Artificial Intelligence and Law (2023) 31:195–212
https://doi.org/10.1007/s10506-021-09306-3

ORIGINAL RESEARCH

Rethinking the field of automatic prediction of court decisions

Masha Medvedeva[1,2] · **Martijn Wieling**[1] · **Michel Vols**[2]

Accepted: 7 November 2021 / Published online: 25 January 2022

Abstract

In this paper, we discuss previous research in automatic prediction of court decisions. We define the difference between outcome identification, outcome-based judgement categorisation and outcome forecasting, and review how various studies fall into these categories. We discuss how important it is to understand the legal data that one works with in order to determine which task can be performed. Finally, we reflect on the needs of the legal discipline regarding the analysis of court judgements.

Keywords Judicial decisions · Machine learning · Case law · Natural language processing

1 Introduction

Automatic analysis of legal documents is a useful, if not necessary task in contemporary legal practice and research. Of course, data analysis should be conducted in a methodologically sound, transparent and thorough way. These requirements are extra important with regard to legal data. The stakes that legal professionals such as lawyers, judges and other legal decision-makers deal with and the cost of error in this field make it very important that automatic processing and analysis are done well. That means that it is essential to understand how the automated systems used in the analysis work, what legal data exactly is analysed and for what purpose.

✉ Masha Medvedeva
m.medvedeva@rug.nl

Martijn Wieling
m.b.wieling@rug.nl

Michel Vols
m.vols@rug.nl

[1] Center for Language and Cognition Groningen, University of Groningen, Groningen, The Netherlands

[2] Department of Legal Methods, University of Groningen, Groningen, The Netherlands

The need for established practices and methodology is becoming more urgent with the growing availability of data. In striving for transparency, many national and international courts in Europe adhere to the directive to promote accessibility and re-use of public sector information[1] and publish their documents online (Marković and Gostojić 2018). This is also the case for many other courts around the world.[2] Digital access to a large amount of published case law provides a unique opportunity to process this data automatically on a large scale using natural language processing (NLP) techniques.

In this paper we review previous work on applying NLP techniques to court decisions, and discuss the methodological issues as well as good practices. While automatic legal analysis is an enormous field which has been around for some time, in this paper we focus solely on the recent development of using machine learning techniques for classifying court decisions. This sub-field has expanded drastically in the past 6 years with papers that attempt to predict decisions of various courts around the world. We subsequently discuss whether it is fair to say that they indeed succeed. Our main finding is that many of the papers under review claiming to predict decisions of the courts using machine learning actually perform one of three different tasks.

In the following section, we define the scope of review we conducted. Next, in Sect. 3 we discuss (our terminology of) different types of tasks within the field of automatic analysis of court decisions and how previous research falls within those categories. We examine the purpose of such research for each task, as well as good practices and potential pitfalls. We then discuss our survey in Sect. 4. In Sect. 5 we summarise and conclude our work.

2 Scope of the review

We limit our review to the papers that use machine learning techniques and claim to be predicting court decisions. The publication dates range from 2015 to (June) 2021.[3] We specifically chose these years, as this is when machine learning in this field became popular. If a paper included in our review attempts multiple tasks, we only focus on the experiment(s) that focus on predicting judicial decisions. While our survey is meant to provide an exhaustive overview, we may have inadvertently missed some research in the field.

While we already mentioned that the research in the field is growing, not all courts share (all) their case law online. Furthermore, the majority of available case law is extremely varied in its outcomes, which may make it harder to set up an outcome prediction task. For this reason, research often focuses on a relatively

[1] https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data, accessed on 11/10/2021.

[2] See, for instance, case law of the Constitutional Court of South Africa available at: https://collections.concourt.org.za.

[3] For description of earlier approaches in automatic prediction of court decision with and without using machine learning we refer to Ashley and Brüninghaus (2009)

restricted set of courts. In this paper, we surveyed publications that use machine learning approaches and focus on case-law of the US Supreme Court (Sharma et al. 2015; Katz et al. 2017; Kaufman et al. 2019), the French court of Cassation (Şulea et al. 2017b; Sulea et al. 2017a), the European Court of Human Rights (Aletras et al. 2016; Liu and Chen 2017; Chalkidis et al. 2019; Kaur and Bozic 2019; O'Sullivan and Beel 2019; Visentin et al. 2019; Chalkidis et al. 2020; Condevaux 2020; Medvedeva et al. 2020a, b; Quemy and Wrembel 2020; Medvedeva et al. 2021), Brazilian courts (Bertalan and Ruiz 2020; Lage-Freitas et al. 2019), Indian courts (Bhilare et al. 2019; Shaikh et al. 2020; Malik et al. 2021), UK courts (Strickson and De La Iglesia 2020), German courts (Waltl et al. 2017), the Quebec Rental Tribunal (Salaün et al. 2020) (Canada), the Philippine Supreme Court (Virtucio et al. 2018), the Thai Supreme Court (Kowsrihawat et al. 2018) and the Turkish Constitutional Court (Sert et al. 2021). Many of these papers achieve a relatively high performance on their specific task using various machine learning techniques.

The distinction between different tasks in this paper is conditional on the data, but is not contingent on the algorithms used. Consequently, we discuss the following papers from the perspective of which data was used, how it was processed and general performance of the systems using particular data for a particular task. We do not go into detail of the algorithms used for achieving that performance. For the specifics of different systems, we therefore refer the interested reader to the papers at hand. For a more detailed explanation of machine learning classification for legal texts in general, see Medvedeva et al. (2020a) and Dyevre (2020).

3 Terminology and types of judgement classification

In papers that use machine learning for classifying court decisions, different terms and types of tasks are often used interchangeably. For the field to move forward, we therefore argue for a more strict use of terminology. Consequently, in this paper, we use 'judgement' to mean the text of a published judgement. While the word 'outcome' is a very general term, for the purposes of distinguishing between different tasks in the legal context, we define outcome as a specific closed class of labels for verdicts (i.e. with a pre-defined limited number of verdicts). For example, in the context of case law concerning the European Convention on Human Rights (ECHR) the outcome will be a *violation* or a *non-violation* of a specific human right. Other examples of outcomes are *eviction* or *non-eviction* in a housing law context (Vols 2019) or the US Supreme Court affirming or reversing a decision of a lower court. We use 'verdict' and 'decision' as synonyms of 'outcome'.

