase) were used, the latter made available by prof. Sergio Bolasco within the EMBeDS Project of the Scuola Superiore Sant'Anna di Pisa.

The significant amount of personal (often sensitive) data included in the analysed rulings creates a series of technical and legal challenges linked to personal data protection.

We endorse the idea that all techniques applied must be pursued following the principles of privacy-by-design and privacy-by-default. In addition to this polar star the dataset has been cured following the relevant technical and organizational measures provided under article 89 EU Reg. 2016/679 and the Italian national implementation (Amram 2020). From this perspective, a series of organizational measures like specific training activities, authorizations, and confidential obligation templates have been implemented within the research team in order to ensure compliance with the current legal framework during the research activities. In addition, the dataset is fully contained from external access and processed locally to minimise any data breach risk.

Furthermore, an infrastructure protected by encryption techniques has been established to store the judgments before being processed, in order to ensuring the availability, confidentiality, and integrity of the collected data according to the results of the data protection impact assessment performed by the scientific coordinators of the project.

The first step we pursued was to pseudonymise data in order to process them for research and statistical purposes in line with art. 89 GDPR prescriptions. A full anonymization approach proved impossible to pursue since it would have rendered almost impossible the achievement of the specific purposes of the research. In fact, our goal, here, is twofold. On the one hand, we aim at establishing a reliable and legally compliant dataset for the specific research purposes. On the other hand, we aim at establishing a viable blend of procedures, safeguards, and protocols to be generalized in statistical and text mining of legal documents containing personal data.

Pseudonymization, is a technical measure that allows to process data without identifying the data subject: it consists of the replacement of ID information with a common code. For example, A for plaintiff (in Italian *Attore*), C for Defendant (in Italian *Convenuto*), X for minors or third parties, T for Witness (in Italian *Testimone*) etc. Other data like the place and date of birth or the taxpayer's code are simply deleted, preserving only general data (gender, age class and country of birth of the parties) because they can be relevant for the judicial reasoning (e.g. the age of the parties could be relevant to quantify the alimony).

For this reason, we established this protocol balancing the level of de-identification with the ability of not losing crucial information. We decided to use TaLTaC as a tool to easily identify some personal information and their removal/masking.

In this phase, we opted for leaving as graphical forms the words starting or containing capital letters (tagged as MAJ); using grammatical and semantic tagging, it was easy to identify as Multiwords tagged as Names. It is worth of notice that in the sampling control we detected also names of firms (in reason of property or even as individual firms) that we included in the anonymization steps.

For the taxpayer's code we tagged all graphical words of length 16 with the tag CFISC, detecting also its writing in 3 or four separate words (blanks inside), identifiable as CAT(MAJ).

All these different graphic modes to write the taxpayer's code were detected during the control step, where no taxpayer's code was detected. In the same step, we also detected a relevant number of taxpayer's codes (referred to the lawyers) and deleted them, because of no

relevance for our analysis. *Viceversa*, we preserved four variables with gender, age class and birth's country, as they are functional to our research outputs (Amram 2020).

4.2 Judgements Data warehouse

The data warehouse (Corpus) in TaLTaC is made up of two datasets:
1) The Vocabulary (matrix of graphical forms and columns, such as the number of frequencies (occurrences) and other annotations (semantic and grammatical).
2) The Judgements matrix (5075 x K variable size matrix) is progressively updated with the results of the text mining actions and forms the basis of the subsequent analyzes (see the following paragraph 5.3).

The Vocabulary has 119.933 rows for a total of 6.087.327 occurrences. The judgement' length is not large: the minimum is 280, the mean is 1204, and the maximum value is 29.709, with a median value of 927.

4.3 Text Mining and Information Extraction Queries

From the steps 6A-8A we obtained a set of variables (Table 1) that can help us to identify and measure the conditions and situations described by the divorce judgements.

