Artificial Intelligence and Law (2020) 28:237–266
https://doi.org/10.1007/s10506-019-09255-y

ORIGINAL RESEARCH

Using machine learning to predict decisions of the European Court of Human Rights

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

Published online: 26 June 2019

Abstract

When courts started publishing judgements, big data analysis (i.e. large-scale statistical analysis of case law and machine learning) within the legal domain became possible. By taking data from the European Court of Human Rights as an example, we investigate how natural language processing tools can be used to analyse texts of the court proceedings in order to automatically predict (future) judicial decisions. With an average accuracy of 75% in predicting the violation of 9 articles of the European Convention on Human Rights our (relatively simple) approach highlights the potential of machine learning approaches in the legal domain. We show, however, that predicting decisions for future cases based on the cases from the past negatively impacts performance (average accuracy range from 58 to 68%). Furthermore, we demonstrate that we can achieve a relatively high classification performance (average accuracy of 65%) when predicting outcomes based only on the surnames of the judges that try the case.

Keywords Machine learning · Case law · European Court of Human Rights · Natural language processing · Judicial decisions

✉ Masha Medvedeva
m.medvedeva@rug.nl

Michel Vols
m.vols@rug.nl

Martijn Wieling
m.b.wieling@rug.nl

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

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

1 Introduction

Nowadays, when so many courts adhere to the directive to promote accessibility and re-use of public sector information[1] and publish considered cases online, the door for automatic analysis of legal data is wide open. The idea of automation and semi-automation of the legal domain, however, is not new. Search databases for legal data, such as Westlaw and LexisNexis have existed since the early 90s. Today computers are attempting automatic summarization of legal information and information extraction (e.g., *DecisionExpress*[2]), categorization of legal resources (e.g., *BiblioExpress*[3]), and statistical analysis (e.g., *StatisticExpress*[4]).

Language analysis has been used in the legal domain and criminology for already a long time. For example, text classification has been used in forensic linguistics. Whereas in earlier times, such as in the Unabomber case,[5] the analysis was done manually, today we can perform many of these tasks automatically. We now have so-called 'machine learning' software which is able to identify gender (Basile et al. 2017), age (op Vollenbroek et al. 2016), personality traits (Golbeck et al. 2011), and even the identity of an author[6] almost flawlessly.

In this study we address the potential of using language analysis and automatic information extraction in order to facilitate statistical research in the legal domain. More specifically, we demonstrate and discuss the possibilities of Natural Language Processing techniques for automatically predicting judicial decisions of the European Court of Human Rights (ECtHR).

Using machine learning (see Sect. 3), we are able to use a computer to perform quantitative analysis on the basis of the words and phrases that were used in a court case and then based on that analysis 'teach' the computer to predict the decision of the Court. If we can predict the results adequately, we may subsequently analyse which words made the most impact on this decision and thus identify what factors are important for the judicial decisions.

It is very important to note that whenever we are talking about *predicting judicial decisions* we are talking exclusively in respect to the data that we have and approaches that we use. We are not claiming that were we to encounter a potential victim of a human rights violation, we would be able to predict what the decision in their case would be. While our research is aimed at getting closer to that goal, this is not the purpose of the present study.

In the following section we will discuss earlier work involving automatic analysis within the legal domain. In Sect. 3 we discuss how machine learning can be used for classification of legal texts. Section 4 is dedicated to describing data we have used

[1] https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information, Accessed on 01/10/2018.
[2] http://www.nlptechnologies.ca/en/decisionexpress.
[3] http://www.nlptechnologies.ca/en/biblioexpress.
[4] http://www.nlptechnologies.ca/en/statisticexpress.
[5] https://archives.fbi.gov/archives/news/stories/2008/april/unabomber_042408.
[6] https://github.com/sixhobbits/yelp-dataset-2017.

for our experiments. In Sect. 5 we describe three experiments that we have conducted for this study and report the results. In Sects. 6 and 7 respectively we discuss the results and draw conclusions.

2 Background

For centuries, legal researchers applied doctrinal research methods, which refer to describing laws, practical problem-solving, adding interpretative comments to legislation and case law, but also 'innovative theory building (systematization) with the more simple versions of that research being the necessary building blocks for the more sophisticated ones' (Van Hoecke 2011, p. vi). Doctrinal legal research 'provides a systematic exposition of the rules governing a particular legal category, analyses the relationship between rules, explains areas of difficulty and, perhaps, predicts future developments' (see Hutchinson and Duncan 2012, p. 101).

One of the key characteristics of the doctrinal analysis of case law is that the court decisions are manually collected, read, summarized, commented and placed in the overall legal system. Quantitative research methods were hardly used to analyse case law (Epstein and Martin 2010). Nowadays, however, due to the massive amount of case law that is published, it is physically impossible for legal researchers to read, analyse and systematize all the international and national court decisions. In this age of legal big data, more and more researchers start to notice that combining traditional doctrinal legal methods and empirical quantitative methods is a promising approach which is able to help us make sense of all available case law (Custers and Leeuw 2017; Derlén and Lindholm 2018; Goanta 2017; Šadl and Olsen 2017).

In the United States of America, the quantitative analysis of case law has a longer tradition than in other parts of the world. There are several quantitative studies of datasets consisting of case law of American courts. Most of these studies use manually collected and coded case law. Many studies use the Supreme Court Database, which contains manually collected and expertly-coded data on the US Supreme Court's behaviour of the last two hundred years (Katz et al. 2017). A large amount of these studies analyse the relationship between gender or political background of judges and their decision-making (see Epstein et al. 2013; Rachlinski and Wistrich 2017; Frankenreiter 2018).

In countries other than the United States, the use of quantitative methods to analyse case law is not very common (see Vols and Jacobs 2017). For example, Hunter et al. (2008, p. 79) state: 'This tradition has not been established in the United Kingdom, perhaps because we do not have a sufficient number of judges at the appropriate level who are not male and white to make such statistical analysis worthwhile'. Still, researchers have applied quantitative methods to datasets of case law from, for example, Belgium (De Jaeger 2017), the Czech Republic (Bricker 2017), France (Sulea et al. 2017a), Germany (Dyevre 2015; Bricker 2017), Israel (Doron et al. 2015), Latvia (Bricker 2017), the Netherlands (Vols et al. 2015; Vols and Jacobs 2017; van Dijck 2018; Bruijn et al. 2018), Poland (Bricker 2017), Slovenia (Bricker 2017), Spain (Garoupa et al. 2012) and Sweden (Derlén and Lindholm 2018).

