IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 February 2024, accepted 5 March 2024, date of publication 11 March 2024, date of current version 15 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3375766

RESEARCH ARTICLE

Explainable Artificial Intelligence for Agile Mediation Propensity Assessment

ENRICO COLLINI, PAOLO NESI, (Member, IEEE), CLAUDIA RAFFAELLI, AND FRANCESCO SCANDIFFIO

Distributed Systems and Internet Technology Laboratory (DISIT), University of Florence, 50139 Florence, Italy

Corresponding author: Paolo Nesi (paolo.nesi@unifi.it)

This work was supported by the Italian Ministry of Justice, with a collaboration between universities and courts for the ASSE I, Obiettivo Specifico 1.4, Azione 1.4.1 of the National Operative Program (PON) Governance and Institutional Capacity 2014–2020, with the aim of producing a better organization of the judicial machine.

ABSTRACT Italian Justice has recently added mechanisms to exploit mediation process. One of the most critical aspects is a reliable identification of litigations which can be successfully mediated outside court procedures. The decision is under responsibility of a judge/court who has to read hundreds of pages and several documents, to be able to take a decision on the basis of few statements. This paper describes both an artificial intelligence solution and a tool to provide a decision support system which could process documents and be capable to: (i) produce reliable suggestions, (ii) produce circumstantiated motivations, thus highlighting statements which could support identified suggestion focusing the work of any judge/court on actual statements and documents with relevant facts, and (iii) provide a web based tool producing suggestions and motivations on demand at service of the involved court and judges, compliant with privacy and security, as to data. To this end, AI and eXplainable AI technologies have been used and a solution has been obtained which meets the above-mentioned objectives and many other detailed requirements. Such a solution has been developed in the context of the research project "*Giustizia Agile",* funded by the Italian National PON Governance and Institutional Capacity, and validated against real cases. The solution has exploited the Snap4City framework for data and AI/XAI management.

INDEX TERMS Artificial intelligence, explainable AI, mediation propensity assessment, decision support system, shortening justice procedures.

I. INTRODUCTION

The Italian justice system is one of the slowest in Europe. According to the report of the European Commission for the Efficiency of Justice (CEPEJ) [1] published in 2022 and referring to year 2020, the main cause of inefficiency in the Italian justice system is the excessive length of court cases, mainly concerning civil and commercial litigations. The measure used by CEPEJ to compare the celerity of judicial systems in different EU countries is called Disposition Time. Despite its name, the Disposition Time index is calculated out of the ratio of the number of pending cases, with respect to the number of solved cases along an observation period, which is typically a year. This index is equivalent to the number of days needed to solve pending cases in that specific court, according to its capability. Although the index has gradually decreased since 2012 until 2018, Italy has recorded the highest Disposition Time in the European Union for the first instance of civil cases, with a Case Duration of as many as 674 days compared to a European average of only 237. A slow justice system may negatively affect any other elements such as the economic growth and the possible appeal of our country in the eyes of foreign investors, and it may also reduce any citizens' trust in institutions, thus nurturing a more than likely criminality increment [2].

For this purpose, as to civil trials, one tool to be used, whenever appropriate, is mediation. Civil mediation is the activity carried out by an accredited, third party - external to the trial and impartial – whose aim is assisting both parties involved in a litigation, while searching for an amicable agreement bringing forth the dispute conclusion. Aside from those cases where Italian law imposes an obligation to attempt mediation before any trial start, a judge may invite parties to attempt mediation at their discretion [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Xianzhi Wang.

If a mediation is successfully concluded, this lightens the workload of courts by removing the need to start or bring a trial to its conclusion; on the contrary, as a result of a failed mediation attempt, the case returns to trial after a given time (usually months), which only ends up adding more time to the trial length. According to the CEPEJ report, in 2020 there were 60,110 mediation attempts in Italy: only 15,013 of them ended in an agreement, whereas, in all other instances, cases returned to court and the time spent trying to reach an agreement contributed to a further duration of the litigation process. If mediation has to be effective, what is needed is an accurate assessment of the current propensity shared by parties to reach an agreement.

To this end, the research described in this paper has aimed to develop a decision support system capable to provide additional information to any judge, in order to determine if a specific dispute could be reasonably solved through mediation. For example, with a classification in classes ''non-propensity to mediate'', ''propensity to mediate'', and ''neutral.'' When propensity for mediation is identified correctly, this can be of great help to avoid managing any dispute into complex and long court mechanisms, thus ensuring a faster dispute resolution, which would reduce the workload of courts and judges. On such basis, the instrument should provide:

- **reliable suggestions** for determining when mediation can be successful. This implies to provide a judge with a score about any likeliness by the parties about accepting a mediation process as a way to find a mutual agreement.
- **circumstantiated motivations** behind any provided suggestion. This implies to provide a judge with some clear evidence about reasons why parties involved in a dispute should be motivated to mediate mutually, for example, by stressing both statements and phrases in any official document that would lead to infer that propensity.
- **web based tool producing suggestions and motivations on demand** at service of both court and judges involved. This implies to integrate the solution in the context of the workflow procedure and instruments available and adopted in Italian Courts. The solution has to respect data privacy according to the GDPR European Union General Data Protection Regulation 2016/679 [4]**.**

For this reason, this current research has focused on the use of artificial intelligence (AI) and NLP (natural language processing) techniques to develop a decision support system at service of both courts and judges. This research activity has developed a solution to process legal documents, decompose them, and use AI and specifically BERT derived (Bidirectional Encoder Representations from Transformers) [5] techniques, to identify, within the large set of documents and statements related to disputes, if there are some elements to push in the direction of mediation, and its related grounds. To this end, our activity focused on: (i) developing a specific data set (training, validation and test sets); (ii) defining a model by using fine tuning techniques on top of a pretrained BERT model for Italian language, to perform a classification of the text and detect the possibility of mediation or its absence, (iii) developing a solution to exploit Explainable AI technique, XAI, namely Shapley approach [6] as tool for providing motivations behind the model; (iv) developing a method and a decision support system as a web based tool for providing on demand responses to courts and judges as to documents which they have on hand, so that it becomes possible to reduce the time needed for human based document analysis. Finally, the same instrument of (iv) has to be able to collect suggestions and comments to further improve its related model (ii) and its related data set (i) for the next version of the solution. More specifically, such a tool has been called XAI4MA (*Explainable Artificial Intelligence tool for Mediation Agile*).

Both approach and solution have been validated by a group of mediation experts, affiliated to the Law Department of the University of Florence. They input a number of court records and evaluated results from our proposed XAI4MA and provided useful feedback. This work has been developed in the framework of the research project ''*Giustizia Agile'', (agile justice)* funded by the Italian National PON Governance and Institutional Capacity, so as to achieve the result of a better organization of the legal machine. The solution we proposed has exploited Snap4City framework for data and AI/XAI management [7].

The paper is structured as explained in **Figure 1** below, which shows the data flow of our paper. **Section II** describes the background and its related work. **Section III** outlines both requirements and goals of the system being produced. **Section IV** discusses the adopted techniques to create our dataset, including data gathering, preprocessing, labeling of sentences, normalization of each example by conforming abbreviations or dates; splitting sentences, string chunking, and finally preparing data sets for training, validation and test. **Section V** details the system architecture, thus giving a brief description of the BERT [5] model used as a foundation and outlining the fine-tuning techniques used for learning the sentence classification task. In addition, the focus is set also on the approach used to obtain document-level propensity classifications, based on predicted scores of sentences.

Furthermore, model explainability methodologies based on Shapley approach [6] are described in **Section VI**. Finally, **Section VII** presents the extension of the model and solution at document level. In **Section VIII**, the online tool for decision support at disposal of any court is presented. The tool has been called XAI4MA and has a graphic user interface designed to enable decision makers in the Department of Justice to exploit AI/XAI solution in different procedure phases, as an expert to be consulted on demand. This tool enabled the collection of additional data that have been used to perform an additional validation of the solution, as reported at the end of **Section VIII**.

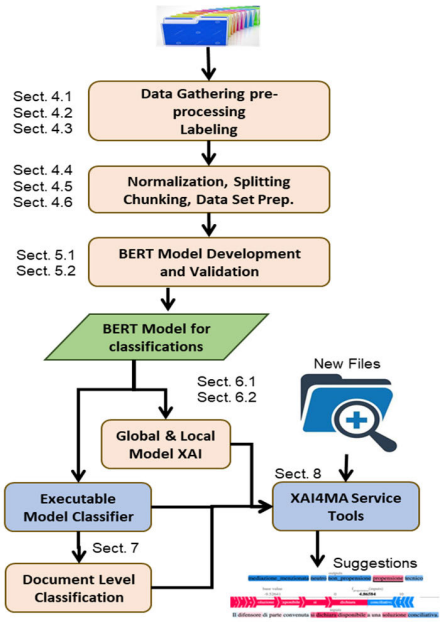


FIGURE 1. Data and process flows in the paper, with details about the section where the different phases/tasks are addressed.