In this paper we will distinguish between three types of tasks: *outcome identification*, *outcome-based judgement categorisation*, and *outcome forecasting*.[4] In simple terms, outcome identification is the task of identifying the verdict in the full text of the published judgements, judgement categorisation is the task of categorising documents based on the outcome, and outcome forecasting is the task of predicting future decisions of a particular court. At present, these task distinctions are not clearly made in the literature, even by ourselves (Medvedeva et al. 2020a). This is potentially problematic as the different tasks have specific uses, which we will discuss below.

The most likely reason for the ambiguity in terminology is the cross-disciplinary nature of the field, combining law with NLP. When using machine learning in the field of NLP, all three tasks are so-called classification tasks. The most commonly used approach in machine learning, and the one all of the reviewed papers have used, is *supervised learning*. This means that the system is trained on some input data (e.g., facts extracted from a criminal case) that is connected to the labels (outcomes), for instance whether the case was won by the defendant or the prosecution. During the training phase, the model is presented with input data together with their labels in order to infer patterns characterising the relationship between the two. To evaluate the system after training, the system is provided with similar data (*not* used during the training phase), such as other criminal cases, and it then *predicts* the label for each document. Since the label in each task is the outcome, identifying the purpose of these systems within NLP as 'predicting court decisions' is appropriate. However, that meaning does not translate in the same way outside of the NLP domain. Specifically, the word *predict* in the legal domain suggests that one can forecast a decision (of the judge) that has not been made yet, whereas in NLP *predict* merely refers to the methodology and terminology of machine learning. The majority of papers on *predicting* court decisions published today, however, do not attempt to predict decisions of the cases that have not been judged yet. Furthermore, the majority of the work in this interdisciplinary field suggests a benefit for legal professionals, but does not explicitly specify what the models that were introduced can be used for.

To circumvent the use of the ambiguous word *predict*, we therefore suggest using terminology that better reflects the different tasks, and thereby also differentiates between objectives. In order to distinguish between outcome identification, outcome-based judgement categorisation and outcome forecasting it is important to carefully assess the data used in the experiments conducted.

When discussing different papers, we will also refer to their performance scores. The conventional way of reporting the performance of a classification system is by

[4] In principle, there are three additional tasks, namely *charge identification*, *charge-based judgement categorisation* and *charge forecasting*. These tasks involve determining the specific sentence or charge. For example, the number of years someone was sentenced to go to prison in criminal court proceedings. These tasks have most often been investigated for various courts in China (Luo et al. 2017; Ye et al. 2018; Jiang et al. 2018; Liu and Chen 2018; Zhong et al. 2018a, b; Li et al. 2019; Chen et al. 2019; Long et al. 2019; Chao et al. 2019; Fan et al. 2020; Cheng et al. 2020; Tan et al. 2020; Huang et al. 2020). The distinction we make between identification, categorisation and forecasting (and the pitfalls and suggestions regarding this distinction) in this paper, however, hold for these cases as well.

using accuracy or the F1-score. Accuracy is how many of the labels (in our case, outcomes) were classified (i.e. identified, categorised, or forecasted) correctly. The F1-score is a harmonic mean of precision and recall, where precision is the amount of judgements for which the assigned outcome is correct and recall is the percentage of cases with a specific outcome which are classified (i.e. identified, categorised, or forecasted) correctly by the system.

In the following subsections we will make the definitions of the three tasks more explicit, and then give examples from published research for each task. We also highlight the distinct uses of the different tasks for legal professionals.

3.1 Outcome identification

Outcome identification is defined as the task of identifying the verdict within the full text of the judgement, including (references to) the verdict itself. In principle, a machine learning system is often not necessary for such a task, as keyword search (or using simple regular expressions) might suffice.

Outcome identification falls under the field of information extraction and when not confused with predicting court decisions is often also referred to as outcome extraction (e.g., Petrova et al. 2020). Given the growing body of published case law across the world, the automation of this task may be very useful, since many courts publish case law without any structured information (i.e. metadata) available, other than the judgements themselves, and often one may require a database where the judgements are connected to the verdicts in order to conduct research. At present and to our knowledge, most of such work is generally done manually, as a human can do this task with 100% accuracy (by simply reading the case and finding the verdict in it).

Automation of outcome identification allows one to save time when collecting this information. While the task is not necessarily always trivial for a machine and depends on how the verdict is formulated (see, for instance, Vacek and Schilder (2017), Petrova et al. (2020) and Tagny-Ngompé et al. (2020)), there is nonetheless an expectation that these automated systems should achieve (almost) perfect performance to justify the automation. However, the approach to outcome identification is highly dependent on the structure of judgements in a particular legal domain or jurisdiction and the language of the case law. As a result, a system that automatically identifies a verdict in a particular set of judgements cannot be applied easily to case law of courts in other legal domains or other jurisdictions.

3.1.1 Research in outcome identification

A total of eight papers that aimed to predict court decisions (see Table 1) were performing the outcome identification task. These papers use the text of the final judgements published by the court that contain references to the verdict or the verdict itself.