Besides that, a growing body of research exists on quantitative analysis of case law of international courts. For example, Behn and Langford (2017) manually collected and coded roughly 800 cases on Investment Treaty Arbitration. Others have applied quantitative methods in the analysis of case law of the International Criminal Court (Holá et al. 2012; Tarissan and Nollez-Goldbach 2014, 2015), the Court of Justice of the European Union (Lindholm and Derlén 2012; Derlén and Lindholm 2014; Tarissan and Nollez-Goldbach 2016; Derlén and Lindholm 2017a, b; Frankenreiter 2017a, b; Zhang et al. 2017) or the European Court of Human Rights (Bruinsma and De Blois 1997; Bruinsma 2007; White and Boussiakou 2009; Christensen et al. 2016; Olsen and Küçüksu 2017; Madsen 2018).

In most research projects, case law is manually collected and hand-coded. Nevertheless, a number of researchers use computerized techniques to collect case law and automatically generate usable information from the collected case law (see Trompper and Winkels 2016; Livermore et al. 2017; Shulayeva et al. 2017; Law 2017). For example, Dyevre (2015) discusses the use of automated content analysis techniques in the legal discipline, using such tools as Wordscores[7] and Wordfish,[8] which are traditionally used to automatically extract political positions using word frequencies in text documents. The author applied these two techniques to analyse a (relatively small) dataset of 16 judgements of the German Federal Constitutional Court on European integration. He found that both Wordscore and Wordfish are able to generate judicial position estimates that are remarkably reliable when compared with the accounts appearing in legal scholarship. Christensen et al. (2016) used a quantitative network analysis to automatically identify the content of cases of the ECtHR. They exploited the network structure induced by the citations to automatically infer the content of a court judgement. Panagis et al. (2016) used topic modelling techniques to automatically find latent topics in a set of judgements of the Court of Justice of the European Union (CJEU) and the ECtHR. Derlén and Lindholm (2017a) used computer scripts to extract information concerning citations in CJEU case law.

A large number of studies (especially outside the USA) present basic descriptive statistics of manually collected and coded case law (e.g., Bruinsma and De Blois 1997; White and Boussiakou 2009; De Jaeger 2017; Madsen 2018; Vols and Jacobs 2017). Other studies present results of relatively basic statistical tests such as correlation analysis (e.g., Doron et al. 2015; Evans et al. 2017; Bruijn et al. 2018). A growing body of papers present results of more sophisticated statistical analyses, including regression analysis of case law (see Dhami and Belton 2016). Most of these papers focus on case law from the USA (see Chien 2011; Epstein et al. 2013), but researchers outside of the USA have conducted such analyses as well (Holá et al. 2012; Behn and Langford 2017; Bricker 2017; van Dijck 2018; Zhang et al. 2017; Frankenreiter 2017a, 2018).

A growing body of research presents the results of citation analysis of case law of courts in the USA (see Whalen 2016; Matthews 2017; Shulayeva et al. 2017; Frankenreiter 2018), where they analyze patterns of citations within case law documents,

[7] http://www.tcd.ie/Political_Science/wordscores/.

[8] http://www.wordfish.org/.

their number and impact. Other scholars have applied this method to case law from European countries, such as Sweden (Derlén and Lindholm 2018). Researchers have also used this method to analyse case law of international courts. Some have performed a citation analysis of the case law of the CJEU (Lindholm and Derlén 2012; Derlén and Lindholm 2014; Tarissan and Nollez-Goldbach 2016; Derlén and Lindholm 2017a, b; Frankenreiter 2017a, b, 2018). Derlén and Lindholm (2017a, p. 260) use this method to compare the precedential and persuasive power of key decisions of the CJEU using different centrality measurements. A number of studies investigated citation network analyses of case law of the ECtHR (Lupu and Voeten 2012; Christensen et al. 2016; Olsen and Küçüksu 2017). Olsen and Küçüksu (2017, p. 19) hold that citation network analysis enables researchers to more easily note the emergence and establishment of patterns in case law that would otherwise have been difficult to identify. Some researchers have combined citation network analysis of case law of both European courts in one study (Šadl and Olsen 2017). Other papers used this method to analyse case law of the International Criminal Court (Tarissan and Nollez-Goldbach 2014, 2015, 2016).

There are studies that focus specifically on text mining arguments of legal cases (among others, Mochales and Moens 2008; Wyner et al. 2010). Being able to identify arguments is essential for automatic analysis of legal data and can be used for predicting court decisions. However, this is a very hard task, and the majority of known approaches to solving it require a large amount of manually annotated data. As we do not have this type of data, we use a data-driven approach which does not use argument mining. Instead, we use as much unprocessed data as possible and build a system that predicts the decisions of the Court and then try to derive what the basis is of this prediction. Similarly, there is number of studies that use the arguments of the court and the decisions to identify the verdict (among others Grabmair 2017; Ruppert et al. 2018; Waltl et al. 2017). Such approaches can be used, for instance, for sorting already published judgements or extracting the verdict out of unstructured legal texts. By contrast, our interest is focused on using the information that is available *before* the court rules on the case, thereby excluding the parts of the judgements that contain the arguments or decisions.

A relatively small number of studies have used machine learning techniques—which we use in this study as well (and is explained in more detail in the following section—to analyse case law (see Evans et al. 2007; Custers and Leeuw 2017; Ashley and Brüninghaus 2009). Again, researchers in the United States were the first to use this technique to predict the courts' decisions or voting behaviour of judges (Katz 2012; Wongchaisuwat et al. 2017). Recently, Katz et al. (2017) developed a prediction model that aims to predict whether the US Supreme Court as a whole affirms or reverses the status quo judgement, and whether each individual Justice of the Supreme Court will vote to affirm or reverse the status quo judgement. Their model achieved an accuracy of 70.2% at the case outcome level and 71.9% at the justice vote level. Outside of the United States, only a few scholars have used machine learning techniques to predict the courts' decisions. Sulea et al. (2017b) used machine learning techniques to analyse case law of the French Court of Cassation. They aimed to predict the law area of a case and the court ruling. Their model achieved an accuracy of over 92%. Aletras et al. (2016) used machine

learning techniques to predict the decisions of the ECtHR. Their model aims to predict the court's decision by extracting the available textual information from relevant sections of the ECtHR judgements. They derived two types of textual features from the texts, N-gram features (i.e. contiguous word sequences) and word clusters (i.e. abstract semantic topics). Their model achieved an accuracy of 79% accuracy at the case outcome level.