II. RELATED WORKS

Most commercial solutions meant to assist lawyers have focused on both indexing any information contained in some closed trial records and allowing to search useful information by means of natural language queries. Examples are Jurimetria [8] and Predictice [9], where it is possible to obtain an estimation of the amount of compensation obtainable under a specific judge. Other systems have focused on automating the analysis of any legal documents, thus predicting the judges' decisions about a particular trial [10]. A selection of such works is reported in Table 1. For example, Katz et al., [11] proposed a system to predict any decisions made by the U.S. Supreme Court using ensemble learning based techniques. Court cases have been modelled with up to 240 variables and most of them were categorial. The used random forest could achieve a 70.2% accuracy in terms of classification of case outcomes. Alghazzawi et al., in [12] used a long short-term memory [13] plus convolutional neural network [14] (LSTM +CNN) model to also forecast lawsuit verdicts for the US Supreme Court. These researchers have applied an oversampling, in order to handle any unbalanced dataset and a feature selection process before training and validating the proposed solution. The model achieved an accuracy on the test set equal to 92.05%. Medveva et al., in [15] exploited Big Data analytics on the judgements carried out by the European Court of Human Rights (ECtHR) to predict whether a case could be judged as a violation of human rights or not. This binary classification problem was assessed by the authors via Support Vector Machines (SVM) [16] and achieved an accuracy of 75% in predicting the violation of 9 articles. Aletras et al., [17], proposed a prediction approach for the ECtHR decisions. The developed AI model was a Support Vector Machines (SVM) where training data consisted of textual features extracted from given cases, while as output there was the actual decision made by judges. The developed statistical NLP framework achieved an accuracy on the test set of 79%. Hsun-Ping Hsieh et al., [18], instead of focusing on classifying final case decisions, have focused on the possibility of mediation success between parties. Authors have proposed a system to predict the success of mediation requests using textual information and case properties, such as location of the dispute, mediator ID, number of participants, etc. The goal is burden reduction on the courts. Results have been obtained by using an LSTM-based framework, called LSTMEnsembler capable of predicting mediation results by assembling multiple classifiers. Results on the test set achieved an Accuracy of 78.8%. Authors have stated that in terms of future developments, they would like to focus their attention on explainability aspects of their proposed framework. One major deficiency of current AI legal reasoning approaches is that they are unable to give a justification of their reasoning in terms of appropriate legal concepts [19]. Branting et al., in [20] focused their efforts on the proposal of an explainable legal decision prediction support system. This solution aimed at highlighting the most relevant portions of case text, which can be considered as the most predictive ones, when it comes to final decisions. Enabling explanations on AI models (also referred as Explainable Artificial Intelligence XAI) for the legal domain will allow judges to maximize their possibility of identifying errors and biases within any algorithms as reported in [21].

Turan et al., in [22] used XGBoost to assess the decisions taken on cases by the Constitutional Court of the Republic of Turkey in a XAI framework using an XGBoost model reaching 93.84% of Accuracy with an interpretable solution. Please note that the European Union's General Data Protection Regulation (GDPR) contains provisions requiring what has been termed as a "right to an explanation." [23].

The field of text analysis has been revolutionized by LLMs (Large Language Models), which is becoming a de facto standard for many tasks and applied also towards other types of data. In fact, in [24], the authors proposed a LLMediator, a LLM based solution for dispute resolution towards the mediation of parties. This model has turned out to be fine-tuned for classification tasks, thus delivering high accuracy, and at present without XAI support.

One of the key aspects to be taken into consideration when developing AI models towards classification tasks is the possibility of introducing bias. In machine learning, a model is affected by bias, when it produces results that are systemically jeopardised due to erroneous assumptions in the decisional process [25]. Although any final decision remains under a judge's responsibility, it is important that the system does not mistakenly produce biased suggestions that are corrupted by discriminatory elements embedded within the training dataset and/or by the model and data/feature selections. A review by Geraghty et al., in [26], has analyzed tools to predict any likelihood of recidivism in people

TABLE 1. Related works table.

Authors	Target	Features	Dataset	XAI	Model	Resutls
Katz et al., 2016 [11]	Justice vote, Case outcome	justice (id), term, natural court, month of argument, petitioner, respondent, manner in which court took jurisdiction, administrative action, court of origin and source of the case, lower court disagreement, reason for granting cert, lower court disposition, lower court direction, issue, issue area	Supreme Court Database	No	RF	RF results / Justice Vote / Case Outcome; Accuracy / 71.9% / 70.2%
Alghazzawi et al., 2022 [12]	Case outcome	Natural Court, Petitioner, Respondent name, Evidence Type, Origin of Case, no public witnesses, Source of Case, Cert Reason, Origin Court dir, Lower Court Decision, Low Court disag, Cont eye witness dis, Precedence Alternation, no defense witnesses, Issue, Area of issue, Direction of issue, argument month, justice court difference, Type of Law, Prosecution motive, Majority Votes, Lower Court Direction res, Justice, Lower/Upper/court/difference, direction Court.	Supreme Court Database	No	LSTM + CNN	LSTM + CNN / Case Outcome; Accuracy / 92.1%
Medveva et al., 2020 [17]	Violation/No Violation of fundamental rights	NLP extracted n-grams from Text information	European Court of Human Rights	No	SVN	SVN / Violation/NonViolation; Accuracy / 75,0%
Hsun-Ping Hsieh et al., 2022 [18]	mediation agreement between parties	Case Information Features, Text information	Tainan, Taiwan mediation committee cases	Future development	LSTME nsembler	LSTMEnsembler / Mediation success or not; Accuracy / 78,9%
Turan et al., 2023 [22]	Violation/No Violation of fundamental rights and freedoms	term frequency–inverse document frequency from Text information	Constitutional Court of the Republic of Turkey	Yes	XGBoost	XGBoost / Case outcome; Accuracy / 93.84%

already convicted or under trial. What emerged is that many tools of this kind were built with data from male subjects, thus resulting in a lower accuracy as to predicting recidivism in women. Another example of this aspect can be COMPAS [27]; it is a commercial solution in use in the U.S. to predict any potential for recidivism among criminal defendants. Indeed, such software has become infamous, because of the use of biased data selection, which produced trained AI models with a tendency to be unfavorable with people belonging to certain ethnic groups. DataJust was a solution with the goal of developing a dataset and a system to propose compensation related to both any suffered harm and the party concerned, harm being considered more or less severe, depending on the characteristics of the person who suffered it. This solution was de-commissioned [28] due to some criticisms about the usage of data that were not totally anonymous. In fact, handling sensitive data is a fundamental point that should not be overlooked during the implementation of any kind of software. This becomes especially true, when dealing with data coming from certain areas, such as health care, financial sector or justice, which naturally referred to individuals.

In some cases, the information needed to take a correct decision is in the facts describing the context, more clearly than in the data body. When this occurs, anonymizing the related dataset allows to preserve any context description and it is a safeguard for user privacy [29]. A correct anonymization of personal data would lead to data where any subject (persoMn) cannot be any longer identified directly or indirectly, neither by the data controller alone, nor in collaboration with any other party [4]. Fully anonymized data can be stored and used without constraint. In addition, data cleaning should remove the information that could introduce bias, such as a person's social status, gender, nationality, ethnic group, etc.

A variety of tools has been proposed in both academic and commercial fields to identify and remove private information within documents [30]. Newer systems make the anonymization procedure be same as NLP task of Named Entity Recognition (NER) [31], [32], namely, to identify within texts particular entities of interest such as names, phone numbers, dates, addresses, and then proceed to replace them with anonymous labels.

Our research aims at contributing to the field of legal decision support systems in the context of mediation process providing a XAI solution based on a large language model. The aim of this research is to enhance both efficiency and effectiveness of the mediation process within the Italian justice system with a transparent and interpretable AI, and via an accessible online tool.

III. REQUIREMENTS ANALYSIS

The main objective of this work is to provide a tool to assist judges in coping with a lawsuit, in order to speed up their conclusions in assessing which disputes may result in a successful mediation. On the basis of an analysis done within Agile Justice project's objectives and critical challenges and according to the workshops performed with justice operators, identifying the requirements needed for a mediation decision support system has become possible.

Before providing a description of requirements, a further analysis and description of the whole context is needed, also to better identify the target and the concept itself of mediation prediction.

A. CONTEXT AND PROCEDURES

According to the law, some topics may imply the obligation to conduct a mediation procedure, where the parties concerned, aided by a mediator, seek to find a possible common ground that will end the dispute. However, there are cases, where it is the judge who may consider the possibility of pushing the parties to such a procedure. This may prove to be an effective tool for the conclusion of the case, if the parties are willing to seek a compromise. On the other hand, the mediation trial may end up adding more time, if the parties do not fully agree, thus contributing to a time extension in the litigation process. With these assumptions and in order for the judge's assessment to be as accurate as possible, it may be worth considering the introduction of a solution/tool capable to estimate the mediation probability of success. This would be a supporting tool for decision making, in a way similar as what the judge could benefit from a mediation expert who might also give some rationales, aside from any suggestion/assessment.

The assessment on the mediation probability of success (here called *mediability*) can be analyzed from different perspectives, each of them focusing on a particular aspect of the opportunity for mediation. In particular, the following three main cases are known:

A. **Probability of success as to the mediation attempt.** The success of a negotiation procedure is affected by many factors, such as: dispute context; personalities and interpersonal relations of litigants, as well as their personal interests; the given dispute severity; the extent of resources available to support any mediation process; and the negotiation skills the assigned mediator has.
B. **Propensity of the judge to submit the case to the mediation procedure.** Judges' inclination to pursue mediation is influenced by some factors such as their personal experience with cases showing a similar pattern, or familiarity with the appointed mediator. Moreover, a judge's propensity is not something easy to be calculated and it does not play a relevant role in those cases dealing with matters subject by law to compulsory mediation (e.g., disputes of condominium nature, or concerning leases, gratuitous loans, or business leases). The judge, in fact does not have full discretion in deciding the proper route for such cases.
C. **Propensity of litigants to engage in mediation.** It analyzes the willingness of the involved parties to actively participate in a mediation attempt, so as to reconcile mutually their different interests. Indeed, an eager spirit to cooperate and seek a shared solution is a determining factor as to any successful mediation outcome.

Without any doubt, **Case A**, appears to be the most effective in determining whether it is worthwhile to choose the mediation path. However, as already described, this option is affected by factors that are difficult to quantify and cannot be deeply analyzed within court records. One fundamental requirement the system under consideration has to meet is its non-reliability on subjective or hard-to-interpret data; it should rely exclusively on reliable documentation like textual court records. In most cases, it is very difficult to understand if a dispute has concluded the mediation or it is still running. A successful mediation is not typically reported as a result in the court. Therefore, solution A is not viable.

Case B predicts judge's propensity in choosing the mediation path and it may not provide a useful tool for decision making, since we would go beyond the individual capabilities of judges in selecting cases to be pushed.

Case C deals with calculating the parties' propensity for active participation in any mediation attempt and this would allow us to quantify any desire by the litigants to engage in any possible dispute resolution. Indicators of the parties' inclination could be identified within court texts, which is the exact opposite of the other two measures, where additional information would be necessary. In this paper, the focus has been on providing a solution for **Case C**.

B. IDENTIFIED REQUIREMENTS

Since the tool under consideration must provide an evaluation from textual court documents, an important requirement to be met is being able to provide unbiased suggestions. On the other hand, the processed data include a great number of sensitive information about the involved subjects, such as names, surnames, social security numbers, addresses and dates. It is important to ensure that these pieces of information do not influence the evaluation carried out by the tool; therefore, it will be of great help the implementation of measures to remove such pieces of information from texts, as the latter are used to develop the system. A similar procedure could also be applied to input documents, as they are input to the system for their evaluation. This would be necessary if these documents were uploaded, for example, on a public cloud for storage purposes. De-identification, i.e., the removal of sensitive information from texts would allow their preservation in compliance with privacy requirements.