One of the earliest papers that tried predicting court decisions using the text of the judgement is Aletras et al. (2016). The authors used a popular machine learning

Table 1 Research that falls under the category of outcome identification, including relevant court, the (best) performance

Paper	Court	Max. performance
Aletras et al. (2016)	ECtHR	79%
Liu and Chen (2017)	ECtHR	88%
Sulea et al. (2017a, b)	French Court of Cassation	99%
Virtucio et al. (2018)	Philippine Supreme Court	59%
Lage-Freitas et al. (2019)	Brazilian courts (appeal)	79% (F1)
Visentin et al. (2019)	ECtHR	79%
Bertalan and Ruiz (2020)	São Paolo Justice Court	98% (F1)
Quemy and Wrembel (2020)	ECtHR	96%

When instead of accuracy, the F1-score (the average between precision and recall) is used as a performance indicator, this is indicated

algorithm, a Support Vector Machine (SVM) to predict decisions of the European Court of Human Rights (ECtHR). Their model aimed to predict the court's decision by extracting the available textual information from relevant sections of the ECtHR judgements and reached an average accuracy of 79% for three separate articles of the ECHR. While the authors did exclude the verdict itself (or the complete section containing the verdict), they still used the remaining text of the judgements, which often still included specific references to the final verdict (e.g., 'Therefore there is a violation of Article 3'). While their work was positioned as predicting the outcome of court cases, the task they conducted was therefore restricted to outcome identification.

Other studies focusing on the ECtHR included Liu and Chen (2017), Visentin et al. (2019), and Quemy and Wrembel (2020). Since Liu and Chen (2017) and Visentin et al. (2019) used the same dataset as Aletras et al. (2016), they also conducted the task of outcome identification. Liu and Chen (2017) used similar statistical methods as Aletras et al. (2016) and achieved an 88% accuracy using an SVM, whereas Visentin et al. (2019) achieved an accuracy of 79% using an SVM ensemble. Whereas Quemy and Wrembel (2020) collected a larger dataset for the same court and performed a binary classification task (violation of any article of the ECHR vs. no violation) using neural models, they did not appear to exclude any part of the judgement, thereby restricting their task also to outcome identification (with a concomitant high accuracy of 96% using a range of statistical methods). These studies show that automatic outcome identification to a large extent is possible for the ECtHR. However, from a legal perspective this task is not very useful, as the verdict has already been categorised on the ECtHR website.

The studies on the basis of the ECtHR illustrate two broad categories of papers which aim at predicting court judgements, but instead are outcome identification tasks. The first category consists of studies which were only partially successful in removing the information about (references to) the verdict. Besides the aforementioned studies of Aletras et al. (2016), Liu and Chen (2017) and Visentin et al. (2019), the studies of Şulea et al. (2017a, b) suffer from the same problem. They focus on the French Court of Cassation and reach an accuracy of up to 96%. While they masked the words containing the verdict, various words which were found to be important for the prediction of their model appeared to be closely related to the outcome description. Consequently, they were not completely successful in filtering out the information about the outcome.

The second category consists of studies which do not filter out any information out of the judgement at all (or do not mention filtering out this type of information), such as Quemy and Wrembel (2020). Virtucio et al. (2018) are explicit in not filtering out the actual court decision of the Philippine Supreme Court (due to a lack of consistent sectioning in the judgement descriptions) when predicting its judgement. Nevertheless, their accuracy was rather low at only 59%. In addition, there is a number of papers that do not specify any pre-processing steps to remove the information that may contain the verdict. Examples are Lage-Freitas et al. (2019) who deal with appeal cases of Brazilian courts (with an F1-score of 79%) and Bertalan and Ruiz (2020) who worked on second-degree murder and corruption cases tried in São Paolo Justice Court (with an F1-score of up to 98%).

3.2 Outcome-based judgement categorisation

Outcome-based judgement categorisation is defined as categorising court judgements based on their outcome by using textual or any other information published with the final judgement, but excluding (references to) the verdict in the judgement. Since the outcomes of such cases are published and no longer need to be 'predicted', this task is mainly useful for identifying predictors (facts, arguments, judges, etc) of court decisions within the text of judgements. To avoid the system *identifying* the outcome within the text of the judgement and in order for it to learn new information any references to the verdict need to be removed.

While an algorithm may perform very well on the categorisation task, the obtained categories are not useful by themselves. As the documents used by the system are only available when the judgements are made and public, the outcome categorisation does not contribute any new information (one can simply extract the verdict from the published judgement). This view is also supported by Bex and Prakken (2021) who insist that the ability to categorise decisions without explaining why the categorisation was made, does not provide any useful information and may even be misleading. The performance of a machine learning model for judgement categorisation, however, may provide useful information about how informative the characteristic features are. To enable feature extraction, it is important that the system is not a 'black box' (such as many of the more recent neural classification models). Therefore, rather than 'predicting court decisions' the main objective of the outcome-based judgement categorisation task should be to identify *predictors* underlying the categorisations.

As we only discuss publications that categorise judgements on the basis of the outcome of the case, we will refer to outcome-based judgement categorisation simply as judgement categorisation.

Table 2 Research that falls under the category of outcome-based judgement categorisation, including relevant court, whether or not the most important features were extracted (FI), and the best achieved performance

Paper	Court	FI	Max. performance
Kowsrihawat et al. (2018)	Thai Supreme Court	×	67%
Chalkidis et al. (2019)	ECtHR	✓	82% (F1)
Kaufman et al. (2019)	SCOTUS	×	77%
Kaur and Bozic (2019)	ECtHR	×	82%
O'Sullivan and Beel (2019)	ECtHR	×	69%
Chalkidis et al. (2020)	ECtHR	×	83% (F1)
Condevaux (2020)	ECtHR	×	88%
Medvedeva et al. (2018, 2020a)	ECtHR	✓	75%
Salaün et al. (2020)	Québec Rental Tribunal	×	85%
Shaikh et al. (2020)	Dehli District Court	×	92%
Strickson and De La Iglesia (2020)	UK highest Court of Appeal	✓	69%
Sert et al. (2021)	Turkish Constitutional Court	×	98% (F1)
Malik et al. (2021)	Indian Supreme Court Court	×	77%
Medvedeva et al. (2021)	ECtHR	×	92% (F1)

When instead of accuracy, the F1-score (the average between precision and recall) is used as a performance indicator, this is indicated

3.2.1 Research in outcome-based judgement categorisation

Most of the papers in the field categorise judgements. The papers surveyed that involve judgement categorisation can be found in Table 2. For all fifteen papers, we indicate the paper itself, the court, whether or not the authors provide a method of analysing feature importance (FI) and consequently identify specific predictors of the outcome within the text, and the maximum performance.