As we focus on the European Court of Human Rights, the study of Aletras et al. (2016) is most relevant for our work. However, in their work they used only a limited number of cases, and due to the unavailability of the application numbers of the cases that they used for their predictions, we were unable to reproduce their results. When using their methods with the same and larger amount of data, however, we consistently achieved lower results than was reported in their paper. Therefore, we start our research by using similar methods using all of the available data and exploring how we can gradually improve on them.

3 Machine learning for legal text classification

There are many possible ways for processing case law, and even though many steps have been taken towards systematisation of the data and automatising the processes, the amount of choices one can make is daunting. Therefore, in this section we discuss one way of automatically processing legal texts.

Legal information of any sort is largely written in natural, although rather specific language. For the most part this information is relatively unstructured. Consequently, to process legal big data automatically we need to use techniques developed in the field of natural language processing.

The goal of this study is to create a system, which is able to automatically predict the category (i.e. a verdict) associated to a new element (i.e. a case). For this task we will employ machine learning. More specifically, we will use *supervised* machine learning. In this type of approach the computer is provided with (textual) information from many court cases together with the actual judgements. By providing many of these examples (in the so-called 'training phase'), the computer is able to identify patterns which are associated with each class of verdict (i.e. violation vs. no violation). To evaluate the performance of the machine learning program, it is provided with a case without the judgement (in the 'testing phase') for which it has to provide the most likely judgement. To make this judgement (also called: 'classification') the program uses the information it identified to be important during the training phase.

To illustrate how supervised machine learning works, let's imagine a non-textual example. Suppose we want to write a program that recognises pictures of cats and dogs. For that we need a database of images of cats and dogs, where each image has a label: either *cat* or *dog*. Then we show the system those pictures with labels one by one. If we show enough pictures, eventually the program starts recognising various characteristics of each animal, e.g., cats have long tails, dogs are generally more furry. This process is called *training* or *fitting the model*. Once the program *learns* this information, we can show it a picture without a label and it will guess which *class* the picture belongs to.

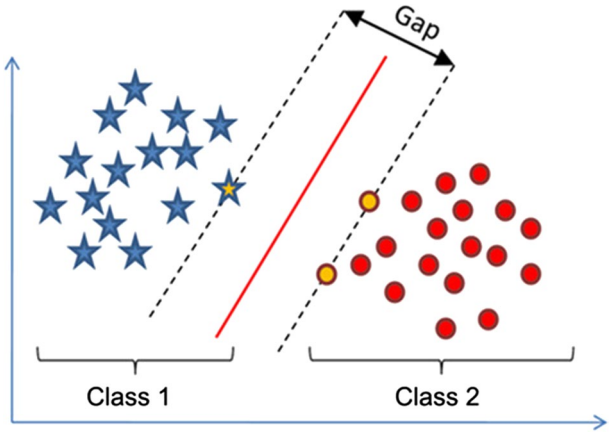

Fig. 1 Illustration of an SVM dividing data into classes. (*Source*: http://diggdata.in/post/940665449 71/support-vector-machine-without-tears)

Very similar experiments can be conducted with text. For instance, when categorising texts into the ones written by men and the ones written by women, the program can analyse the text and the style it was written in. Research conducted on social media data shows that when training such models, we can see that men and women generally talk about different things. For example, women use more pronouns than men (Rangel and Rosso 2013), while men swear more often (Schwartz et al. 2013).

For the present study, we wrote a computer program that analyses texts of judgements of ECtHR cases available on the Court's website[9] and predicts whether any particular article of ECHR was violated.

As we have mentioned in Sect. 2, techniques from machine learning have not often been used in the legal domain. Nevertheless, the data that we have is well-suited for automatic text classification. We have a very large amount of semi-structured cases (which are almost impossible to process manually) that we can roughly split into facts, arguments and decisions. By providing the machine learning program with the facts, we may predict the decisions (i.e. the label).

For this task we use a particular approach (i.e. an algorithm) used in machine learning called a Support Vector Machine (SVM) Linear Classifier. It sorts data based on labels provided in the dataset (i.e. the *training data*) and then tries to establish the simplest equation that would separate different data points from each other according to the labels with the least amount of error.

We can see an example of how the system works in Fig. 1. The algorithm decides on the best hyperplane (i.e. a line in multiple dimensions) to separate the data. In the figure this is the middle line separating the stars and the circles. The support vectors are the data points nearest to this line. The goal of the SVM algorithm is to choose the position

[9] https://hudoc.echr.coe.int/.

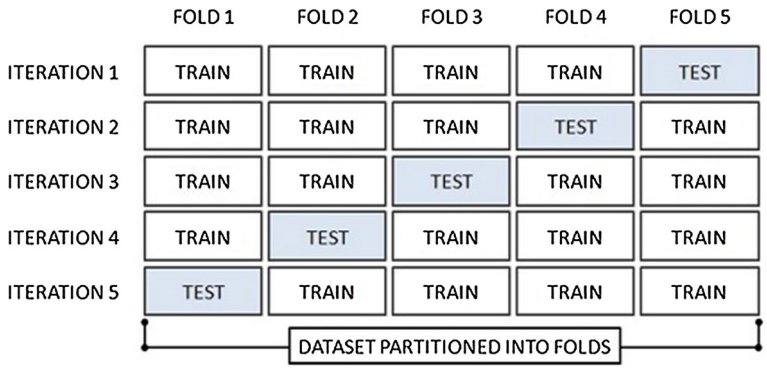

Fig. 2 Example of fivefold cross-validation. (*Source*: http://www.dummies.com/programming/big-data/data-science/resorting-cross-validation-machine-learning/)

of the hyperplane in such a way that the largest possible margin with respect to the points is achieved. This allows for a greater chance of classifying new data correctly.

After training an SVM, we use a separate set of cases that have not been used during training (*test set*) to evaluate the performance of the machine learning approach. We let the program indicate for every case whether it classifies it as a violation or not, and then compare this decision to the actual decision of the court. We then measure the performance of the system as the percentage of correctly identified decisions. We will discuss the choice of cases for the test set in the next section.