Table 1 Variables obtained from Text Mining of Judgements

Variable	**Values**	**Type**
Y_i : amount of money by month	1-max	Real (€)
Z_i : amount of money as lump sums	1-max	Real (€)
X_1 : low formal education of spouse1	1 when first level formal education; 0 otherwise	dummy
X_2 : low formal education of spouse2	1 when first level formal education; 0 otherwise	dummy
X_3 : high school education of spouse1	1 when second or tertiary level formal education; 0 otherwise	dummy
X_4 : high school education of spouse2	1 when second or tertiary level formal education; 0 otherwise	dummy
X_5: Presence of underage children	1 when X_9>5 ; 0 otherwise	dummy
X_6: Spouses born in Italy	0 if both Foreign ; 2 if both Italian; 1 otherwise	integer
X_7: spouse1 age class	five-year age classes	ordinal
X_8 : spouse2 age class	five-year age classes	ordinal
X_9 : number of underage children	0-max	integer
X_{10} : number of adult children dependent	0-max	integer
X_{11} : number of adult children independent	0-max	integer
X_{12}: Presence of children	1 when X_9>5 OR X_{10} >0 OR X_{11} >0: 0 otherwise	dummy
X_{13}: Not Consensual Divorce	1 when CATSEM(*contenzioso*) is TRUE; 0 otherwise	dummy

As an example, the process obtaining the dummy variable "Figli" (sons and daughters) started from the graphical forms like "figl*" and exporting the list for the semantic tagging. During the control step, other graphical forms were added (like *figila* instead of *figlia*, for example), and it was added another semantic tag "NoFigli" in presence of a list of MW like "*dal matrimonio non erano nati figli*" or "*in assenza di figli*"[2].

2 During the control check, we decided to sample the judgements with the presence of values for the variable "NumFigli" (number of the same semantic tag in the same judgment) and with the presence of the tag "NoFigli": a systematic sampling with step 30 (on the about 1000 sentences with "NoFigli") with expert reading revealed the complete accuracy of the algorithm, being the "Figli" cited (but with a maximum value of 5) born from (previous) marriages of one or two spouses.

As another example, to detect the amount of alimony or maintenance, we started with a query listing the segments before or after a value tagged by TaLTaC as NUM and containing a graphical form for "€", searching also for time indication (month, year, una tantum).

5. Preliminary Results

5.1 Results from non-supervised classification

In order to classify the sentences with respect to the issues present in PQM, we first proceeded with a lexical-textual analysis, to define and select the keywords. This step is characterized by grammatical tagging and identification of multiword expressions (MWEs). In particular, the 3.998 keywords were selected in the previous step of lexical analysis, through grammatical features, selecting only the forms noted as nouns and adjectives. Then, the identification of the main specializations contained in the Corpus is carried out through a combination of correspondence analysis and cluster analysis. Multidimensional analysis allows us to observe the similarity of records based on their lexical content. The word-based clustering phase is a non-supervised classification that reflects the semantic similarity among records. This conceptual homogeneity reflects the semantic level of similarities between the different PQM. The conceptual homogeneity expresses the semantic theme, or semantic trait, prevalent in that group of fragments, which can be summed up in a category, not defined before but obtained through this analysis. Accordingly, 3.398 words and MWEs were selected and used to define the textual matrix records × keywords (5.075×3.398), to be processed through a factor analysis. Cluster Analysis on the results of the factor analysis disentangles 5 clusters of PQM.

5.2 Results from Text Mining and Information Extraction Queries

The goal of our methodology is to identify the situations or conditions that may affect the final decisions on separation and divorce maintenance and alimony (e.g. children, adult children dependents ones), the socio-economical context (nationality, education, age of spouses), the judicial procedure (consensual separation/divorce -where spouses file the request to validate an agreement before the court- or dispute one – where the conditions on marriage dissolution are stated by the judge) and nature of the adopted economical measure (monthly sum or lump sum). We can measure the impact of the family composition (i.e. presence of children in 3196 judgments), the reasons upon the refusal to award the spousal support (spouse maintenance or alimony) in 1630, and the ones upon the different amounts awarded in the other 3345 judgements.