Within these studies, two broad categories can be distinguished depending on which type of data they use. On the one hand, most studies use the raw text, explicitly selecting parts of the judgement which does not include (references to) the verdict. On the other hand, there are (fewer) studies which manually annotate data and use that as a basis for the categorisation.

Kowsrihawat et al. (2018) used the raw text to categorise (with an accuracy of 67%) the documents of the Thai Supreme Court on the basis of the facts of the case and the text related to the legal provisions in the cases such as murder, assault, theft, fraud and defamation using a range of statistical and neural methods. Medvedeva et al. (2018), Medvedeva et al. (2020a) categorised (with an accuracy of at most 75%) decisions of the ECtHR using only the facts of the case (i.e. a separate section in each ECtHR judgement). Notably, Medvedeva et al. (2020a) identified the top predictors (i.e. sequences of one or more words) for each category, which was possible due to the (support vector machine) approach they used. Strickson and De La Iglesia (2020) worked on categorising judgements of the UK Supreme Court and compared several systems trained on the raw text of the judgement (without the verdict) and reported an accuracy of 69%, while also presenting the top predictors for each class. Sert et al. (2021) categorised cases of the Turkish Constitutional Court related to public morality and freedom of expression using a traditional neural multi-layer perceptron approach with an average accuracy of 90%. Similarly to Medvedeva et al. (2020a), Chalkidis et al. (2019) also investigated the ECtHR using the facts of the case, and proposed several neural methods to improve categorisation performance (up to 82%). They additionally proposed an approach (a hierarchical attention network) to identify which words and facts were most important for the classification of their systems. In their subsequent study Chalkidis et al. (2020) used a more sophisticated neural categorisation algorithm which was specifically tailored for legal data (LEGAL-BERT). Unfortunately, while their approach did show an improved performance (with an F1-score of 83%) it was not possible to determine the best predictors of the outcome due to the system's complexity. Medvedeva et al. (2021) reproduced the algorithms in Chalkidis et al. (2019) and Chalkidis et al. (2020) in order to compare their performance for categorisation and forecasting tasks (see below) for a smaller subset of ECtHR cases, and achieved an F1-score of up to 92% for categorising judgements of 2019. The scores however varied throughout the years. For example, categorisation of cases from 2020 did not surpass 62%. Several other categorisation studies (with accuracies ranging between 69 and 88%) focused on the facts of the ECtHR, but likewise did not investigate the best predictors (Kaur and Bozic 2019; O'Sullivan and Beel 2019; Condevaux 2020). Malik et al. (2021) used neural methods to develop a system that categorised Indian Supreme Court Decisions achieving 77% accuracy. As their main focus was to develop an explainable system, they used an approach which allowed them to investigate the importance of their features, somewhat similar to the approach of Chalkidis et al. (2020).

Manually annotated data was used by Kaufman et al. (2019) who focused on data from the US Supreme Court (SCOTUS) Database (Spaeth et al. 2014) and achieved an accuracy of 75% using statistical methods (i.e. AdaBoosted decision trees). However, they did not investigate the most informative predictors. Shaikh et al. (2020) also used manually annotated data to categorise the decisions of murder-cases of the Delhi District Court with an accuracy of up to 92% using classification and regression trees. These authors manually annotated 18 features, including whether the injured is dead or alive, the type of evidence, the number of witnesses et cetera. Importantly, they analysed the impact of each type of feature for each type of outcome.

Finally, Salaün et al. (2020) essentially combined the two types of predictors, by not only extracting a number of characteristics from the cases of Rental Tribunal of Quebec (including the court location, judge, types of parties, et cetera), but also using the raw text of the facts (as well as the complete text excluding the verdict), achieving a performance of at most 85% with a French BERT model, FlauBERT.

Notably, the performance of Sert et al. (2021) was very high. Despite the high success rate of their system, however, the authors warn against using it for decision-making. Nevertheless, they do suggest that their system can potentially be used for prioritising the cases that have a higher likelihood to end up in a violation. This suggestion mirrors the proposition made by Aletras et al. (2016) for potentially using their system to prioritise cases with human rights violations. In both cases, however, the experiments were conducted using data extracted from the final judgements of the court, and the performance of these systems using data compiled before the verdict was reached (i.e. information necessary to prioritise cases) is unknown. Making these types of recommendations is therefore potentially problematic.

Many categorisation papers shown in Table 2 claim to be useful for legal aid. However, as we argued before, categorisation as such is not a useful task, given that the verdict can simply be read in the judgement text. To be useful, it is essential that categorisation performance is supplemented with the most characteristic features (i.e. predictors). Unfortunately, only a minority of studies provides this information. And even if they do, the resulting features, especially when using the raw text (i.e. characteristic words or phrases), may not be particularly meaningful.

In an attempt to be maximally explainable, Collenette et al. (2020) suggest using Abstract Dialectical Framework instead of machine learning. They apply this framework to deducing the verdict from the text of judgements of the ECtHR regarding Article 6 of the ECHR (the right to a fair trial). The system requires the user to answer a range of questions, and on the basis of the provided answers, the model determines whether there was a violation of the right to a fair trial or not. The questions for the system were derived by legal experts, and legal expertise is also required to answer these questions (Collenette et al. 2020). While their system seemed to perform flawlessly when tested on ten cases, we face the same issue as with the machine learning systems. Specifically, the main input data is based on the final decision that has already been made by the judge. For instance, one of the questions that the model requires to be answered is whether the trial was independent and impartial, which is a question that has to be decided on by the judge. While this type of tool may potentially 1 day be used for judicial support, for example, as a checklist for a judge when making a specific decision, it is unable to actually forecast decisions in advance, or point to external factors that are not identified by legal experts.