Another way to evaluate the performance is by using *k-fold cross-validation*. For that, we take all the data that we have available for the model to learn characteristics of the cases, and we split this set into k parts. then we take one part out and train the model using the remaining part of this set. Once the model is trained we evaluate it by obtaining the decisions of the program on the basis of the cases in the withheld part. Then we repeat this procedure using another part of the data (i.e. leave it out and train on the rest, evaluate on the withheld part), etc. We repeat this k times until we evaluated the model using each of the k withheld parts. For instance, if $k = 5$ we will perform fivefold cross-validation and train and test the model 5 times (see Fig. 2). Each time the withheld part consists of 20% (1/5th) and the training phase is done using the remaining 80% of the data.

Using cross-validation allows us to determine the optimal parameters of the machine learning system, as well as evaluating if it performs well when being evaluated using different samples of data. In this way, the model is more likely to perform better for unseen cases.

4 Data

4.1 Collecting the data

As we will build on the results by Aletras et al. (2016), we use the publicly available data published by the ECtHR. In order to understand the data we are going to be working on it is important to have some idea about the composition of the court itself and the structure of its documents.

The European Court of Human Rights is an international court that was established in 1959. It deals with individual and State applications that claim violation of various right laid out in the European Convention on Human Rights. Applications are always brought against a State or multiple States that have ratified the ECHR, but not against individuals.

The number of judges in the Court is equal to the number of States Parties to the Convention, which at the moment of publication of this paper is 47. The judges are currently elected for 9-year terms with no possibility of re-election. The cases are tried in Sections that contain 7-member Chambers or in a 17-member Grand Chamber, where the judge from the State that is being accused of the violation (the 'national judge') is always present.

In order to be tried by a Chamber the case has to pass the admissibility stage by a single judge (who is not the 'national judge'). If the case is found admissible, it will be judged based on merit by a Chamber within one of the 5 Sections of the Court or, in exceptional circumstances, by the Grand Chamber.

The rulings of the Court are available online and have a relatively consistent structure.

A judicial decision of the ECtHR contains the following main parts:

- *Introduction*, consisting of the title (e.g., Lawless vs. Ireland), date, Chamber, Section of the Court and its constitution (i.e. judges, president, registrar);
- *Procedure*, containing the procedure that took place from lodging and application until the judgement by the Court;
- *Facts*, consisting of 2 parts:
  - *Circumstances*, containing a relevant background information on the applicant and events and circumstances that led them to seek justice due to alleged violation of their rights in accordance to ECHR;
  - *Relevant Law*, containing relevant provisions from legal documents other than the ECHR (these are typically domestic laws, as well as European and international treaties);
- *Law*, containing legal arguments of the Court with each alleged violation discussed separately;
- *Judgement*, containing the decision of the Court per alleged violation;
- *Dissenting/Concurring Opinions*, containing the additional opinions of judges, explaining why they voted with the majority (concurring opinion) or why they did not agree with the majority (dissenting opinion).

Table 1 Initial distribution of cases (in English) Obtained from HUDOC on September 11, 2017

Article	Title	'Violation' cases	'Non-violation' cases
2	Right to life	559	161
3	Prohibition of torture	1446	595
4	Prohibition of slavery and forced labour	7	10
5	Right to liberty and security	1511	393
6	Right to a fair trial	4828	736
7	No punishment without law	35	47
8	Right to respect for private and family life	854	358
9	Freedom of thought, conscience and religion	65	31
10	Freedom of expression	394	142
11	Freedom of assembly and association	131	42
12	Right to marry	9	8
13	Right to an effective remedy	1230	170
14	Prohibition of discrimination	195	239
18	Limitation on use of restrictions on rights	7	32

Frequently, there is no section about the dissenting or concurring opinions, but the other parts are typically included. They can be of different length and detail level.

In order to create a database which we can use for our experiments, we had to automatically collect all data online. We therefore created a program that automatically downloaded all documents in English from the HUDOC website.[10] Our database[11] contains all texts of admissible cases available on HUDOC as of September 11, 2017. Cases which were only available in French or another language were excluded. We used a rather crude automatic extraction method, so it is possible that a few cases might be missing from our dataset. However, this does not matter, given that we have extracted a large enough sample. For reproducibility, all of the documents that we obtained are available online together with the code we used to process the data.

In this study, our goal was to predict whether there were any violations of each article of the European Convention on Human Rights separately. We therefore created separate data collections with cases that involved specific articles, and whether or not the court ruled that there was a violation. As many of the cases consider multiple violations at once, some of the cases appear in multiple collections. The information about a case being a violation of the specific article or not was automatically extracted from the metadata on the HUDOC website.

From the data (see Table 1) we can see that most of the admissible cases considered by the European Court of Human Rights result in a decision of 'violation'

[10] https://hudoc.echr.coe.int/.

[11] https://www.dropbox.com/s/lxpvvqdwby30157/crystal_ball_data.tar.gz.

Table 2 Final number of cases per Article of ECHR

Article	'Violation' cases	'Non-violation' cases	Total	Test set
Article 2	57	57	114	398
Article 3	284	284	568	851
Article 5	150	150	300	1118
Article 6	458	458	916	4092
Article 8	229	229	458	496
Article 10	106	106	212	252
Article 11	32	32	64	89
Article 13	106	106	212	1060
Article 14	144	144	288	44

by the state. The specific distribution, however, depends on the article that is being considered.

4.2 Balanced dataset

The machine learning algorithm we use learns characteristics of the cases based on the text it is presented with as input. The European Court of Human Rights often considers multiple complaints within one case, even though they might be related to the same article of the ECHR. However, we conduct this experiment as a binary task only predicting two possible decisions: 'violation' of an article and 'non-violation' of the article. While some cases may have both decisions for one article if there are multiple offences, here we only focus on cases in which there is a single ruling ('violation' or 'non-violation'). We do this to obtain a clearer picture of what influences the two separate decisions of the Court.

While excluding cases which have both decisions makes the task more limited, the goal of our study is to determine the patterns that are specific for a 'violation' or 'no violation' of a particular article of the Convention. Limiting our task helps us obtain a clear picture.

Crudely speaking, until a certain amount, the more data is available for the training phase, the better the program will perform. However, it is important to control what sort of information it learns. If we blindly provide it with all the cases, it might only learn the distribution of 'violation'/'non-violation' cases rather than more specific characteristics. For example, we might want to train a program that predicts whether there is a violation of Article 13, and feed it all 170 'non-violation' cases together with all 1230 'violation' cases. With such a clear imbalance in the number of cases per type, it is likely that the program will learn that most of the cases have a violation and then simply predict 'violation' for every new case (the performance will be quite high: 88% correct). In order to avoid this problem, we instead create a *balanced* dataset by including the same number of violation cases as the number of non-violation cases. We randomly removed the violation cases such that the distribution of both classes was balanced (i.e. 170 violation cases vs. 170 non-violation cases). The excluded violation cases were subsequently used to test the system.