According to our data, the presence of children increases the awarded amounts for spousal support, despite the fact that children's maintenance is a different measure, with different requirements to be met. From this perspective, the statistical model emerging from this research will be able in the future to estimate how all the discussed legal conditions are concretely affecting the "horizontal support" for marriage dissolution (namely maintenance and alimony) in Italy. In fact, from the annotation algorithm we obtained other elements like the nationality of the spouses (Italian or foreign citizen), the nature of the measure (lump sum or monthly one), level of formal education and the number of children that combined with the composition of the family have affected the judicial reasoning and improved the legal system.

5.3 Comparing Results from Text Mining and non-Supervised Clusterization

In the steps 10-11, we combined the obtained results: the matrix judgment $_x$ variables issued by TaLTaC has been integrated by the cluster to which each judgment belongs. This matrix

was the input for an MCA in which the cluster was considered as supplementary variable. The graph of the 2-dimensional solution is shown in Fig. 2; the first dimension reflects the presence of children (YesChds on the negative axis of the abscissa; and NoChds on the positive axis), while the second differentiates the presence of minor children (Yes Minor on the positive axis of the ordinate) and the divorces between foreigners / Italians.

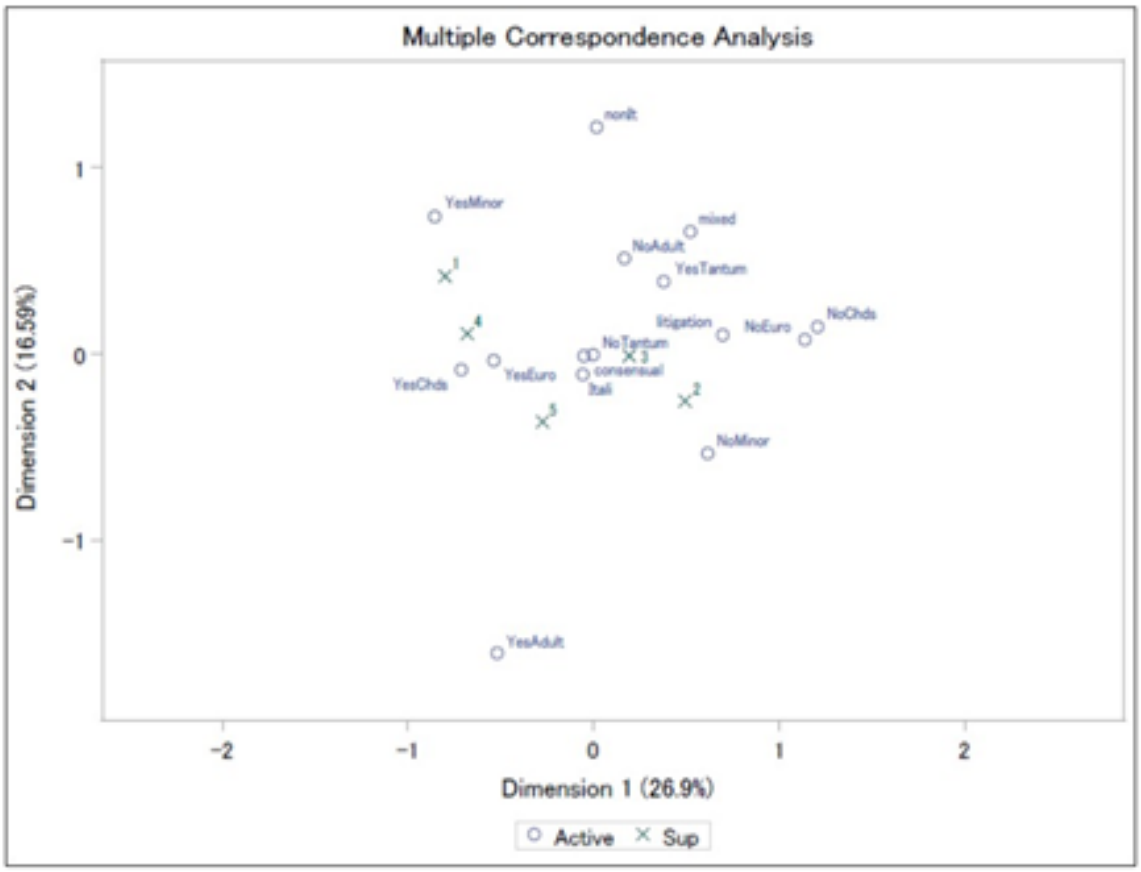

Fig. 2 Configuration of MC Analysis

The position of the clusters in this configuration leads us to expect that in cluster 1 there will be a greater presence of minor children, while in cluster 2 the opposite position should be expected; the variable that identifies the presence of adult children is distant from all clusters, but closer to cluster 5.

To assess whether and how the two types of analysis converge, the result of the unsupervised cluster was examined, identifying the subgroups within each cluster based on the variables used for the MCA, and also taking into account the non-euro values present in the clustered matrix. For each of the 5 clusters, Table 2 shows the number, a description of the cluster based on the most influential words, the number of the largest number subgroups with equal values on the variables deriving from the annotation phases with text mining. The "updated" description of the cluster is summarized in the last column.

For the clusters 1 and 5, the variables derived from text mining are described by 4 and 3 subgroups which respectively include 89% and 86% of the sentences that form part of it.

Leaving aside cluster 4 (consisting of only 34 sentences), in cluster 2 (the one with the highest number) only in 6 subgroups about 88% of the sentences are described, while for cluster 3 the sentences are less clearly defined with the 5 subgroups identified.