3.3 Outcome forecasting

Outcome forecasting is defined as determining the verdict of a court on the basis of textual information about a court case which was available *before* the verdict was made (public). This textual information can, for instance, be submissions by the parties, or information (including judgements) provided by lower courts to predict the decisions of a higher court, such as the US Supreme Court. Forecasting thereby comes with the essential assumption that the input for the system was not influenced in any way by the final outcome that it forecasts. In contrast to *outcome-based judgement categorisation*, it is useful to evaluate how well the algorithm is able to predict the outcome of cases. For example, individuals may use such algorithms to evaluate how likely it is that they will win a court case. Similarly to *judgement categorisation*, determining the factors underlying a well-performing model is useful as well. While identification and categorisation tasks only allow one to extract information and analyse already made court decisions, forecasting allows one to predict future decisions that have not been made yet. Note that whether or not a model was trained

Table 3 Research that falls under the category of outcome forecasting, including relevant court, the data used for forecasting, the best performance

Paper	Court	Data	Max. performance
Sharma et al. (2015)	SCOTUS	Court of Appeal info	70%
Katz et al. (2017)	SCOTUS	Court of Appeal info	70%
Waltl et al. (2017)	German Court of Appeal (Tax Law)	Decision of the lower (fiscal) courts	57% (F1)
Medvedeva et al. (2020b)	ECtHR	Facts as communicated to the parties	75%
Medvedeva et al. (2021)	ECtHR	Facts as communicated to the parties	66% (F1)

When instead of accuracy, the F1-score (the average between precision and recall) is used as a performance indicator, this is indicated

on older cases than it was evaluated on (e.g., the ‘predicting the future’ experiment conducted by Medvedeva et al. 2020a) does not affect its classification as a judgement categorisation as opposed to a judgement forecasting task. Only the type of data affects which task it is. Since Medvedeva et al. (2020a) use extracted data from the court judgements, their task is still an outcome-based judgement categorisation task.

3.3.1 Research in outcome forecasting

Table 3 lists the papers that focus on forecasting court verdicts. While many publications focus on ‘predicting court decisions’, only five papers satisfy our criteria for outcome forecasting. We can observe that the performance of these studies is lower than for the categorisation and identification tasks. This is not surprising as forecasting can be expected to be a harder task. Given the small number of papers, we discuss each of these in some detail.

The advantage of working with the US Supreme Court databases is that it attracts much attention. Consequently, all data from the trials are always systematically and manually annotated by legal experts with many variables immediately after the case was tried. Sharma et al. (2015) and Katz et al. (2017) both use variables available to the public once the case was moved to the Supreme Court, but before the decision was made to forecast decisions of SCOTUS. Sharma et al. (2015) use neural methods, whereas Katz et al. (2017) use the more traditional technique of random forests. Both approaches resulted in forecasting 70% of the outcomes correctly, which was a small improvement over the 68% baseline accuracy where the petitioner always wins (suggested by Kaufman et al. 2019). Moreover, Sharma et al. (2015) present the importance of various variables in their model, therefore potentially enabling a more thorough legal analysis of the data. The variables used in both studies contained information about the courts and the proceedings but hardly any variables pertaining to the facts of the case.

Waltl et al. (2017) attempted to forecast decisions of the German appeal court in the matters of Tax Law (Federal Fiscal Court). The authors used the documents and meta-data of the case (e.g., year of dispute, court, chamber, duration of the case, et cetera) from the court of first instance. They extracted keywords from the facts and (lower) court reasoning to forecast decisions. They tried a range of methods, but selected the best-performing naive Bayes classifier as their final model. Their relatively low F1-score of 0.57 indicates that it may have been a rather difficult task, however.

Medvedeva et al. (2020b) used raw text and facts within documents that were published by the ECtHR (sometimes years) before the final judgement. These documents are known as 'communicated cases'. Specifically they used the facts as presented by the applicant and then communicated by the Court to the State as a potential violator of the human rights. The communicated cases reflect the side of the potential victim, and are only communicated when no similar cases have been processed by the court before. Consequently, these documents include a very diverse set of facts, and different issues (although all within the scope of the European Convention on Human Rights) are covered in them. Medvedeva et al. (2020b) reported an accuracy of 75% using SVMs on their dataset (the model is re-trained and run again every month). This system is integrated in an online platform that also highlights the sentences or facts within the text of these (communicated) cases that are most important for the model's decision.[5]Medvedeva et al. (2021) used a slightly different dataset of the same documents (i.e. only cases with the judgement in English were included, but the dataset was expanded by adding cases that resulted in inadmissibility based on merit) and retrained the model per year (as opposed to per month in Medvedeva et al. (2020b). The authors compared how the state-of-the-art algorithms for this court, BERT (Chalkidis et al. 2019), LegalBERT (Chalkidis et al. 2020), and SVMs (Medvedeva et al. 2020a, b) perform on data available *before* the final judgement and *with* the final judgement. The results showed that forecasting is indeed a much harder task, as the models achieved a maximum F1-score of 66% as opposed to 92% for categorisation of the same cases.

4 Discussion

It is clear that 'predicting court decisions' is not an unambiguous task. There is therefore a clear need to carefully identify the *objective* of the experiments before conducting them. We believe such an objective has to be rooted within the specific needs of the legal community to prevent systems being developed of which the authors believe them to be useful, whereas they do not have any meaningful application in the legal field at all. The purpose of our paper was to provide some terminology which may be helpful for this.

While researchers may believe they are 'predicting court decisions', very infrequently this involves actually being able to predict the outcome of *future*

[5] https://jurisays.com.

judgements. In fact, predicting court decisions sometimes (likely inadvertently, due to sub-optimal filtering or insufficient knowledge about the exact dataset) ended up not being anything other than identifying the outcome from the judgement text. While sophisticated approaches were often put forward in those cases, a simple keyword search might already have resulted in a higher performance for this identification task. Most often, however, predicting court decisions was found to be equal to the task of categorising the judgements according to the verdicts. This is not so surprising given the available legal datasets, which more often contain complete judgements than documents which were produced before the verdict was known.