We decided to withhold 20% of the data in order to use it in future research (i.e. as a test set after several different systems have been developed, one of which is discussed in this paper). These cases were randomly selected and removed from the dataset. These missing cases are available online.[12]

The results of the present study are evaluated using the violation cases that were not used for training the system. The number of cases can be found in Table 2 (test set). Only for Article 14, there were more 'non-violation' cases than 'violation' cases. Consequently, here the test set consists of 'non-violation' cases.

For example, for Article 2 we had 559 cases with 'violation' and 161 'non-violation'. 90 of these cases had both at the same time. After removing those we are left with 469 cases with only 'violation' and 71 'non-violation'. We want to have the same amount of cases with each verdict, so we have to reduce the amount of cases with 'violation' to 71 as well, leaving us with 142 cases in total and a test set of 398 'violation' cases for Article 2. Then we removed 20% of the cases (14 'violation' cases and 14 'non-violation'), leaving us with 114 cases for the training phase.

A machine learning algorithm requires a substantial amount of data. For this reason, we excluded articles with too few cases. We included only articles with at least 100 cases, but also included Article 11 as an estimate of how well the model performs when only very few cases are available. The final distribution of cases can be seen in Table 2.

5 Experiments

In this Section we describe the experiments that we conducted in this study. In Experiment 1 we investigate the possibilities of using words and phrases in the text of the cases in order to predict judicial decisions. In Experiment 2 we use the approaches from the first experiment in order to estimate the potential of predicting *future* cases. In Experiment 3 we test if we can make predictions based solely on objective (although limited) information. Specifically, we will evaluate how well we are able to predict the court's judgements only using the surnames of the judges involved.

5.1 Experiment 1: textual analysis

5.1.1 Set-up

The data we provided to the machine learning program does not include the entire text of the court decision. Specifically, we have removed decisions and dissenting/concurring opinions from the texts of the judgements. We have also removed the *Law* part of the judgement as it includes arguments and discussions of the judges

[12] See test20 at https://www.dropbox.com/s/lxpvvqdwby30157/crystal_ball_data.tar.gz.

that partly contain the final decisions. See, for instance, the statement from the Case of Palau-Martinez v. France (16 December 2003):

> 50. The Court has found a violation of Articles 8 and 14, taken together, on account of discrimination suffered by the applicant in the context of interference with the right to respect for their family life.

From this sentence it is clear that in this case the Court ruled for a violation of Articles 8 and 14. Consequently, if we let our program predict the decision based on this information, it will be unfair as the text already shows the decision ('found a violation'). Moreover, the discussions that the *Law* part contains are not available to the parties before the trial and therefore, predicting the judgement on the basis of this information is not very useful. Other information we have removed, is the information in the beginning of the case description which contains the names of the judges. We will, however, use this data in Experiment 3. The data we used can be grouped in five parts: *Procedure*, *Circumstances*, *Relevant Law*, the latter two together (*Facts*) and all three together (*Procedure* + *Facts*).

The important characteristic of the aforementioned parts is that they are available years before the decision is made. The Facts part of the case become available after the case is ruled to be admissible. One might argue that because the cases available on HUDOC were written after the trial, the information that was found irrelevant and was dismissed during the trial might not appear in the text. However, the Facts of the case are available and remain unchanged several years *before* the final judgement is made (i.e. after the application is found admissible based on the formal criteria). As part of the procedure, the Court 'communicates' the application to the government concerned and lists the facts stated by the applicant, in order for the government to be able to respond to the accusation. The Procedure part lists the information regarding what happened before the case was presented to the Court and thus is also available before the decision is made. The rest of the text of the case is only available when the final judgement is made by the Court, and thus should not be used to predict the decisions in advance.

Until now we have ignored one important detail, namely how the text of a case is represented for the machine learning program. For this we need to define features (i.e. an observable characteristic) of each case. In terms of the cats-and-dogs example, features of each picture would be the length of a tail as a proportion of the total body length, being furry or not, the number of legs, etc. The machine learning program then will determine which features are the most important for classification. For the cats-and-dogs example, the relative tail length and furriness will turn out to be important features in distinguishing between the two categories, whereas having four legs will not be important. An essential question then becomes how to identify useful features (and their values for each case). While it is possible to use manually created features, such as particular types of issues that were raised in the case, we may also use automatically selected features, such as those which simply contain all separate words, or short consecutive sequences of words. The machine learning program will then determine which of these words or word sequences are most characteristic for either a violation or a non-violation. A contiguous sequence of one or more words in a text is formally called a word

n-gram. Single words are called unigrams, sequences of two words are bigrams, and sequences of three consecutive words are called trigrams.

For example, consider the following sentence:

> By a decision of 4 March 2003 the Chamber declared this application admissible.

If we split this sentence into bigrams (i.e. 2 consecutive words) the extracted features consist of:

> By a, a decision, decision of, of 4, 4 March, March 2003, 2003 the, the Chamber, Chamber declared, declared this, this application, application admissible, admissible.

Note that punctuation (e.g., a point at the end of the sentence) is also interpreted as being a word. For trigrams, the features consist of:

> By a decision, a decision of, decision of 4, of 4 March, 4 March 2003, March 2003 the, 2003 the Chamber, the Chamber declared, Chamber declared this, declared this application, this application admissible, application admissible.

While we now have shown which features can be automatically extracted, we need to decide what values are associated with these features for each separate case. A very simple approach would be taking all n-grams from all cases and using a binary feature value: 1 if the n-gram is present in the case description and 0 if it is not. But of course, we then throw away useful information, such as the frequency with which an n-gram occurs in a document. While using the frequency as a feature value is certainly an improvement (e.g., 'By a': 100, '4 March': 1, 'never in': 0) some words simply are more common and therefore used more frequently than other words, despite these words not being characteristic for the document at all. For example, the unigram 'the' will occur much more frequently than the word 'application'. In order to correct for this, a general approach is to normalise the absolute n-gram frequency by taking into account the number of documents (i.e. cases) in which each word occurs. The underlying idea is that characteristic words of a certain case will only occur in a few cases, whereas common, uncharacteristic words will occur in many cases. This normalized measure is called *term frequency-inverse document frequency* (or *tf-idf*). We use the formula defined by `scikit-learn` Python package (Pedregosa et al. 2011): $\text{tfidf}(d,t) = \text{tf}(t) * \text{idf}(d,t)$, where $\text{idf}(d,t) = \log(n/\text{df}(d,t)) + 1$, n is the total number of documents, and $\text{df}(d,t)$ is the document frequency. Document frequency is the number of documents d that contain term t. In our case the terms are n-grams. So, let's say we have a document of a 1000 words containing the word *torture* 3 times, the term frequency (i.e. *tf*) for *torture* is $(3/1000) = 0.003$. Now let's say we have 10,000 documents (cases) and the word *torture* appears in 10 documentsm. Then the inverse document frequency (i.e. *idf*) is $\log(10,000/10) + 1 = 4$. The resulting tf-idf score is $0.003 * 4 = 0.012$. This is the score (i.e. weight) assigned to the word *torture*. Note that this score is higher than using the *tf* score, as it reflects that the word does not occur often in other documents.