Table 2 Comparison between results from cluster analysis and variables from text mining

Clu - ster	**N**	**Cluster analysis based on lexical similarity**	**Subgroups (k - n - %)**	**Interpretation from variables extracted through text mining**
1	**1465**	Judgements with minors or older dependents with the agreements and allocation criteria description	**4 - 1298 - 88,6%**	This cluster includes consensual divorce or separation procedures (i.e. settled with an agreement found or validated before the court). In 9% of cases one spouse is Italian and another one has a foreign nationality, in 2% both spouses are foreign citizens. Considering the variable "children" in 18% of cases, families include adult children dependents, while in 85% they are minors. In 9 % of cases, agreements refer only to personal relations, without involving economical aspects. Where one of the spouse is a foreign citizen, the amount of the maintenance/alimony is lower, as Q3 goes from 500 to 645€. In this cluster, a lump sum is never awarded.
2	2047	Judgments with older children and specific agreements on expenses and obligations.	**6 - 1791 - 87,5%**	This cluster includes in 98% cases consensual separation or divorce procedures. In 11%, one of the spouses is a foreign citizen; 3% both spouses are foreign citizens; in 60% families are composed only of the couple; in 27% the couple has also teenagers/adults; in 14% there are minors. In 52% cases there is money transfer in the decision. The sixth group refers also to lump sums provisions. In case of one foreign citizen with minors, monthly cheques are lower also in this cluster, despite the cases that include 16 judgments where the family is composed of one foreign spouse and adult children or teenagers where amounts are higher (Q3=800€), including lump sums.
3	**1291**	Conflictual disputes	**5 – 881 – 68,2%**	This cluster includes 76% of judgments that decide a conflictual dispute on separation and divorce. In 40% of cases, there are no economical provisions. In 38% of cases, families also include minors, in 25% teenagers or older children, in 25% they refer only to the couple. In 3% cases, spouses are both foreign citizens, in 10% one of the spouses is foreign citizen.
4	**34**	Judgments with specified agreements on expenses (medical ones, educational ones, sport-related ones etc.)	**2 – 26 – 76,5%**	Only consensual separation and divorce procedures are included in this cluster. Families are composed of an Italian couple in 94%, with minors (71 %) and 26% teenagers and older children. In 85% cases the judgment provides a monthly cheque, no lump sums are included.
5	**238**	Judgments real-estate details related	**3 – 205 – 86,1%**	In this cluster, judgments are settled or introduced with a consensual procedure, unless 2 cases. In 91% of cases, couples are Italian citizens. Families composition is very well balanced (32% no children, 38% with teenagers/older dependent ones, 32% with minors). Only 7 decisions do not provide money transfer, and they are all concerning adult children.