In sum, to identify the exact task, and the concomitant goals which are useful from a legal perspective, it is essential that researchers are well aware of the type of data they are analysing. Unfortunately, this is frequently not the case. For example, several researchers (Chalkidis et al. 2019; Quemy and Wrembel 2020; Condevaux 2020) have recently started to develop (multilabel classification) systems, which are able to predict which articles were invoked in an ECtHR case. However, this task is not relevant from a legal perspective, as articles which are potentially violated have to be specified when petitioning the ECtHR.

Therefore when creating a new application, for instance, using data from another court, one should clearly determine the goal of such a system first, and then review whether the data for the established task is available. Specifically, one needs full judgements for the outcome identification task. In case of a judgement categorisation task, full judgements from which the outcomes can be removed are necessary. If the system needs to perform a forecasting task, it requires data available before the judgement is made.

For all of the above tasks, explainability (i.e. being able to determine the importance of various features when determing the model's outcome) helps to better analyse the performance and gain insight into the workings of the system. However, explainability is *essential* for judgement categorisation, as this task is reliant on the ability to investigate which features are related to the outcome.

As we mentioned before, the identification task does not always require the use of machine learning techniques. This task can often be solved with a keyword search which does not require any annotated data. Using machine learning is necessary when the judgement text is not very structured, and when more complex descriptions of the outcome need to be extracted. For both the judgement categorisation task and the forecasting task, statistics may be useful to assess the relation between predetermined factors and the outcome, whereas for categorisation task machine learning techniques allow for discovering new patterns and factors within the judgements that may have not been considered previously. Similarly, machine learning can be used to forecast future court decisions by training the system on the decision that the court has made in the past. To illustrate these three tasks, their goals and requirements, a flow-chart is shown in Fig. 1.

Finally, we would like to emphasise that while the approaches discussed in this paper can be suitably used in legal analysis, and for example to try to understand past court decisions, none of the systems capable of solving any of the discussed tasks are appropriate for making court decisions. Judicial decision-making requires

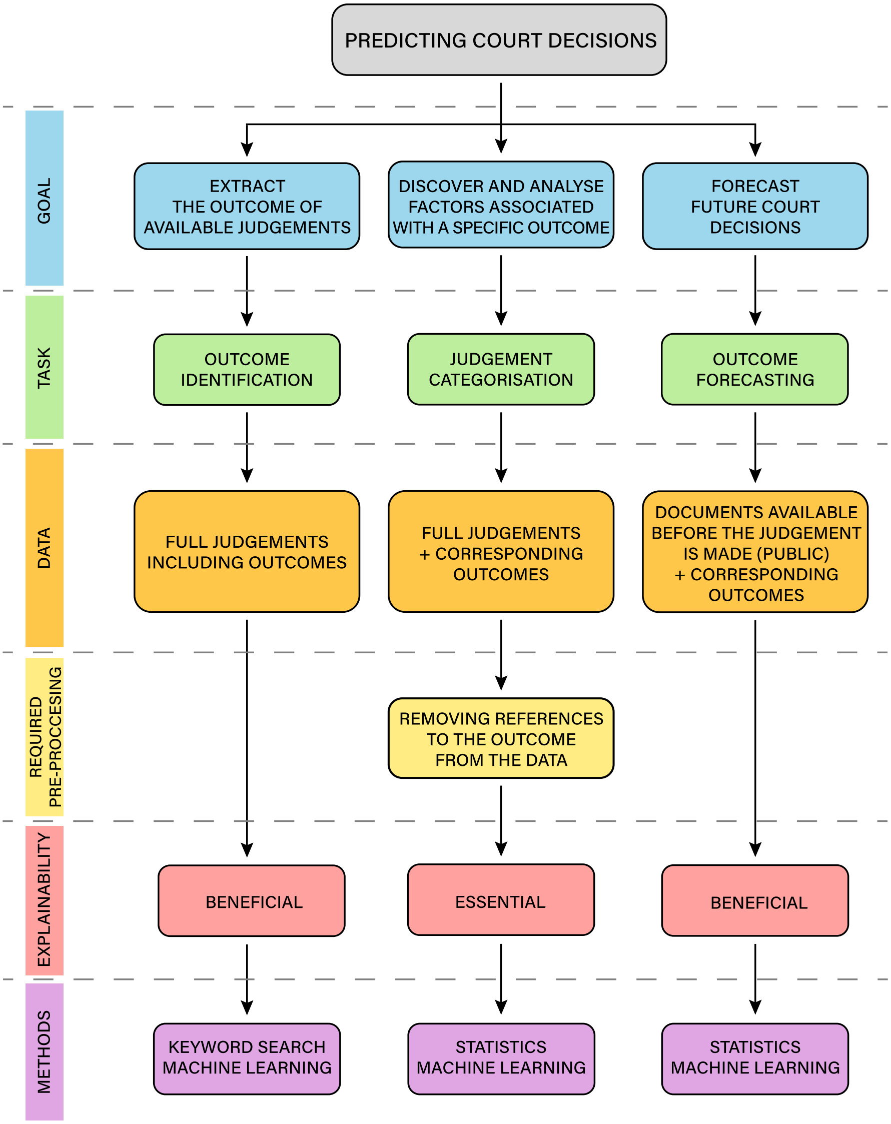
PREDICTING COURT DECISIONS
GOAL
TASK
DATA
REQUIRED PRE-PROCESSING
EXPLAINABILITY
METHODS
EXTRACT THE OUTCOME OF AVAILABLE JUDGEMENTS
DISCOVER AND ANALYSE FACTORS ASSOCIATED WITH A SPECIFIC OUTCOME
FORECAST FUTURE COURT DECISIONS
OUTCOME IDENTIFICATION
JUDGEMENT CATEGORISATION
OUTCOME FORECASTING
FULL JUDGEMENTS INCLUDING OUTCOMES
FULL JUDGEMENTS + CORRESPONDING OUTCOMES
DOCUMENTS AVAILABLE BEFORE THE JUDGEMENT IS MADE (PUBLIC) + CORRESPONDING OUTCOMES
REMOVING REFERENCES TO THE OUTCOME FROM THE DATA
BENEFICIAL
ESSENTIAL
BENEFICIAL
KEYWORD SEARCH MACHINE LEARNING
STATISTICS MACHINE LEARNING
STATISTICS MACHINE LEARNING

Fig. 1 Flowchart illustrating the goals and requirements for the three court decision prediction tasks

(among others) knowledge about the law, knowledge about our ever-changing world, and arguments to be weighed. This is very different from the (sometimes very sophisticated) pattern-matching capabilities of the systems discussed in this paper.