Table 3 List of evaluated parameter values

Name	Values	Description
ngram_range	(1, 1), (1, 2), (1, 3), (1, 4), (2, 2), (2, 3), (2, 4), (3, 3), (3, 4), (4, 4)	Length of the n-grams; e.g., (2, 4) contains bigrams, 3-grams and 4-grams
lowercase	True, False	Lowercase all words (remove capitalization for all words)
min_df	1, 2, 3	Exclude terms that appear in fewer than *n* documents
use_idf	True, False	Use inverse document frequency weighting
binary	True, False	Set term frequency to binary (all non-zero terms are set to 1)
norm	None, 'l1', 'l2'	Norm used to normalize term vectors[a]
stop_words	None, 'English'	Remove most frequent words in English language from documents. *None* to keep all words
C	0.1, 1, 5	Penalty parameter for the SVM[b]

[a]We can use normalization to account for the bias towards high frequencies of certain words as well as the length of the texts. For more information on the differences between L1- and L2-norms see http://blog.christianperone.com/2011/10/machine-learning-text-feature-extraction-tf-idf-part-ii/

[b]C-parameter determines the trade-off between training error and model complexity. If C is too small, it will increase the number of training errors, while a large C might lead to a model that cannot generalize and is thus unable to predict well the decisions of the cases it has never seen before (Joachims 2002)

Table 4 Selected parameters used for the best model

Article	Parts	N-grams	Remove capitalisation	Remove stopwords
Article 2	Procedure + facts	3–4	✓	✗
Article 3	Facts	1	✓	✗
Article 5	Facts	1	✓	✗
Article 6	Procedure + facts	2–4	✓	✗
Article 8	Procedure + facts	3	✓	✗
Article 10	Procedure + facts	1	✗	✗
Article 11	Procedure	1	✗	✓
Article 13	Facts	1–2	✗	✗
Article 14	Procedure + facts	1	✓	✓

In order to identify which sets of features we should include (e.g., only unigrams, only bigrams, only trigrams, a combination of these, or even longer n-grams) we evaluate all possible combinations. It is important to realise that longer n-grams are less likely to occur (e.g., it is unlikely that one full sentence occurs in multiple case descriptions) and therefore are less useful to include. For this reason we limit the maximum word sequence (i.e. n-gram length) to 4.

However, there are also other choices to make (i.e. parameters to set), such as if all words should be converted to lowercase, or if the capitalisation is important. For these parameters we take a similar approach and evaluate all possible combinations. All parameters we have evaluated are listed in Table 3.[13] Because we had to evaluate all possible combinations, there were a total of 4320 different possibilities to evaluate. As indicated above, cross-validation is a useful technique to assess (only on the basis of the training data) which parameters are best. To limit the computation time, we only used threefold cross-validation for each article. The program therefore trained 12,960 models. Given that we trained separate models for 5 parts of the case descriptions (*Facts*, *Circumstances*, etc.), the total number of models was 64,800 for each article and 583,200 models for all 9 articles of the ECHR. Of course we did not run all these programs manually, but rather created a computer program to conduct this so-called grid-search automatically. The best combination of parameters for each article was used to evaluate the final performance (on the test set). Table 4 shows the best settings for each article.[14]

During this type of search, we identify which parameter setting performs best by testing each combination of parameters a total of 3 times (using random splits to determine the data used for training and testing) and selecting the parameter setting which achieves the highest average performance. We use this approach to make sure

[13] For more detailed description of the parameters see http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html and http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html.

[14] The choice of all parameters per article can be found online: https://github.com/masha-medvedeva/ECtHR_crystal_ball.

Table 5 Cross-validation (tenfold) and test results for Experiment 1

	Art 2	Art 3	Art 5	Art 6	Art 8	Art 10	Art 11	Art 13	Art 14	Average
Cross-val	0.73	0.80	0.71	0.80	0.72	0.61	0.83	0.83	0.75	0.75
Test	0.82	0.81	0.75	0.75	0.65	0.52	0.66	0.82	0.84	0.74

that we did not just get 'lucky', but that overall the model performs well. Of course, it is still possible that the model performs worse (or better) on the test data.

For most articles unigrams achieved the highest results, but for some longer n-grams were better. As we already expected, the *Facts* section of the case was the most informative and selected for 8 out of 9 articles. For many articles the *Procedure* section was also informative. This is not surprising, as *Procedure* contains important information on the alleged violations. See, for instance, a fragment from the procedure part of the Case of Abubakarova and Midalishova v. Russia (4 April 2017):

> 3. The applicants alleged that on 30 September 2002 their husbands had been killed by military servicemen in Chechnya and that the authorities had failed to investigate the matter effectively.

5.1.2 Results

After investigating which combinations of parameters worked best, we used these parameter settings together with tenfold cross-validation to ensure that the model performed well in general and was not overly sensitive to the specific set of cases on which it was trained. When performing tenfold cross-validation instead of threefold cross-validation, there is more data available to use for training in each fold (i.e. 90% rather than 66.7%). The results can be found under 'cross-val' in Table 5. Note that as we used a balanced dataset (both for the cross-validation and the test set), the number of 'violation' cases is equal to the number of 'non-violation' cases. Consequently, if we would just randomly guess the outcome, we would be correct in about 50% of the cases. Percentages substantially higher than 50% indicate that the model is able to use (simplified) textual information present in the case to improve the prediction of the outcome of a case. Table 5 shows the results for the cross-validation in the first row.