In summary, we have identified the following system requirements. The solution should:

R1. produce a comprehensive assessment of each dispute regarding the parties' propensity to actively participate in the mediation process. Please note that each dispute is described by a set of documents. Some of them may be very significant for the assessment, whereas others could be just bureaucratic statements, such as: receipt of documents, records of some legal steps the involved parties have gone through, etc.
R2. produce the assessment by considering only court textual records, with no need to rely on other sources. Optionally other documents or single documents could be assessed independently.
R3. classify each dispute document in one of the following classes: propensity to mediate (M), not propensity to mediate (NM), and neutral (N). The neutral class would serve if there were inadequate information to

determine the propensity of the parties or may be just non-significant for the assessment.

R4. provide a score confidence about the produced classification of **R3**: M, NM and N.

R5. provide an explanation for the provided suggestion /classification of **R3, R4,** at level of sentence.

R6. ensure that the system is developed without the introduction of bias, for instance by removing sensitive information from documents used as a knowledge basesupporting AI-Ethics [33], [34], and Data-Ethics [35], [36].

R7. ensure that all the documents, being input to the tool, are properly stored, while respecting the privacy of the concerned subjects.

R8. identify meaningful sentences for M and NM classes of interest and make them accessible.

R9. provide a simple interface so that users without advanced computer expertise can get results/suggestions, to be used by a judge or a team to take any final decision and remain accessible on paper.

R10. provide an interface to qualified users, so that they can provide their corrections /suggestions, to be used by the solution to improve the model in later versions.

In reality, the solution should provide its assessment on the basis of a set of documents for each case. It is quite common that, whether taken individually, documents may express conflicting claims about the parties' propensity to take part in a mediation process and a large part of such documents may be neutral. In the same dispute consisting of X documents, some of them might be individually assessed as falling into class M, others as NM, while most of them as N. Typically, the judge would shift through each document looking for significant phrases revealing different aspects of the inclination or lack of it. An **additional requirement** of the proposed system must therefore be to provide an overall classification of litigations into M, NM, and N classes, to filter out sentences, within texts, highlighting the ones that are significant for the two classes of interest: M and NM. This additional piece of information can be relevant in deciding, while providing an explanation, in terms of the relevant text fragments which could support the reason why the framework automatically produced such classification. The same approach can be performed at level of single document: it may contain several statements which are neutral and yet some of them may be oriented as NM/M; only the latter ones are significant to a decision making process.

IV. DATA ASSESSMENT AND PROCESSING

According to typical approaches for machine learning, training and test sets must be produced. Moreover, the ML model produced by training would be profitably usable in execution to provide suggestions, if the data shapes provided in training are very similar to the ones which can be fed in execution by judges. Therefore, this section is dedicated to describing the steps undertaken to build the dataset used for training and testing. Some of the subphases are also used in the final solution, when it comes to producing suggestions executing the model on new documents, **R9, R10.**

The overall pipeline process is displayed in **Figure 2**, and in the next subsections we have provided detailed descriptions of such steps. Specifically, **SubSection IV.A** quantifies and describes data in their raw format. **SubSection IV.B** gives an in-depth description of the de-identification procedure which data need to be subjected to. **SubSection IV.C** is dedicated to the labeling procedure, in order to obtain annotated texts. **SubSection IV.D** describes the normalization to clean up and uniform the text data. **SubSection IV.E** reports the process adopted for sentence splitting and chunking to make sentences addressable by the BERT learning phase. **SubSection IV.F** describes the partitioning of the data set into the training, validation and test sets.

A. FILE GATHERING

The data we discuss in this section have been made available through an agreement between the University of Florence and the civil court of Florence. This has granted the authorization to access and process the content of such civil case files. Dossiers are made of a set of Italian-language documents of varying length and quantity. They consist of transcripts of sentences, hearing transcripts, and documents drafted by attorneys that entered into the trial record. More rarely a dossier can also contain a mediation transcript. The low number of mediation transcripts, however, should not be taken as an indication of whether mediation can be successful or not: parties often only verbally communicate one another that they have reached an agreement, without providing the documents produced by that mediation. In addition, clearly each trial may follow a more or less lengthy process depending on the issue complexity, or the interests of the involved parties. The data made available to us come directly from the court information system named SICID (District Civil Litigation Information System), which stores court documents, in the form of PDF files [37]. As the digital transition from paper format has occurred only recently in Italy, many of the documents uploaded in the application are not digital-native, but scans of paper documents. We therefore decided to exclude all non-digital-native PDF files from our selection, as it would have overloaded the pre-processing phase with the risk of introducing inaccuracies into texts.

The selected raw dataset consisted of 74 dossiers with a total of 474 documents. As already explained, dossiers contain a varying number of documents depending on both number of hearings and progress of the trial. In our case the size of a single dossier ranges from 2 up to 13 files, with a median value of 6. Even documents have a variable length, with number of pages being in a range from 1 to 40. Documents are digital-native PDFs. This process corresponds to **step 1 of Figure2**.

B. DATA DE-IDENTIFICATION

Requirement **R6** recommended making sure that this system would develop according to data-ethics and AI-ethics. Such issues can arise, when the system based on input data, namely

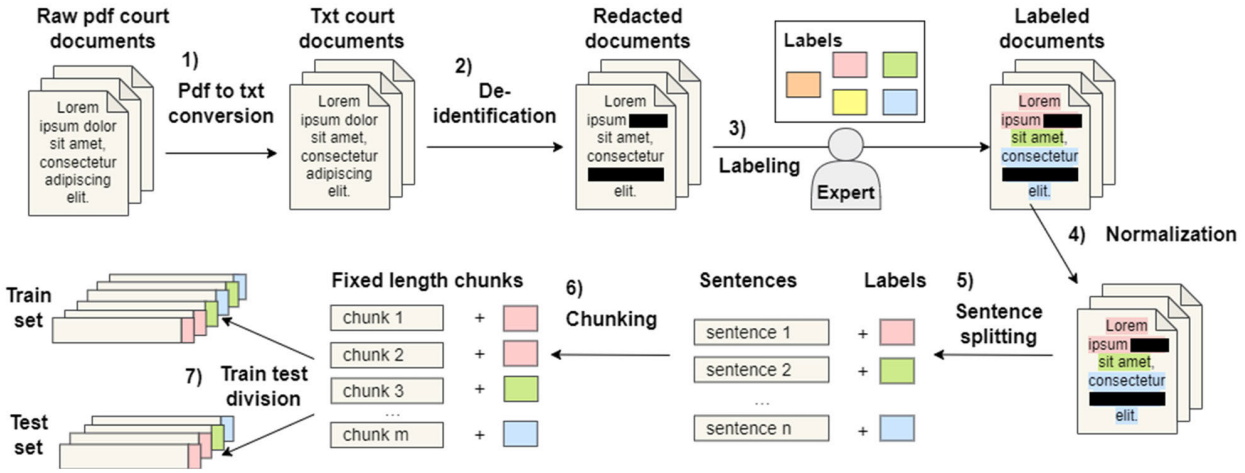


FIGURE 2. Dataset building procedure. Numbers refer to the steps which are described in the paper. The role of expert(s) has been to provide their assessment to both sentences and documents according to a classification model.

the ones contained within the dataset, picks up incorrect patterns. Information such as names, surnames, social security numbers, addresses, dates of birth or other events are particularly at stake and the risk is misinterpretation. *Annotation bias* is the name given to such an issue and it occurs precisely when the system is induced to create an association between labels and any irrelevant information contained in the examples used as ground truth, like any personal data. The presence of this piece of information within texts could be associated by the system, for example, with certain offenses or to a likelihood of propensity. Suppose a particular organization, such as a bank, is involved in processes where all parties always agree to resolve the controversy through a mediation procedure: leaving the name of the bank in plain text would lead the system to misclassify all sentences containing the name of that bank, without taking into account the rest of the sentence. Even if creating association metadata between a given entity and a specific mediation propensity would provide useful data for statistical analysis, the aim of this research is to create a text analysis model that is generic and unbiased enough to be used outside the Florence court.

The presence of such personal data within texts is not at all relevant for the purposes of this dataset, namely the calculation of mediation propensity. The presence of personal information could also be a problem especially under the umbrella of current legislation on data privacy. Works such as [38] aimed at proposing privacy models for document sanitization. Others proposed guidelines for processing personal data in legal documents to be GDPR compliant as in [39]. As to our work, these aspects are covered by the anonymization process specific to the SICID. Furthermore, the estimation of mediation propensity must be based solely on the events and facts reported in the documents, and not on the characteristics of the parties at the trial. To avert the formation of such assumptions, we have described in this section the procedure undertaken to remove such pieces of information from documents, referring to the process as *de-identification*. This involves the removal of all sensitive information that could lead to the identification of any person or organization involved in the case. Obviously, this process must ensure that texts remain understandable and their meaning intact, which is to say semantically meaningful. The removal of identification data is an important part of the preparatory process to: (i) avoid the formation of potential bias in the system knowledge base, (ii) improve the privacy level as to the involved individuals, whose sensitive data deserve special treatment and care.

This removal process corresponds to **step 2 of Figure 2**. This process should not be confused with a phase of anonymization, since we need to preserve the meaning of phrases. In order to clarify this aspect, we need to stress that there are several ways of identification removal:

1. **Redaction** (also known as purification or brutal approach) is the simplest approach, which involves replacing all personal data with a single label, such as OMISSIS [30]. This method ensures total de-identification; however, it loses entirely the information associated with the category to which such removed data belong.
2. Alternatively, using **entity-related labels**, such as #PERSON or #ORGANIZATION, allows to anonymize and keep the sentence context almost unchanged.
3. As a further extension, it might be possible to assign **numbered labels,** so as to maintain an anonymous distinction between instances of a given Entity, such as #PERSON1 for Mario Rossi and #PERSON2 for Luigi Verdi, and for all other identities involved in the related documents and for the whole document sets with the same labels.