We are aware of the need to refine our semantic annotation and tagging algorithms, and yet we think that these very preliminary results demonstrate the potential of the proposed analysis strategy, as the possibility to merge the different variables provides a unique opportunity to analyze both the social and economic dimension of the legal tools applied to manage the crisis of the family.

For example, the textual analysis identified 5 clusters, but through the text mining analysis that may include additional variables, we may identify interesting social and legal features to interpret the case-law. In particular, information related to the family composition described in a cluster based on legal term may design a pattern within the family dynamics. This is a unique contribution on policy-making and law-making.

For instance, more than 52% of the decisions include families with teenagers or older dependent ones (i.e. teenagers or >18 years old still economically dependent to the parents with higher expenses to be allocated, clusters 5). In these cases, it seems that agreements, also reached during the judgment, prevail on the "pure" judiciary conflictual solution. We may assume that once children have grown up, couples are more open to settle the economical profiles emerging during the crisis of their horizontal relationship. Therefore, in terms of policy-making, the target group to be sensitized for alternative dispute resolutions (family mediation, collaborative law) could be younger couples with younger children.

Another example might be given by cluster 1. In about 26% of cases the judgment on spousal maintenance and alimony are issued in the context of personal provisions also on child custody (visiting rights, education, care-related issues). This means that vertical issues (i.e. personal and economic issues related to parent-child relationships) affect "horizontal issues" in terms of requests and in terms of final provisions. It would be interesting to verify whether the algorithm may also express if they have a monetary impact on the maintenance/alimony amounts in order to address specific measures to make the separation and divorce with children more sustainable both for horizontal and vertical relationships.

Again, in disputes where one of the spouses is absent, because he/she has disappeared (in another country for example or in case of extreme discomfort), the algorithm might be able to analyze whether or not this scenario reflects a specific trend in socio-economic terms (like lower education or different nationalities etc.).

Another interesting result refers to the real-estate related issues: case-law may include either conditions where the maintenance/alimony is affected by the existence of a mortgage distribution (in this scenario it is interesting to verify conditions that may replace/reduce the monthly cheque for maintenance or alimony) or conditions where a provision that includes a property transfer between (former) spouses may act as lump sum. The impact of such an analysis on tax law and alternative dispute resolution models is a relevant output.

6. Discussion and further research

These preliminary results obtained are interesting: first of all, the identification of a pseudonymisation protocol through the text annotation techniques could be easily applied to any big dataset including sensitive data.

Secondly, from the given case-study on the maintenance and alimony measures, the annotation protocol outlined interesting insights from the combination of the analysed queries in terms of policy-making like the difference between the lump sum and the monthly one or the percentage cross borders couples. These outputs may play an essential role in terms of policy-making, procedural protocols and also to stimulate reforms.

In this regard, the Italian Parliament is currently analysing a bill to reform the alimony provision (AC 506 bill). In particular, according to the new proposal, the state of need will be

deleted, highlighting the discretionary power of the judge to evaluate the concrete positions of the spouses in light of new criteria, that shall be introduced in order to overcome some bias which make the current legal framework no more suitable to fulfil the so called post-marriage solidarity (Amram 2019). Considering the evolution of the societal and family relationships, in fact, alimony shall be determined in light of the family contribution considered as the role of each spouse on children as well as elderly's care. Health conditions and age of both former-spouses shall have a relevant position in the assessment of individual and common assets. In this perspective, the true history of the couple will be assessed in terms of impact on the individual assets after the divorce in order to re-establish a kind of economic equilibrium between the (former) spouses.