In order to evaluate if the model predicts both classes similarly, we use precision, recall and f-score to estimate the performance. *Precision* is the percentage of cases for which the assigned label is correct (i.e., 'violation' or 'no violation'). *Recall* is the percentage of cases having a certain label which are identified correctly. The F-score can be described as the harmonic mean of precision and recall.[15] Table 6 shows these measures per class for each model.

[15] The exact description of the metric used can be found here: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

Table 6 Precision, recall and f-score per class per article achieved during tenfold cross-validation

Art #	Class	Precision	Recall	F-score
Art 2	Non-violation	0.72	0.68	0.70
Art 2	Violation	0.70	0.74	0.72
Art 3	Non-violation	0.80	0.77	0.79
Art 3	Violation	0.78	0.81	0.80
Art 5	Non-violation	0.77	0.75	0.76
Art 5	Violation	0.76	0.77	0.77
Art 6	Non-violation	0.78	0.87	0.82
Art 6	Violation	0.85	0.76	0.80
Art 8	Non-violation	0.69	0.76	0.72
Art 8	Violation	0.73	0.66	0.69
Art 10	Non-violation	0.63	0.66	0.65
Art 10	Violation	0.64	0.61	0.63
Art 11	Non-violation	0.86	0.78	0.82
Art 11	Violation	0.80	0.88	0.8
Art 13	Non-violation	0.83	0.86	0.85
Art 13	Violation	0.85	0.83	0.84
Art 14	Non-violation	0.77	0.76	0.77
Art 14	Violation	0.77	0.77	0.77

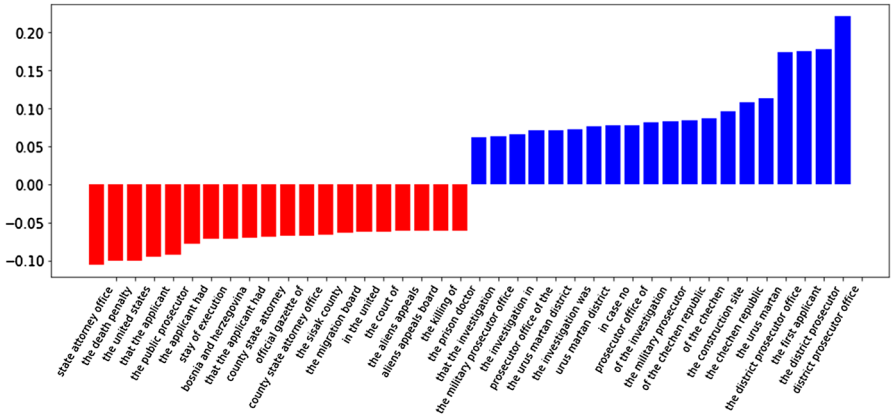

Fig. 3 Coefficients (weights) assigned to different n-grams for predicting violations of Article 2 of ECHR. Top 20 'violation' predictors (blue on the right) and top 20 'non-violation' predictions (red on the left). (Color figure online)

As we can see from the table, 'violation' and 'non-violation' are predicted very similarly for each article.

During the training phase, different weights are assigned to the bits of information it is given (i.e. n-grams) and a hyperplane is created which uses support vectors to maximise the distance between the two classes. After training the model, we may inspect the weights in order to see what information had the most impact

Table 7 Confusion matrix for tenfold cross-validation for Article 6

	Actual: non-violation	Actual: violation
Predicted: non-violation	397	112
Predicted: violation	61	346

on the model's decision to predict a certain ruling. The weights represent the coordinates of the data points, such as stars and circles we've seen in Fig. 1. The further the data point is from the hyperplane, the more positive the weight is for the violation class or the more negative the weight is for the non-violation class. These weights can then be used to determine how important the n-gram was for the separation. The n-grams that were most important hopefully may yield some insight into what might influence Court's decision making. The idea behind the approach is that we will be able to determine certain keywords and phrases that are indicative of specific violations. For instance, if a case mentions a minority group or children, then based on the previous cases with the same keywords, the machine learning algorithm will be able to determine the verdict better.

In Fig. 3 we visualize the phrases (i.e. 3 and 4-grams) that ranked the highest to identify a case as a 'violation' blue on the right) or a 'non-violation' (red on the left) for Article 2. If we look at the figure, we can notice for instance that the Chechen Republic is an important feature in relation to 'violation' cases, while Bosnia and Herzegovina has a higher weight on the 'non-violation' side.

We also observe many relatively 'meaningless' n-grams, like *in case no* or *the court of*. This can be explained by the simplicity of our model and the lack of filtering of unnecessary information. As we have mentioned, the only processing of the text that we do is lowering the case of the words for some articles and filtering out some of the more common words such as *the*, *a*, *these*, *on*, etc. for some of the articles.

After tuning and evaluating the results using cross-validation we have tested the system using the 'violation' cases (for Article 14: non-violation cases) that have not been used in determining the best system parameters. These are the 'excessive' cases The results can be found in Table 5 (second results line). For several articles (6, 8, 10, 11, 13) the performance on the test set was worse, but for a few it performed quite a lot better (e.g., Articles 2, 5, 14). Discrepancies in the results may be explained by the fact that sometimes the model learns to predict non-violation cases better than violation cases. By testing the system on cases that only contain violations, performance may seem to be worse. The opposite happens when the model learns to predict violations better. In that case, the results on this violation-only test set become higher. Note that the test set for Article 14 contains non-violations only, and an increase in the performance here indicates that the model has probably learned to predict non-violations better. Nevertheless, the test results overall seem to be relatively similar to the cross-validation results, suggesting the models are well-performing, despite having only used very simple textual features.

Table 8 Comparison of selected n-grams within top-100 tf-idf scores for correctly predicted and mislabelled documents for Article 13

	Actual: non-violation	Actual: violation
Predicted: non-violation	Applicant	
	Police	**Police**
	Security	**Security**
	Commission	
	Imprisonment	Prison
Predicted: violation	**Prosecutor**	**Prosecutor**
	Criminal	**Criminal**
		Military
	Ukraine	**Ukraine**
		Russian

Bold is used to highlight the words that are the the same in two columns to indicate how the machine learning model may have made the error in classifying the document

5.1.3 Discussion

The results, with an average performance of 0.75, show substantial variability across articles. It is likely that the differences are to a large extent caused by differences in the amount of training data. The lower the amount of training data, the less the model is able to learn from the data.