As to building up the dataset for text analysis via machine learning, it is very important to perform a de-identification to avoid including personal data, while preserving meaningful aspects, so as to identify any elements relevant to mediation. Therefore, it is imperative to maintain the information which allows to distinguish among entities such as: litigant, judge,

and lawyer, thus preferring modalities of data removal able to preserve context. For this reason, **Case 1** of redaction is excluded. Adding distinguishing numbers as suffixes to labels greatly increases the complexity of an automated anonymization process, especially considering that a file is composed of multiple documents where the identifier's consistency must be maintained. In addition, each number represents a variation to a label that would otherwise be identical for all examples/documents, leading to a consequent increase in features to be learned by the model. Such features do not represent a really relevant difference for text analysis purposes, being those numbers exclusively useful for a better understanding of the context from a human being perspective. For these reasons, the strategy illustrated in **Case 3** is also ruled out. These observations led us to consider the use of context-related labels without numbers (**Case 2**) the ideal choice for this research.

According to **Case 2**, a supervised replacement tool driven by SICID defined metadata has been developed (this is the typical way legal documents are anonymized in Italy) directly into the SICID tool provided for each court in Italy. For each document a set of labels is mapped to placeholder. For example, the entity ''e-mail'' can relate to several subjects quoted in texts, and thus we have prepared a set of ''e-mail'' labels, one for each possible subject: plaintiff, defendant, third party, judge, plaintiff's lawyer, defendant's lawyer, third party's lawyer, witness, ctp/ctu (technical consultants), public notary, administrator, bank, generic company, etc. This is applied to all types of entities, such as first and last names, date of birth, place of birth, social security number, and residential addresses; and to entities not strictly related to a subject, such as other places, dates, or codes (SSN, VAT, Fiscal Codes, Chambre of Commerce codes, etc.).

C. DATASET ELEMENT LABELING

This section describes the annotation procedure used to create the ground truth relevant to model training and test (**step 3 of Figure 2**), only in preparation of the learning and test sets. The choice to annotate documents at this early stage of pre-processing is designed to avoid corrupting annotations, in the event we would decide to revise text modification steps. As to the labeling task, we made use of a tool named Doccano [40], which, thanks to its graphic user interface, allows quick annotation of text substrings in documents. We refer to these substrings as *sentences*, usually portions of text ranging from period to period. However, with this term we also refer to other cases, if deemed appropriate by the expert annotator, e.g., a substring that, though belonging to a larger period, requires a different label than the remaining period, or even blocks of text composed of several contiguous periods and belonging to the same class, and which therefore for convenience are labeled together as a single sentence. The adopted labels to manually annotate sentences are as follows:

- **Propensity to mediate:** sentences out of which the willingness of parties for active participation in the mediation procedure is clearly evident.
- **Not propensity to mediate**: sentences where the unwillingness of at least one party to be involved in the mediation process clearly emerges.
- **Technician involved**: sentences where technical experts, such as court-appointed technical consultants (CTUs) and partisan technical consultants (CTPs), are mentioned. This implies that one of the parties or the judge have asked for their presence.
- **Mentioned mediation:** sentences where the mediation procedure is referred, while excluding the ones where either a negative or positive orientation has emerged.
- **Neutral sentences** all those not classified as above.

The output of the labeling process is a series of files in JSONL format, one for each dossier. Each file contains as many JSON as the number of documents annotated in that case file. Each JSON has the following keys: id, text, label. Where text is the text of the entire document and label is a list of annotations. Each annotation includes a starting and ending index with respect to the document text, and a label.

Please note that, according to the distribution of labels assigned to sentences, we have registered a large number of documents which are Neutral, and thus not significant for the production of any suggestion. The portions of text relevant to the analysis of mediation propensity are small in size compared to the overall size of the documents.

D. TEXT NORMALIZATION

This Section describes **step 4 of Figure 2** regarding normalization, which is necessary because the technical jargon used in the legal field is rich of abbreviations and legal references that have to be normalized to avoid any inconsistent behavior during the later tokenization process, and also to avoid biasing according to the adopted jargon forms. Moreover, abbreviations also provide points which may be misinterpreted as full stops. They should not be considered as sentence breaks. There are also different abbreviations (by style or by typos) referring to a single extended version: for example, the abbreviations s.r.l s.r.l. srl. All types of S.r.l abbreviation are referring to the extended version of ''SRL Società a Responsabilità Limitata'' (i.e., Limited Liability Company). To normalize abbreviations, we exploited a mapping table and replaced all contracted versions with their respective extended alternative.

The data analysis has also revealed that there was plenty of specific information that was not relevant to the analysis of interest, such as dates, times, and other numerical data. Such data could introduce bias into the network, since they were also contained in some relevant sentences. Even when appearing in neutral sentences, their specificity results in reduced similarity among sentences. We therefore decided to use regex to replace such occurrences with placeholders as DATE, TIME, CODE. Regarding dates and times, we also have noted the absence of standard in formatting, with periods, colons and spaces used interchangeably as separators. In addition, if the numbers of hours or days

were single digits, the standard was not met by adding a prefixed 0, resulting in a change in patterns and requiring an increase in the complexity of the regex function used for substitution. As to codes and amounts, we have substituted any numeric sequence, optionally interspersed with commas, spaces, and periods, if not already identified by any date and time patterns.

The normalization phase has to retain symbols necessary for any logical separation into periods, such as commas and colons, while removing others if considered not essential for text understanding, such as quotation marks and parentheses. To normalize this condition, we have decided to separate these texts into sentences, by separating the phrases at each dot and semicolon. Therefore, every punctuation mark separated from words and non-needed whitespaces has been removed. Reducing and merging several words with a single alias or variant, as well as separating words from punctuation marks has the additional benefit of reducing the number of tokens into which the text will be converted. The smaller is the size of the token space, the easier the learning of BERT becomes.

E. SENTENCE SPLITTING AND CHUNKING

According to the normalization process described in previous section, the phase of sentence separation as **step 5 of Figure 2** is performed by splitting text blocks at each dot and semicolon. The result of this preprocessing step is a list of examples consisting in pairs of the type <sentence, training label>. Since the BERT architecture operates with input tokens, it is necessary to tokenize the text. Therefore, the process of text chunking (step 6 of **Figure 2**) to transform sentences into chunks of tokenized text is presented below. See **Section V** for BERT architecture networks.

By the term chunking we refer to that preprocessing **step 6 of Figure 2**, during which human readable text is transformed into blocks composed of tokens for the training of the machine learning model. The block size, or chunks must be of at most 512 tokens, which is a limit imposed by the BERT architecture, as better described in **Section V**. This process tokenizes the sentences independently, so that each chunk contains only tokens derived from a single sentence. To define the maximum number of tokens in a chunk, an analysis has been carried out as to the distribution of the number of tokens for sentences of the whole dataset. Therefore, according to the distribution the limit of 128 tokens has allowed to find a compromise from actual chunk size and the occupancy of the block. The calculated statistics reveal a distribution that follows a negative exponential trend. Most sentences contain fewer than 100 tokens, and the frequency significantly decreases as the token count exceeds 200. Following these results, we have determined $N = 128$ as a favorable compromise between padding minimizing and efficiently utilizing computational resources during this training process. Given N as the maximum number of tokens, chunks are obtained as follows. Let K be the number of tokens in a sentence, we define M as the number of chunks derived from a sentence as

$$M = \lceil K/N \rceil.$$

When $M > 1$, we decided to split the text only at indexes containing spaces, so as to not split a word into different chunks. In order to do this, we used the text length of the sentence L to compute the estimated length of the chunk E

$$E = \lceil L/M \rceil.$$

The length E is used as a reference point for a backward search of whitespaces. A whitespace is a separation between two words, so it can be used to safely split a sentence into substrings without splitting a word into two parts. If operating the cut on the identified space results in a substring with a number of tokens greater than N, the search is repeated for subsequent spaces. If the sentence has no spaces (a borderline case that has never occurred), the algorithm cuts the sentences according to the size closest to the estimated substring length E, thus, the number of tokens generated is less than the established number of max tokens. The final result is a list of pairs <text_chunk>, <label> where <text_chunk> is a text that, when tokenized with the HuggingFace BertTokenizerFast, with padding enabled, yields a list of exactly N=128 tokens. <label> is the same label of the sentence where the chunk is derived from, and it is one of the 5 labelling classes described in Section IV-C.

F. TRAINING-VALIDATION-TEST SPLIT

Instead of the usual 60-20-20 rule for the training-validation-set division, we have divided our dataset according to the 80-10-10 ratio. The high variability of textual examples, especially for neutral class, associated with the low quantity of examples for relevant classes, led us to decide to reduce the size of validation and test sets, so as to favor a larger training set; moreover, we have planned since the very beginning of our research a second level of model validation, to be carried out by experts in the legal field. As described in **Section VIII**, a second validation has been performed by using new data when legal experts exploited the solution and also provided their comments and assessment using our tools as described in the following.

As described in **Section IV-C,** the dataset has been labeled by experts at level of cases and at the level of a single sentence. As reported in **Table 2**, most sentences have been labeled as supposed neutral since the very beginning. Considering a generic case file, documents such as any parties' subpoenas do not - by their nature - contain any useful information about the propensity for mediation. Even when a document includes some relevant information, it is always limited in a relatively small number of sentences within the entire document length.

Since the number of neutral examples represents more than 90% of the entire dataset, if this whole dataset for training was used, it would result in an unbalanced model, providing satisfactory results only for neutral class. On the contrary, our goal is to produce a model capable to correctly

TABLE 2. Dataset composition in terms of number of sentences: training, validation and test sets, with respect to their classification.