5 Conclusion

In this paper, we have proposed several definitions for analysing court decisions using computational techniques. Specifically, we discussed the difference between forecasting decisions, categorising judgements according to the verdict and identifying the outcome based on the text of the judgement. We also highlighted the specific potential goals associated with each of these tasks and illustrated that each task is strongly dependent on the type of data used.

The availability of enormous amounts of legal (textual) data in combination with the legal discipline being relatively methodologically conservative (Vols 2021) has enabled researchers from various other fields to attempt to analyse these data. However, to conduct meaningful tasks, we argue for more interdisciplinary collaborations, not only involving technically skilled researchers, but also legal scholars to ensure meaningful legal questions are answered, and this new and interesting field is propelled forwards.

References

Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V (2016) Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. PeerJ Comput Sci 2:e93

Ashley KD, Brüninghaus S (2009) Automatically classifying case texts and predicting outcomes. Artif Intell Law 17(2):125–165

Bertalan VGF, Ruiz EES (2020) Predicting judicial outcomes in the Brazilian legal system using textual features. In: DHandNLP@ PROPOR, pp 22–32

Bex F, Prakken H (2021) On the relevance of algorithmic decision predictors for judicial decision making. In: Proceedings of the 19th international conference on artificial intelligence and law (ICAIL 2021). ACM Press

Bhilare P, Parab N, Soni N, Thakur B (2019) Predicting outcome of judicial cases and analysis using machine learning. Int Res J Eng Technol (IRJET) 6:326–330

Chalkidis I, Androutsopoulos I, Aletras N (2019) Neural legal judgment prediction in English. In: Proceedings of the 57th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 4317–4323. https://doi.org/10.18653/v1/P19-1424. https://www.aclweb.org/anthology/P19-1424

Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I (2020) LEGAL-BERT: "preparing the muppets for court". In: Proceedings of the 2020 conference on empirical methods in natural language processing: findings, pp 2898–2904

Chao W, Jiang X, Luo Z, Hu Y, Ma W (2019) Interpretable charge prediction for criminal cases with dynamic rationale attention. J Artif Intell Res 66:743–764

Chen H, Cai D, Dai W, Dai Z, Ding Y (2019) Charge-based prison term prediction with deep gating network. arXiv preprint arXiv:1908.11521

Cheng X, Bi S, Qi G, Wang Y (2020) Knowledge-aware method for confusing charge prediction. In: CCF international conference on natural language processing and Chinese computing. Springer, pp 667–679

Collenette J, Atkinson K, Bench-Capon TJ (2020) An explainable approach to deducing outcomes in European Court of Human Rights cases using ADFs. In: COMMA, pp 21–32

Condevaux C (2020) Neural legal outcome prediction with partial least squares compression. Stats 3(3):396–411

Dyevre A (2020) Text-mining for lawyers: how machine learning techniques can advance our understanding of legal discourse. Available at SSRN 3734430

Fan Y, Zhang L, Wang P (2020) Leveraging label semantics and correlations for judgment prediction. In: China conference on information retrieval. Springer, pp 70–82

Huang YX, Dai WZ, Yang J, Cai LW, Cheng S, Huang R, Li YF, Zhou ZH (2020) Semi-supervised abductive learning and its application to theft judicial sentencing. In: 2020 IEEE international conference on data mining (ICDM). IEEE, pp 1070–1075

Jiang X, Ye H, Luo Z, Chao W, Ma W (2018) Interpretable rationale augmented charge prediction system. In: Proceedings of the 27th international conference on computational linguistics: system demonstrations, pp 146–151

Katz DM, Bommarito MJ II, Blackman J (2017) A general approach for predicting the behavior of the Supreme Court of the United States. PloS One 12(4):e0174698

Kaufman AR, Kraft P, Sen M (2019) Improving Supreme Court forecasting using boosted decision trees. Polit Anal 27(3):381–387

Kaur A, Bozic B (2019) Convolutional neural network-based automatic prediction of judgments of the European Court of Human Rights. In: AICS, pp 458–469

Kowsrihawat K, Vateekul P, Boonkwan P (2018) Predicting judicial decisions of criminal cases from Thai Supreme Court using bi-directional GRU with attention mechanism. In: 2018 5th Asian conference on defense technology (ACDT). IEEE, pp 50–55

Lage-Freitas A, Allende-Cid H, Santana O, de Oliveira-Lage L (2019) Predicting brazilian court decisions. arXiv preprint arXiv:1905.10348

Li Y, He T, Yan G, Zhang S, Wang H (2019) Using case facts to predict penalty with deep learning. In: International conference of pioneering computer scientists. Springer, Engineers and Educators, pp 610–617

Liu YH, Chen YL (2018) A two-phase sentiment analysis approach for judgement prediction. J Inf Sci 44(5):594–607

Liu Z, Chen H (2017) A predictive performance comparison of machine learning models for judicial cases. In: 2017 IEEE symposium series on computational intelligence (SSCI). IEEE, pp 1–6

Long S, Tu C, Liu Z, Sun M (2019) Automatic judgment prediction via legal reading comprehension. In: China national conference on Chinese computational linguistics. Springer, pp 558–572