Our analysis, using both text mining and unsupervised techniques, could be useful to analyse society in the physiological development of private relationships starting from the judicial disputes that represent the pathological one. Therefore, it promises significant developments in several private law domains (tort, contract, property, consumer, insurance, family law etc).

References

Amram D. (2019), *Lo scioglimento dell'unione civile: assegno di mantenimento*, in A. Cagnazzo, L'assegno nella separazione e nel divorzio, Milano, 2019, 93 ff.

Amram D. (2020), *Building up the "Accountable Ulysses" Model. The Impact of GDPR and National Implementations, Ethics, and Health-Data Research: Comparative Remarks*, Computer Law & Security Review, CLSR_105413.

Ash E. et Daniel C. (2018), *What Kind of Judge is Brett Kavanaugh? A Quantitative Analysis*. In Cardozo L. Rev. De Novo, 70-100.

Boukchina E., Mellouli S. et Menif E. (2018), *From Citizens to Decision-Makers: A Natural Language Processing Approach in Citizens' Participation*. In *International Journal of E-Planning Research*, vol. 7, no. 2, pp. 20-34.

Keith C., Livermore M.A. et Rockmore D. (2016), *A Quantitative Analysis of Writing Style on the U.S. Supreme Court*, 93 Wash. U. L. Rev. 1461 (2016).

Ciampi C., *La ricerca «concettuale» e quella «testuale» nella documentazione giuridica automatica: un antico problema*, Informatica e diritto, XVIII annata, Vol. I, 1992, n. 1-2,

Comandè G. (2011), *Legal comparison and measures: it's logic to go beyond numerical comparative law*, in *Studi in onore di Aldo Frignani, Nuovi orizzonti del diritto comparato europeo e transnazionale*, Napoli, Jovene, p. 173.

ID (2017), *Regulating Algorithms' Regulation? First Ethico-Legal Principles, Problems, and Opportunities of Algorithms* in (T. Cerquitelli, D. Quercia, F. Pasquale eds) *Towards glass-box data mining for Big and Small Data*, 169 ff

ID (2018), *Responsabilità ed accountability nell'era dell'Intelligenza Artificiale* in "*Giurisprudenza e Autorità Indipendenti nell'epoca del diritto liquido*", a cura di F. Di Ciommo, O. Troiano, La Tribuna S.R.L. ISBN: 9788893179799, pp. 1001-1013

ID (2018). *The Rotting Meat Error: From Galileo to Aristotle in Data Mining?* In *European Data Protection Law Review*, 4 (3), 270-277.

ID (2019), *Intelligenza artificiale e responsabilità tra liability e accountability. Il carattere trasformativo dell'IA e il problema della responsabilità*, in *Analisi Giuridica dell'economia, 169 - 176.*

ID (2019), *Multilayered (Accountable) Liability for Artificial Intelligence*, in *Liability for Artificial Intelligence and the Internet of Things*, Lohsse S., Schulze R., Staudenmayer D. (eds), Hart Publishing Nomos, 165-187.

Livermore MA et Rockmore D. N (eds) (2019), Law as Data, Computation, Texts& the Future of Lefal Analysis, SFI, Press.

Rey GM (a cura di) (2017), *La mafia come impresa. Analisi del sistema economico criminale e delle politiche di contrasto*, FrancoAngeli Editore, Roma.

Romano MF (2018), *Methodological innovations to estimate illegal economy*, invited contribution to the ISTAT XIII National Conference, Rome 2-4 July 2018.

Romano MF, Baldassarini A, Pavone P (2020), *Text Mining of Public Administration documents: preliminary results on judgements*, in: Iezzi D.F., D. Mayaffre, M. Misuraca (eds) (2020) ADVANCES AND CHALLENGES IN TEXT ANALYTICS Springer.