	Whole dataset	rebalanced	Training set	Validation set	Test set
Neutral	14115 (90.58%)	1800 (55.10%)	1416 (54.65%)	192 (56.80%)	192 (56.80%)
Technician involved	702 (4.51%)	702 (21.49%)	562 (21.69%)	70 (20.71%)	70 (20.71%)
Mentioned mediation	582 (3.74%)	582 (17.81%)	466 (17.99%)	58 (17.16%)	58 (17.16%)
Propensity to mediate	123 (0.78%)	123 (3.76%)	99 (3.82%)	12 (3.55%)	12 (3.55%)
Not propensity to mediate	60 (0.39%)	60 (1.84%)	48 (1.85%)	6 (1.78%)	6 (1.78%)
Total	15582	3267	2591	338	338

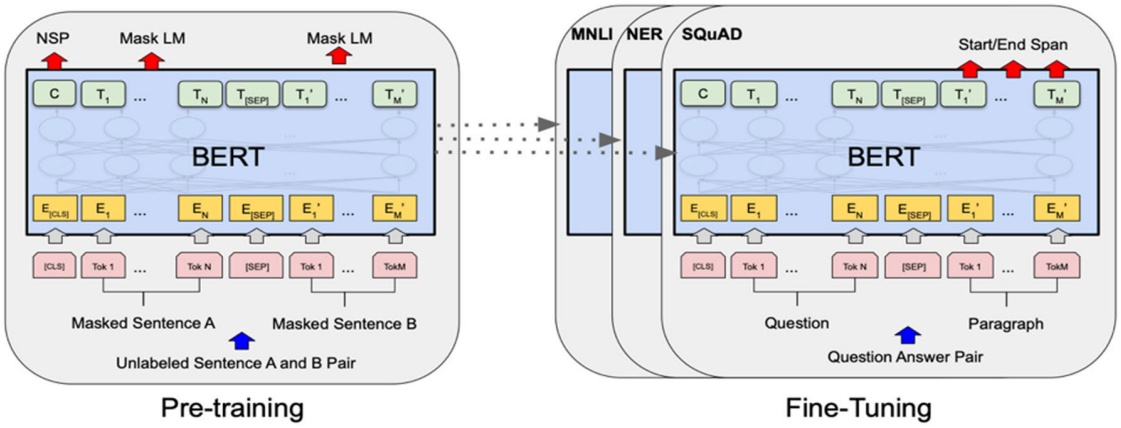


FIGURE 3. Overall BERT pre-training and fine-tuning typical procedures.

classify propensity and non-propensity statements among many other classes. Therefore, a rebalanced training set has been produced, keeping 1800 neutral sentences out of the 14115 totally available, while all other classes have remained unchanged. The number of neutral examples to retain, K, was determined as follows. Starting from the balancing principle that the number of examples belonging to the most populous class K should be equal to 10 times the number of examples in the least frequent class, we set K = 600 as the limit for the number of examples. However, it is essential to consider that not only is the high number of neutral sentences representative of the actual composition of the documents, but it also contains a high variability typical of any natural language. Moreover, the final model in execution would find a large number of neutral examples, so that a compromise is needed. Therefore, we identified a reduction value to find a balance between dataset rebalancing and preserving the variability of neutral examples. As a result, the final training set has been composed of 1800 examples from the neutral class, equivalent to 12.75% of the total number of neutrals originally available, and 30 times the number of examples in the smallest class. The resulting rebalanced dataset from this reduction is shown in **Table 2**, and it has been split into training-validation-test sets according to 80-10-10 percentages (thus resulting in **Step 6 of Figure 2**).

V. MODEL DEVELOPMENT AND VALIDATION

In this section, the training of ML model to classify sentences according to the above mentioned 5 classes is discussed. Sentence classification has allowed a granularity of classification similar to the one which would be obtained from natural language logic analysis (**Subsection V-B**). If considering that a file is composed of several documents with very different content, it has been useful to implement a mechanism, so as to map results in a document-level classification (**Section VI**). Sentence classification has been also assessed by using Shap explainability AI techniques, XAI (see **Section VI**) [6].

A. MODEL DEVELOPMENT

Traditional NLP models were often trained using task-specific labeled datasets trained from scratch on the specific task, thus requiring a relevant amount of labeled data for each task individually, and large resources. Transformers [41] have an encoder-decoder architecture leveraged by a self-attention mechanism to capture dependencies among different parts of the input sequence. This result is obtained by weighting with a relevance of the different input elements, while making predictions. By giving attention to relevant parts of the input, transformers can effectively model long-range dependencies and capture contextual information more effectively than previously done, thus making such models particularly useful for text-classification. Introduced by Devlin et al., in 2018 [5], the BERT approach leverages bidirectional context to capture a more comprehensive understanding of text, unlike previous models that predominantly relied on unidirectional language modeling. BERT's innovation lies in its pre-training and fine-tuning approach. In **Figure 3,** a diagram of BERT architecture is reported. During the pre-training phase, the model has been trained on large-scale corpora using a masked language modeling (MLM) objective 41]. In this process, a certain percentage of input tokens is randomly masked and

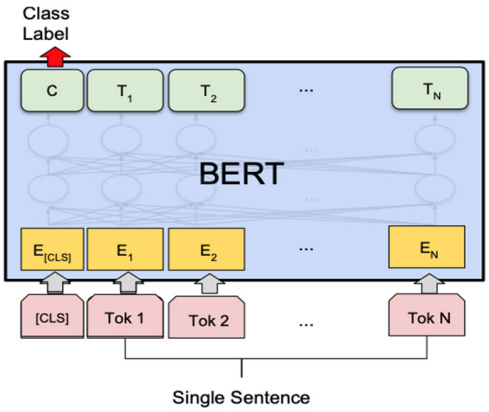


FIGURE 4. Fine-tuning BERT on single sentence classification.

the model is tasked with predicting these masked tokens on the basis of the surrounding context. By training vast amounts of text data, BERT learns rich representations that capture deep semantic and syntactic features. Essentially, this allows the model to learn high quality representations/patterns of natural language. These representations allow to model a language in a general way, learning the semantic relationships between words and structures that are specific to a given language. The pre-training phase is followed by a fine-tuning on the specific task, such as sentiment analysis or named entity recognition, just to mention a few of them.

For our purpose, a fine-tuning approach has been adopted, having its focus on creating a model for text-classification (see **Figure 4**). To this end, as pre-trained model we have used the Italian BERT XXL *Cased* model [42], which is has been trained on Italian texts from Wikipedia [43], OPUS [44] and OSCAR project [45] data collections. The final training corpus was of 81GByte including more than 13 billion of tokens. The *Cased* version has been used, rather than the *Uncased*, since the former one does not remove capital letters and accents. In Italian language, capitalized words are used only at the beginning of sentences and for proper nouns, and because of the uniqueness of legal context, all nouns have been replaced with anonymous labels (please note that in some Italian surnames are also generic substantives or adjectives, for example: Rossi, Bellini; here capital letter may be of help). Such labels are simply very common capitalized words to which is prepended a hash symbol, e.g., #JUDGE and #LAWYER. Knowing how to recognize accented words can also be critical to understand a sentence; just consider, for example, that only an accent distinguishes the word 'e' i.e., *and* conjunction from 'è', i.e., *is*, third person singular of the verb to be.

Hyperparameters of the pre-trained Italian BERT XXL Cased model are: 12 attention heads, 768 as hidden size dimension, 12 hidden layers, 512 as maximum embedding dimension and a vocabular size of 31102. BERT uses a WordPiece tokenizer to prepare textual input. The tokenizer splits statements, so as to have one word per token or into word pieces where one word is broken into multiple

TABLE 3. Range of hyperparameters for BERT Model fine tuning.

Hyperparameter	Search domain
Early stop	15
Learning rate	from 1e-6 to 1e-3
Weight decay	from 0.005 to 0.01
Batch size	[2,4,8,16]

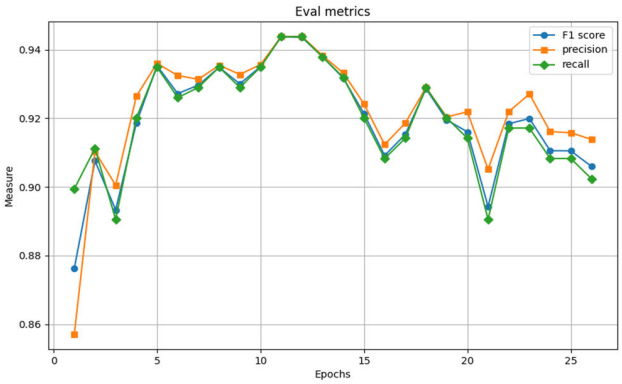


FIGURE 5. *Trends of F1 score, precision and recall as a function of the number of the epochs in the training / validation phase.*

tokens. Another limitation of the adopted pretrained model has forced us to a fixed size for inputs, thus the input had to be regularized by introducing padding and/or truncation sentences. Therefore, the mentioned pre-trained model has been fine-tuned for the purpose of creating a classifier and the number of tokens has been set to 128, due to the motivations reported in **Section IV-E**. A hyper-parameterization to maximize the F1-score on the basis of the validation has been performed. The range of the hyperparameters is reported in **Table 3**. The AdamW optimizer has been selected since it has provided the best results. Best results have been obtained with a Learning Rate of 3.15E-05, Weight decay of 7.85E-03, and batch size of 16; obtaining an F1 score of 0.944.

In **Figure 5**, trends of F1-score, precision and recall in classification are reported as a function of the epoch. According to the graph, the best F1 score turned out to be at 11th epoch. Such a result has been confirmed by the trend of the loss to avoid overfitting.

B. MODEL VALIDATION

This could be possible only because based on results yielded by a sentence-level classification model. Specifically, for each sentence in the document, we have used probabilistic scores in each class produced by this model classification. As a first step, for each sentence, scores of the non-characterizing classes are summed up together: in other words, the scores of the classes "neutro", "tecnico" e "mediazione menzionata" are all summed into "neutro_sum". This grouping is justified by the fact that the content of sentences classified as "tecnico" and "mediazione menzionata," though being more relevant than a generic neutral, is not distinctive enough to be used during this automatic analysis.