Luo B, Feng Y, Xu J, Zhang X, Zhao D (2017) Learning to predict charges for criminal cases with legal basis. In: Proceedings of the 2017 conference on empirical methods in natural language processing. Association for Computational Linguistics, Copenhagen, Denmark, pp 2727–2736. https://doi.org/10.18653/v1/D17-1289. https://www.aclweb.org/anthology/D17-1289

Malik V, Sanjay R, Nigam SK, Ghosh K, Guha SK, Bhattacharya A, Modi A (2021) ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. arXiv preprint arXiv:2105.13562

Marković M, Gostojić S (2018) Open judicial data: a comparative analysis. Soc Sci Comput Rev 38, 295-314

Medvedeva M, Vols M, Wieling M (2018) Judicial decisions of the European Court of Human Rights: looking into the crystal ball. In: Proceedings of the conference on empirical legal studies

Medvedeva M, Vols M, Wieling M (2020a) Using machine learning to predict decisions of the European Court of Human Rights. Artif Intell Law 28:237–266

Medvedeva M, Xu X, Wieling M, Vols M (2020b) Juri says: prediction system for the European Court of Human Rights. In: Legal knowledge and information systems: JURIX 2020: the thirty-third annual conference, Brno, Czech Republic, December 9-11, 2020. IOS Press, vol 334, p 277

Medvedeva M, Üstun A, Xu X, Vols M, Wieling M (2021) Automatic judgement forecasting for pending applications of the European Court of Human Rights. In: Proceedings of the fifth workshop on automated semantic analysis of information in legal text (ASAIL 2021)

O'Sullivan C, Beel J (2019) Predicting the outcome of judicial decisions made by the European Court of Human Rights. In: AICS 2019—27th AIAI Irish conference on artificial intelligence and cognitive science

Petrova A, Armour J, Lukasiewicz T (2020) Extracting outcomes from appellate decisions in US State Courts. In: Legal knowledge and information systems: JURIX 2020: the thirty-third annual conference, Brno, Czech Republic, December 9-11, 2020. IOS Press, vol 334, p 133

Quemy A, Wrembel R (2020) On integrating and classifying legal text documents. In: International conference on database and expert systems applications. Springer, pp 385–399

Salaün O, Langlais P, Lou A, Westermann H, Benyekhlef K (2020) Analysis and multilabel classification of Quebec court decisions in the domain of housing law. In: International conference on applications of natural language to information systems. Springer, pp 135–143

Sert MF, Yıldırm E, İrfan Haşlak (2021) Using artificial intelligence to predict decisions of the Turkish Constitutional Court. Soc Sci Comput Rev

Shaikh RA, Sahu TP, Anand V (2020) Predicting outcomes of legal cases based on legal factors using classifiers. Procedia Comput Sci 167:2393–2402

Sharma RD, Mittal S, Tripathi S, Acharya S (2015) Using modern neural networks to predict the decisions of Supreme Court of the United States with state-of-the-art accuracy. In: International conference on neural information processing. Springer, pp 475–483

Spaeth H, Epstein L, Ruger T, Whittington K, Segal J, Martin AD (2014) Supreme Court database code book

Strickson B, De La Iglesia B (2020) Legal judgement prediction for UK courts. In: Proceedings of the 2020 the 3rd international conference on information science and system, pp 204–209

Sulea OM, Zampieri M, Malmasi S, Vela M, Dinu LP, Van Genabith J (2017a) Exploring the use of text classification in the legal domain. In: Proceedings of the 2nd workshop on automated semantic analysis of information in legal texts (ASAIL 2017)

Şulea OM, Zampieri M, Vela M, van Genabith J (2017b) Predicting the law area and decisions of French Supreme Court cases. In: Proceedings of the international conference recent advances in natural language processing, RANLP 2017. INCOMA Ltd., Varna, Bulgaria, pp 716–722

Tagny-Ngompé G, Mussard S, Zambrano G, Harispe S, Montmain J (2020) Identification of judicial outcomes in judgments: a generalized Gini-PLS approach. Stats 3(4):427–443

Tan H, Zhang B, Zhang H, Li R (2020) The sentencing-element-aware model for explainable term-of-penalty prediction. In: CCF international conference on natural language processing and Chinese computing. Springer, pp 16–27

Vacek T, Schilder F (2017) A sequence approach to case outcome detection. In: Proceedings of the 16th edition of the international conference on artificial intelligence and law, pp 209–215

Virtucio MBL, Aborot JA, Abonita JKC, Avinante RS, Copino RJB, Neverida MP, Osiana VO, Peramo EC, Syjuco JG, Tan GBA (2018) Predicting decisions of the Philippine Supreme Court using natural language processing and machine learning. In: 2018 IEEE 42nd annual computer software and applications conference (COMPSAC). IEEE, vol 2, pp 130–135

Visentin A, Nardotto A, O'Sullivan B (2019) Predicting judicial decisions: a statistically rigorous approach and a new ensemble classifier. In: 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI). IEEE, pp 1820–1824

Vols M (2019) European law and evictions: property, proportionality and vulnerable people. Eur Rev Priv Law 27(4):719–752

Vols M (2021) Legal research. Eleven Publishing, The Hague

Waltl B, Bonczek G, Scepankova E, Landthaler J, Matthes F (2017) Predicting the outcome of appeal decisions in Germany's tax law. In: International conference on electronic participation. Springer, pp 89–99

Ye H, Jiang X, Luo Z, Chao W (2018) Interpretable charge predictions for criminal cases: learning to generate court views from fact descriptions. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 1854–1864. https://doi.org/10.18653/v1/N18-1168. https://www.aclweb.org/anthology/N18-1168

Zhong H, Guo Z, Tu C, Xiao C, Liu Z, Sun M (2018a) Legal judgment prediction via topological learning. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 3540–3549

Zhong H, Xiao C, Guo Z, Tu C, Liu Z, Sun M, Feng Y, Han X, Hu Z, Wang H et al (2018b) Overview of cail2018: legal judgment prediction competition. arXiv preprint arXiv:1810.05851