TABLE 4. Class-level evaluation of results on test-set.

classes	Precision	Recall	F1-score
Propensity to Mediate	1.000	0.923	0.956
Not propensity to Mediate	0.500	1.000	0.667
Neutral	0.959	0.989	0.974
Technician involved	1.000	0.915	0.956
Mentioned mediation	0.944	0.864	0.903
Global	0.958	0.950	0.952

The fine-tuned model presented in **Section V-A** has been assessed on test set (see **Table 2**). On a test set including 338 sentences, 321 were classified correctly. The metrics of precision, recall and F1-score calculated at individual class level are reported in **Table 4**, as shown in equations (1), (2), and (3). Considering as class values ''Negative'', ''Neutral'', ''Positive'', recalling the definitions of True Positive (TP) as an outcome where the model correctly identifies the class, False Positive (FP) as an outcome where the model incorrectly identifies the considered class, and False Negative (FN) as an outcome where the model incorrectly identifies the not considered class, the following metrics have been computed per each class:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1score = 2x\frac{Precision\ x\ Recall}{Precision + Recall} \tag{3}$$

A global assessment has been estimated by using the weighted-average score calculated by taking the mean of all per-class scores and considering the support of each class (see last row of **Table 4**). The support of a class is defined as the number of true occurrences for that class. The 'weight' is the proportion of each support of each class relatively to the sum of all supports. With weighted averaging, the output considers the distribution of cases for each class weighted by the number of instances of a given class. This is particularly useful in the cases of unbalanced data sets among categories, as it occurred in our case. Moreover, accuracy has been estimated as the ratio from total number of correctly classified sentences with respect to number of sentences, and in this case, the accuracy turned out to be equal to 94.9% outperforming the 'related works' results as reported in Table 1.

From those results, it can be observed that the model has produced very good F1-score results for Neutral, Technician Involved and Propensity to Mediate. Given the system's objective of assisting the judge in identifying elements of propensity and non-propensity, we believe that these two latter classes are the most significant ones where any evaluation attention and effort should be focused on. It is extremely important that the model facilitates and speeds up any identification of those few relevant sentences (for those classifications) which are typically contained in long, predominantly irrelevant documents. For this reason, it is important that the model provides a high recall for those two classes, as in fact happens with 0.923 for propensity to mediate and 1.0 for the non-propensity to mediate. However, we have observed that the non-propensity to mediate also experiences a precision of 0.5.

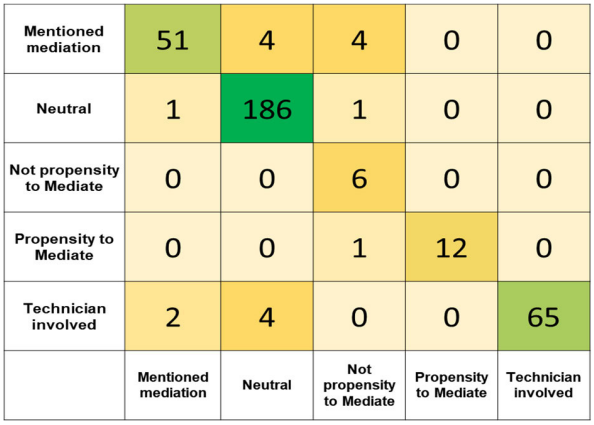


FIGURE 6. Confusion matrix computed over the test-set. Actual Values are those reported on X-axis. For example, propensity to mediate does not present any error.

Based on the predicted results, a **confusion matrix** has been created over the 5 considered classes and classical metrics have been calculated such as precision, recall, f-score aggregating all classes (see **Figure 6**). According to the confusion matrix, it is possible to observe that for the class not prone to mediation, 6 errors are found, of which 4 involved a misclassification towards the class of mentioned mediation. The following consideration can be made:

- the number of errors is relatively low if compared to the total number of analyzed sentences. This means that if a user would have liked to identify all the sentences expressing a feeling of disinclination, he/she could have identified all of them, simply by analyzing 12 sentences (the 6 correctly identified with recall 1.0 and the 6 incorrect ones), instead of analyzing 338 (the total number of sentences in the dataset);
- the detected errors mainly involved this mentioned mediation class, whose examples are actually similar to those of the non-propensity class.

VI. RESULTS EXPLAINABILITY, XAI

A tool for decision support has the duty to motivate the computed suggestions / assessments, thus explaining to users any result produced as identified in **R7** (should identify meaningful sentences for M and NM classes of interest, **R8**). To this end, we have adapted a XAI technique to identify and provide evidence about which features and how they influenced classification results.

Based on [46] the type of explainer for determining how model outputs relate to inputs belonging to the Scoop-based class. A recent survey [47] on XAI and Law reported that this type of approach prevents neural networks, that perform extremely well, from behaving opaquely. Indeed, in [48], a selected group of lawyers were charged with the task

of assessing different explainability methods. Results have shown similar outcomes, and all of them have pointed out the need to provide explainability, so as to assist their work. Lundberg et al., in [6], compared different methodologies to provide this type of explanation and found a much stronger agreement between human explanations and SHAP (SHapley Additive Explanations) [6]. Shap approach is based on game theory to calculate the importance of each feature in determining model output. According to the BERT approach, the adopted features are words in the text, and the technology chosen for the explainability has been Shap to offer a coherent local and global explanation. The global version aims to identify the most influential words in giving a classification according to the model, whereas the local version aims to obtain explanations of the statements / texts sent into inference and why each text is classified in a certain way according to the above defined classification classes.

A. GLOBAL EXPLAINABILITY

Global explanation aims to represent the overall impact of certain features on the final classification of the developed model, in terms of importance towards the class prediction determined by specific words in the considered sentences. For this analysis, the test set reported in **Section IV-F** has been considered. The XAI library SHAP assigns a Shap value representing a quantification of the contribution towards the classification of the sentence in one of the considered classes. This value can be positive (the word contributed positively to the classification of the determined class) or negative. As a result, for each class, the Shap values regarding words that contributed to the classification, have been computed, too. Results are reported in **Figure 7** where the top 10 most relevant words per class are reported in a word cloud. The word cloud is characterized by a representation of the words in terms of color and size. The bigger the word is, the more important is its associated Shap value. Color represents the associated class. Results reported in **Figure 7** are useful to understand the model functioning. For example: the key presence of CTP/CTU keywords in the classification of a statement referring to a technical consultant, which in most cases demands a further technical analysis of the dispute topic, before any decision taking.

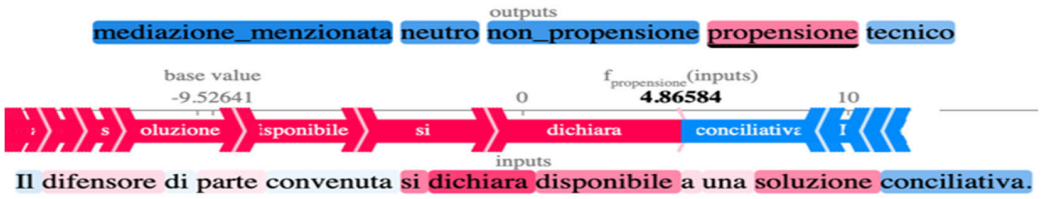
B. LOCAL EXPLAINABILITY

Aside from a global interpretation on the model produced with respect to the whole test set, a specific assessment can be performed when the model is adopted to classify single sentences. In this case, the most relevant words identified by higher Shap values are those related to what mostly can influence the suggested decision/classification, positively or negatively. This is the local Shap approach, each sentence is analyzed independently and each word in the sentence corresponds to a feature. This provides evidence of the compliance with **R5**.

From the test set, 2 key examples have been selected regarding sentences that are classified as "Propensity to mediate" and "Non propensity to mediate" to go deeper into the functioning of the developed model. The graphs generated and reported in **Figure 8** are useful for: (i) assessing model mechanisms, (ii) communicating to decision makers which are the keywords that mainly contributed to the computing of the produced suggestion, in each specific class (and more particularly, in the produced suggestion as the most probable class identified by the model/classifier).

FIGURE 7. Feature cloud, in Italian language since the text of data sets are Italian. (CTP/CTU are the acronym to identify a technical consultant, consulente → Consultant, assista → assisted, mediazione → mediation, contumace → contumacious, remissive → submissive, etc.)

In the represented visualization for each sentence above we can see all possible classes, among which the one suggested by the model is highlighted in red. Within each sentence, words in red are positively associated with the label chosen by the model, and in blue the ones contributing to the classification in a different direction. We also observe the intensity with which features are highlighted: higher intensity corresponds to greater relevance, that is, greater weight (size of the arrow bar below) with which the word affects the model's prediction for the sentence under consideration. In the reported examples, sentences are in Italian, as our case study was the Florence Court where data are mainly in Italian language. For Case (i), the statement analyzed was "*Il difensore di parte convenuta si dichiara disponibile a una soluzione conciliativa*" which becomes in English as: "*The defendant's lawyer declares himself available for a conciliatory solution.*". The XAI highlighted as really important towards the classification of "Propensity to mediate" the following words [*declares himself available for a conciliatory*] indicating the propensity towards mediation. In Case (ii), the sentence "*parte convenuta non ha partecipato*" can be translated as "*respondent did not participate*" where the explanatory visualization justified his/her non propensity towards mediation, his/her lack of interest in the procedure. This XAI feature has two main advantages, one for the development part of the project to understand the functioning of the AI model and the second one for final users to better understand the decision support system. This XAI tool has been integrated into the developed decision support system prototype, whose details are reported in the next section.



(i) Propensity to mediate. which becomes in English as: "The defendant's lawyer declares himself available for a conciliatory solution.".

(ii) Non-propensity to mediate, which can be translated as "respondent did not participate".

FIGURE 8. Local Shap diagrams related to two examples of explainability. Top example (i) for "propensity to mediate" and bottom to (ii) "non propensity to mediate."

VII. DOCUMENT LEVEL ASSESSMENT

The description reported in the previous section focused on classification of single statements among those which are included in several documents related to a litigation case. Among the identified requirements, the ability to classify documents of litigation case (**R3**) can be very useful to guide the court/judges to the relevant documents. The classification at level of document should be according to classes: propensity to mediate (M), non-propensity to mediate (NM), and neutral (N). We aimed to provide also a score confidence about the produced classification (**R4**).

As above mentioned, the presence of statements classified as "*Propensity to mediate*" and "*Non propensity to mediate*" is sporadic in the document. Since the goal of the document classification is to identify those documents which may help to take the decision, the classes "*Neutral*", "*Technician involved*" and "*Mentioned mediation*" are considered as "*neutral_sum*". This grouping is justified by the fact that the sentences classified as "*Technician involved*" and "*Mentioned mediation*" are *defacto* non oriented statements and thus *Neutral*. Thus, if at least one sentence is classified as either propensity or non-propensity, we consider only those sentences to classify the document. We then used the scores of these sentences to calculate the weighted average in the three categories: NM, M, and N. Depending on results of scores for each class/grouping, we distinguished between the following cases:

- **no sentence in the document has been classified as propensity, nor as non-propensity.** Thus, all the sentences in the document are classified as "*Neutral*", "*Technician involved*" or "*Mentioned mediation*", and document classification can be estimated on the basis of the weighted average for these classes.
- **at least one sentence is classified as propensity or non-propensity**, sentences will be considered to classify the document oriented on *propensity to mediate* (M), *non-propensity to mediate* (NM), or "*neutral_sum*", and both document classification and confidence are estimated on the basis of the weighted average of the corresponding statements in one of the 3 classes.

For both cases the obtained result is a set of three weighted averages, which are transformed into percentages and represent the confidence of document classification.

VIII. XAI4MA TOOL FOR DECISION SUPPORT

In this section, the decision support system delivered to the Court of Florence and named XAI4MA (Explainable Artificial Intelligence tool for Mediation Agile) is presented. As described in Figure 1, the tool exploits: (i) BERT Model for classification at level of sentences and documents, (ii) results of the XAI in Shapely approach to provide support to court and judges, in order to make a decision as to sending a litigation to mediation or not. According to the rules imposed by the GDPR to enforce privacy protection over European citizens' data, we need to perform an extensive and detailed anonymization operation on the extremely sensitive data contained in court documents, before being allowed to process them. Indeed, neither the analysis service nor the learning process of the model need any personal data about litigants. XAI4MA accepts the anonymized documents coming from SICID which is a national court tool in Italy and is also currently used by the staff at the Court of Florence. This choice allowed us to fulfil **R6** and to guarantee fast document processing times.

The architecture of XAI4MA is reported in **Figure 9.** It allows to process anonymized data coming from SICID, as well as any other kinds of documents either court or judge would like to assess. Through XAI4MA, users can:

- Upload anonymized textual documents to be analyzed, typically those produced by the SICID tool. This simplifies the possibility of assessing separated documents, see **R2**.
- Upload also non anonymized documents which are anonymized on the fly according to the SICID standard

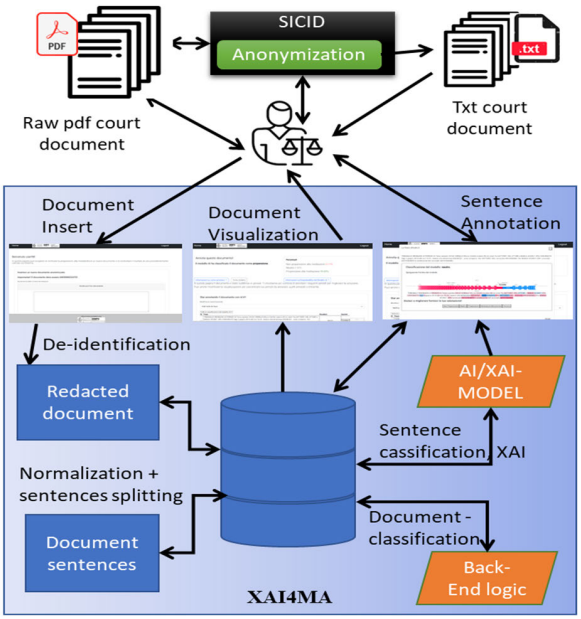


FIGURE 9. XAI4MA architecture, which is based on Snap4City Infrastructure [7], [49] for the data ingestion and management.

of the Italian govern, see **R6.** Both received documents and produced results are managed in the Snap4City platform which is GDPR compliant [49], **(R7).**

- Receive results with explainability hints on the sentence-level classifications.
- Visualize and print results of both sentence-level and document-level classifications, to be used during discussion in the court with judges (**R9, R10**).
- Provide classification at level document, to help judges to identify the most relevant documents, and in those documents the most relevant statements which can be used to take a decision about either sending or not litigation to mediation (**R9, R10**).
- Optionally, experts or judges can assess the provided results, and thus validate results from the model, confirming or proposing a modification to the classifications provided at level of sentences.
- Organize documents in dossiers, in a similar way as in the SICID system, see **R1**.

The XAI4MA architecture is a multi-user application where the front-end of the system can be used to receive the documents/text. In XAI4MA both documents and processed sentences are stored in a database. The processing by the AI/XAI model generates classification and explanation at sentence level and a processing business logic handles the final classification of the model, as reported in **Figure 10**.

In particular, the process is completed by using a monitoring interface to access the document with the classification results shown in the top right part of the interface, where percentages of the classes are reported, as well as explainability of the AI model with its related class in the box in the middle of the user interface (see **Figure 10**).

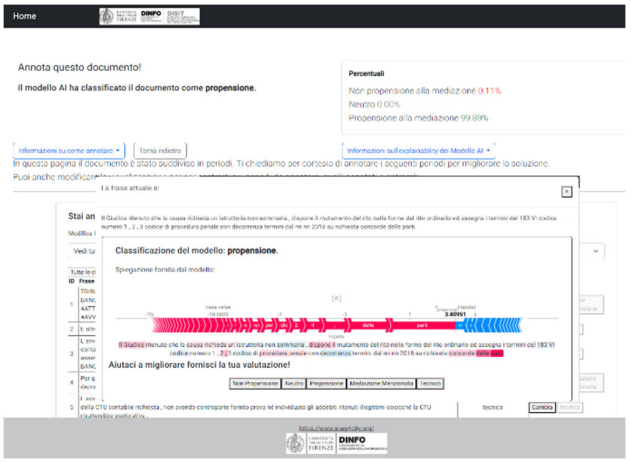


FIGURE 10. XAI4MA main document page, showing main results in terms of XAI explanation about the suggested assessment of the statement into the document. The XAI assessment is provided in the form of Figure 9.

The XAI4MA can collect provided documents to be processed, save proposed assessment and collect possible corrections for judges and experts. The XAI4MA tool has been successfully used by several experts to perform an additional validation. To this end, 25 new documents for a total of 6060 new sentences have been assessed (of which more than 5600 have been classified neutral by experts). In this additional validation, the global weighted Precision has been of 0.99, the weighted recall of 0.97, the F1-score of 0.98, and the Accuracy of 97%, going beyond the expectations described in **Table 4** and outperforming the related works results. On the other hand, despite its high Accuracy, the single class recall for Propensity to Mediate has been limited to 70%.

The XAI4MA system is hosted on a Linux virtual machine with 16 assigned virtual cores from an Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz, 16GB of RAM and 500GB of space. The AI model processing documents is in execution in background in a Dell Precision 5820 Tower equipped with a NVIDIA GeForce RTX 3090 with 24 GB of memory and 10496 cores.

IX. LIMITATIONS

One of the main motivations of our study is to address the problem of identifying the limited number of specific sentences in legal documents as part of a dossier and use them as indicative of propensity or non-propensity to mediate. Even when a document contains relevant information, it is often limited to a relatively small number of sentences, while the others are neutral, etc. This data limitation could impact the utility of AI approaches introducing possible bias towards the neutral class. This has been addressed by producing a rebalanced training set, maintaining a significant number of neutral sentences while keeping the other classes unchanged. The final assessment of accuracy has been estimated also on an unbalanced test set and it has obtained even higher accuracy.

It may be easy to think that the proposed solution is focused on solving the specific case of assessing propension to mediate. The goal of XAI4MA is to simplify the workload of judges who have to go through a lot of documents in a dossier to find the sentences in favor of mediation. The solution provides a confidence scoring at document level to enable the identification of target documents in a dossier which contain sentences classified by the model as indicative of propensity or non-propensity to mediate. On the other hand, most sets of documents around civil or penal disputes, administrative and insurance disputes, etc., share similar problems: *many documents where the significant statements are only a small percentage*. In all these cases, the judge team has to spend a lot of time in skipping neutral and marginal information to identify relevant aspects/statements. Thus, in all the above-mentioned cases, the XAI4MA proposed solution can be applied to reduce processing time and offer decision support.

X. CONCLUSION AND FUTURE DIRECTIONS

The protracted disposition time ranking amongst the highest in Europe remains a significant challenge within the Italian justice system. In civil trials, the adoption of a mediation process offers the potential for a more expedited resolution, enabling the involved parties to amicably conclude disputes outside the formalities of court procedures.

This decision is under the responsibility of judges/courts, entailing the exhaustive perusal of extensive documentation, often of hundreds of pages and various legal materials, and final decision-making on the basis of few sporadic statements. To address this challenge, this study introduces an artificial intelligence solution in the form of an innovative decision support system known as XAI4MA (Explainable Artificial Intelligence tool for Mediation Agile). This tool not only facilitates the assessment of mediation prospects with an accuracy of 97% at sentence level but, more significantly, through the utilization of XAI, it elucidates the specific clauses and segments within documents that could affect the decision-making process. The proposed system could help judges in the final decision process and provide them with the so called "*right to an explanation*" required by the GDPR, as well as take care of the data de-identification procedure towards generalization of applicability and privacy concerns.

Future directions of this current work are focused on: (i) extending the solution to work also on cases related to different kinds of dossier, for examples insurances; (ii) extending and generalizing the solution and extending the training set; (iii) using the tool to train mediators as well.

ACKNOWLEDGMENT

The authors would like to thank Giustizia Agile, national project and partners ([50]). Their heartfelt thanks to Prof. Paola Lucarelli who has guided and introduced them to the world of mediation, thanks also to many experts which have contributed to the development of the annotated data sets, and to the general assessment of the solution. Snap4City (https://www.snap4city.org) is an open technology and research by DISIT Laboratory, University of Florence, Italy.

REFERENCES

[1] (2020). *European Judicial Systems—CEPEJ Evaluation Report—2022 Evaluation Cycle*. Accessed: Mar. 2, 2024. [Online]. Available: https://www.coe.int/en/web/cepej/special-file-report-european-judicial-systems-cepej-evaluation-report-2022-evaluation-cycle-2020-data-

[2] *Direzione Generale Di Statistica E Analisi Organizzativa (DG-Stat)*. Accessed: Mar. 2, 2024. [Online]. Available: https://webstat.giustizia.it/SitePages/Home.aspx

[3] (2010). *DECRETO LEGISLATIVO 4 Marzo 2010*. Accessed: Mar. 2, 2024. [Online]. Available: https://www.gazzettaufficiale.it/eli/id/2010/03/05/010G0050/sg

[4] *Data Protection*. Accessed: Mar. 2, 2024. [Online]. Available: https://gdpr.eu/what-is-gdpr/

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[6] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Presented at the 30th NIPS*, Long Beach, CA, USA, 2017.

[7] C. Garau, P. Nesi, I. Paoli, M. Paolucci, and P. Zamperlin, "A big data platform for smart and sustainable cities: Environmental monitoring case studies in Europe," in *Proc. 20th ICCSA*, Cagliari, CA, Italy, Jul. 2020, pp. 393–406.

[8] LA LEY. *Jurimetria*. Accessed: Mar. 3, 2024. [Online]. Available: https://jurimetria.laleynext.es/content/QueEs.aspx

[9] Predictice. *Accédez à toute Toute L'information Juridique*. Accessed: Mar. 3, 2024. [Online]. Available: https://predictice.com/fr

[10] M. Medvedeva, M. Wieling, and M. Vols, "Rethinking the field of automatic prediction of court decisions," *Artif. Intell. Law*, vol. 31, no. 1, pp. 195–212, Mar. 2023, doi: 10.1007/s10506-021-09306-3.

[11] D. M. Katz, M. J. Bommarito, and J. Blackman, "A general approach for predicting the behavior of the supreme court of the United States," *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0174698, doi: 10.1371/journal.pone.0174698.

[12] D. Alghazzawi, O. Bamasag, A. Albeshri, I. Sana, H. Ullah, and M. Z. Asghar, "Efficient prediction of court judgments using an LSTM+CNN neural network model with an optimal feature set," *Mathematics*, vol. 10, no. 5, p. 683, Feb. 2022, doi: 10.3390/math10050683.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[14] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/j.patcog.2017.10.013.

[15] M. Medvedeva, M. Vols, and M. Wieling, "Using machine learning to predict decisions of the European court of human rights," *Artif. Intell. Law*, vol. 28, no. 2, pp. 237–266, Jun. 2020, doi: 10.1007/s10506-019-09255-y.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/bf00994018.

[17] N. Aletras, D. Tsarapatsanis, D. Preotiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European court of human rights: A natural language processing perspective," *PeerJ Comput. Sci.*, vol. 2, p. e93, Oct. 2016, doi: 10.7717/peerj-cs.93.

[18] H.-P. Hsieh, J. Jiang, T.-H. Yang, R. Hu, and C.-L. Wu, "Predicting the success of mediation requests using case properties and textual information for reducing the burden on the court," *Digit. Government, Res. Pract.*, vol. 2, no. 4, pp. 1–18, Oct. 2021, doi: 10.1145/3469233.

[19] J. Collenette, K. Atkinson, and T. Bench-Capon, "Explainable AI tools for legal reasoning about cases: A study on the European court of human rights," *Artif. Intell.*, vol. 317, Apr. 2023, Art. no. 103861, doi: 10.1016/j.artint.2023.103861.

[20] L. K. Branting, C. Pfeifer, B. Brown, L. Ferro, J. Aberdeen, B. Weiss, M. Pfaff, and B. Liao, "Scalable and explainable legal prediction," *Artif. Intell. Law*, vol. 29, no. 2, pp. 213–238, Jun. 2021, doi: 10.1007/s10506-020-09273-1.

[21] A. Deeks, "The judicial demand for explainable artificial intelligence," *Columbia Law Rev.*, vol. 117, no. 7, pp. 1829–1850, Nov. 2019.

[22] T. Turan, E. Kuçuksille, and N. K. Alagöz, "Prediction of Turkish constitutional court decisions with explainable artificial intelligence," *Bilge Int. J. Sci. Technol. Res.*, vol. 7, no. 2, pp. 128–141, Sep. 2023, doi: 10.30516/bilgesci.1317525.
[23] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation,'" *AI Mag.*, vol. 38, no. 3, pp. 50–57, Sep. 2017, doi: 10.1609/aimag.v38i3.2741.
[24] H. Westermann, J. Savelka, and K. Benyekhlef, "LLMediator: GPT-4 assisted online dispute resolution," 2023, *arXiv:2307.16732*.
[25] B. van Giffen, D. Herhausen, and T. Fahse, "Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods," *J. Bus. Res.*, vol. 144, pp. 93–106, May 2022, doi: 10.1016/j.jbusres.2022.01.076.
[26] K. A. Geraghty and J. Woodhams, "The predictive validity of risk assessment tools for female offenders: A systematic review," *Aggression Violent Behav.*, vol. 21, pp. 25–38, Mar. 2015, doi: 10.1016/j.avb.2015.01.002.
[27] T. Brennan, W. Dieterich, and B. Ehret, "Evaluating the predictive validity of the compas risk and needs assessment system," *Criminal Justice Behav.*, vol. 36, no. 1, pp. 21–40, Jan. 2009, doi: 10.1177/0093854808326545.
[28] E. Marzolf. *Exclusif: Le Ministère De La Justice Renonce á Son Algorithme DataJust*. Accessed: Mar. 7, 2024. [Online]. Available: https://acteurspublics.fr/articles/exclusif-le-ministere-de-la-justice-renonce-a-son-algorithme-datajust
[29] G. M. Csányi, D. Nagy, R. Vági, J. P. Vadász, and T. Orosz, "Challenges and open problems of legal document anonymization," *Symmetry*, vol. 13, no. 8, p. 1490, Aug. 2021, doi: 10.3390/sym13081490.
[30] Super AI. *Document Redact*. Accessed: Mar. 7, 2024. [Online]. Available: https://super.ai/super-redact/document-redact
[31] F. Hassan, J. Domingo-Ferrer, and J. Soria-Comas, "Anonymization of unstructured data via named-entity recognition," in *Proc. 15th MDAI*, Mallorca, Spain, 2018, pp. 296–305.
[32] D. Licari and G. Comandé, "ITALIAN-LEGAL-BERT: A pre-trained transformer language model for Italian law," in *Proc. KM4LAW Workshop*, Bolzano, Italy, Sep. 2022, pp. 1–16.
[33] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, and H. Anderson, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," 2018, *arXiv:1802.07228*.
[34] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. Accessed: Mar. 7, 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[35] A. Richterich. (2018). *The Big Data Agenda: Data Ethics and Critical Data Studies*. Univ. Westminster Press. London, U.K. Accessed: Mar. 7, 2024. [Online]. Available: https://mediarep.org/server/api/core/bitstreams/98fb60cd-d9ba-47a3-afe4-3e3b43fee176/content
[36] M. Butterworth, "The ICO and artificial intelligence: The role of fairness in the GDPR framework," *Comput. Law Secur. Rev.*, vol. 34, no. 2, pp. 257–268, Apr. 2018, doi: 10.1016/j.clsr.2018.01.004.
[37] *SICID (District Civil Litigation Information System)*. Accessed: Mar. 2, 2024. [Online]. Available: https://ca-salerno.giustizia.it/cmsresources/cms/documents/PTEL-MU-RA-028-NS-Manuale_SICID.pdf
[38] D. Sánchez and M. Batet, "C-sanitized: A privacy model for document redaction and sanitization," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 1, pp. 148–163, Jan. 2016, doi: 10.1002/asi.23363.
[39] N. Arajarvi, and L. Holden. (2022). *GDPR Compliant Guidelines for Processing Personal Data in Legal Documents*. Accessed: Mar. 7, 2024. [Online]. Available: https://culturalexpertise.net/wp-content/uploads/2022/01/gdprcompliantguidelinesforlegaldocuments.pdf
[40] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang. *Doccano: Text Annotation Tool for Human*. Accessed: Mar. 7, 2024. [Online]. Available: https://github.com/doccano/doccano
[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," " in *Proc. 30th NIPS*, Long Beach, CA, USA, Dec. 2017, pp. 1–11.
[42] Y. A. Chung, Y. Zhang, W. Han, C. C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*. IEEE, Dec. 2021, pp. 244–250.
[43] Wikipedia Contributors. (2024). *Wikipedia, The Free Encyclopedia*. Accessed: Mar. 7, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Main_Page
[44] OPUS. *CORPUS, Open Parallel Corpora*. Accessed: Mar. 7, 2024. [Online]. Available: https://opus.nlpl.eu
[45] *OSCAR*. Accessed: Mar. 7, 2024. [Online]. Available: https://oscar-project.org/
[46] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101805, doi: 10.1016/j.inffus.2023.101805.
[47] K. M. Richmond, S. M. Muddamsetty, T. Gammeltoft-Hansen, H. P. Olsen, and T. B. Moeslund, "Explainable AI and law: An evidential survey," *Digit. Soc.*, vol. 3, no. 1, pp. 1–33, Apr. 2024, doi: 10.1007/s44206-023-00081-z.
[48] L. Górski and S. Ramakrishna, "Explainable artificial intelligence, lawyer's perspective," in *Proc. 18th Int. Conf. Artif. Intell. Law*, Jun. 2021, pp. 60–68, doi: 10.1145/3462757.3466145.
[49] C. Badii, P. Bellini, A. Difino, and P. Nesi, "Smart city IoT platform respecting GDPR privacy and security aspects," *IEEE Access*, vol. 8, pp. 23601–23623, 2020, doi: 10.1109/ACCESS.2020.2968741.
[50] R. A. Ruggeto. *Giustizia Agile Project of PON Governance and Institutional Capacity*. Accessed: Mar. 7, 2024. [Online]. Available: https://www.unitus.it/it/unitus/mappatura-della-ricerca/articolo/giustizia-agile

ENRICO COLLINI is currently pursuing the Ph.D. degree with the DINFO Department, University of Florence. He is also an engineer. His research interests include deep learning, mobility, and data models.

PAOLO NESI (Member, IEEE) is a Full Professor with the DINFO Department, University of Florence, where he is also a Chief of DISIT Laboratory. He is and has been a coordinator of several research and development multipartner international research and development projects. His research interests include machine learning, massive parallel and distributed systems, physical models, the IoT, mobility, big data analytic, semantic computing, formal model, and machine learning. He has been the chair of several international conferences.

CLAUDIA RAFFAELLI is a Research Fellow and a Software Engineer with DINFO, University of Florence. Her research interests include artificial intelligence and deep model intellectual property protection.

FRANCESCO SCANDIFFIO is currently pursuing the master's degree in computer engineering with the University of Florence. He is also a Research Scholar with DINFO, University of Florence. His research interests include artificial intelligence and quantum internet protocols.

Open Access funding provided by 'Università degli Studi di Firenze' within the CRUI CARE Agreement