To analyse how well the model performs, it is useful to investigate the *confusion matrix*. This matrix shows in what way the cases were classified correctly and incorrectly. For example, Table 7 shows the confusion matrix for Article 6. There were 916 cases in the training set for Article 6, half of which (458 cases) had a violation verdict, and the other half a non-violation verdict. In the table we see that from 458 cases with a non-violation, 397 were identified correctly, and 61 were identified as cases with a violation. Additionally, 346 cases with a violation were classified correctly and 112 cases were identified as a non-violation. Given that the amount of non-violation and violation cases is equal, it is clear from this matrix that the system for Article 6 is better at predicting non-violation cases than violation cases, as we could also see in Table 6.

Cases themselves may also influence the results. If there are many similar cases with similar decisions, it is easier to predict the judgement of another similar case. Whenever there are several very diverse issues grouped under a single article of the ECHR, the performance is likely lower. This is likely the cause of the relatively low performance of Article 8 (*Right to respect for private and family life*), which covers a large range of cases. The same can be said for Article 10 (*right to freedom of expression*), as the platforms for expression are growing in variety, especially since the spreading of the Internet.

To investigate the prediction errors made by each system, we focus on Article 13 having the highest accuracy score. Our approach was to first list the n-grams having the top-100 tf-idf scores for the incorrectly classified documents (separately for a violation classified as a non-violation and vice versa). We then only included the

Table 9 Number of cases

	Art. 3	Art. 6	Art. 8
Train set	356	746	350
2014–2015	72	80	52
2016–2017	140	90	56

n-grams occurring in at least three incorrectly classified documents for each of the two types. For the correctly identified documents, we did the same (also two types: violation correctly classified and non-violation correctly classified) and then looked at the overlap of the n-grams in the four lists. While those lists contained very different words and phrases, we were also able to observe some general tendencies. For instance, phrases related to prison (e.g., 'the prison', 'prisoner', etc.) generally appeared in cases with no violation. Consequently cases with violation which do contain these words are likely to be incorrectly classified (as non-violation). Similarly, words related to prosecutors (e.g.,'public prosecutor', 'military prosecutor', 'the prosecutor') are found more often within cases with a violation and therefore non-violation cases with such phrases may be mislabelled. Table 8 shows a subjective selection of similar-behaving words.

Note that the error analysis remains rather speculative. It is impossible to pinpoint what exactly makes the largest impact on the prediction, as the decision for a document is based on all n-grams in the document.

In the future more sophisticated methods that include semantic analysis should be used in order to not just predict the decisions, but to identify the factors behind the choice that the machine learning algorithm makes.

5.2 Experiment 2: predicting the future

5.2.1 Set-up

In the first experiment the test set was simply random sampled without considering the year of the cases. In this section we will assess how well we are able to predict future cases, by dividing the cases used for training and testing on the basis of the year of the case. Such an approach has two advantages. The first is that this would result in a more realistic setting, as there is no practical use of predicting the outcome of a case for which the actual outcome is already known. The second advantage is that times are changing, and this affects the law as well. For example, consider Article 8 of the ECHR. Its role is to protect citizens' private life which includes, for instance, their correspondence. However, correspondence 40 years ago looked very different from that of today. This also suggests that using cases from the past to predict the outcome of cases in the future might reflect a lower, but a more realistic performance than the results reported in Table 5. For this reason, we have set up an additional experiment to check whether this is indeed the case and how sensitive our system is to this change. Due to more specific requirements of the data for this experiment, we have only considered datasets with the largest amounts of

Table 10 Results for Experiment 2

Period	Art. 3	Art. 6	Art. 8	Average
2014–2015	0.72	0.64	0.69	0.68
2016–2017	0.70	0.63	0.64	0.66
2016–2017 (10 year gap)	0.69	0.59	0.46	0.58
Experiment 1*	0.78	0.78	0.72	0.76
Experiment 1	0.80	0.80	0.72	0.77

cases (i.e. Articles 3, 6, and 8) and have divided them into smaller groups on the basis of the year of the cases. Specifically, we evaluate the performance on cases from either 2014–2015 or 2016–2017 while we use cases up to 2013 for training. Because violation and non-violation cases were not evenly distributed between the periods, we had to balance them again. Where necessary we used additional cases from the 'violations' test set (used in the previous experiment) to add more violation cases to particular periods. The final distribution of the cases over these periods can be found in Table 9.

We have performed the same grid-search of the parameters of tf-idf and SVM on new training data as we did in the first experiment. We did not opt to use the same parameters as these were tailored to predict mixed-year cases. Consequently, we performed the parameter tuning only on the data up to and including 2013.

The two periods are set up in such a way that we may evaluate the performance for predicting the outcome of cases that follow directly after the ones we train on, versus those which follow later. In the latter case, there is a gap in time between the training period and testing period. Additionally we have conducted an experiment with a 10 year gap between training and testing. In this case, we have trained the model on cases up to 2005 and evaluated the performance using the test set of 2016-2017.

In order to be able to interpret the results better we conducted one additional experiment. For Experiment 1* we have reduced the training data from Experiment 1 to a random sample with the size equal to the amount of cases available for training in Experiment 2 (i.e 356 cases for Article 3, 746 cases for Article 6, etc.), but with all time periods mixed together. We compare cross-validation results on this dataset (Experiment 1*) to the results from the 2014–2015 and 2016–2017 periods in this experiment. The reason we conduct this additional experiment is that it allows us to control for the size of the training data set.

5.2.2 Results

As we can see from Table 10 training on one period and predicting for another is harder than for a random selection of cases (as has been done in Experiment 1). We can also observe that the amount of training data does not influence the results substantially. Experiment 1 produced an average of 0.77 for the chosen articles and Experiment 1* had a result almost as high. However, testing on separate periods resulted in a much lower accuracy. This suggests that predicting future judgements

is indeed a harder task, and it gets harder if the gap between the training and testing data increases.

5.2.3 Discussion

The results of Experiment 2 suggest that we have to take the changing times into account if we want to predict future cases. Therefore, while we can predict the decisions of the past year relatively well, performance drops when there is a larger gap between the period on the basis of which the model was trained and tested. This shows that a continuous integration of published judgements in the system is necessary in order to keep up with the changing legal world and maintain an adequate performance.

While there is a substantial drop in performance on the basis of the 10-year gap, this is likely also caused by a large reduction in training data. Due to the limit on the period, the number of cases used as training data was reduced to 112 for Article 3 (instead of 356), 354 (instead of 754) for Article 6, and 144 (instead of 350) for Article 8. Nevertheless, the large drop for Article 8, suggests that the issues covered by this Article has evolved more in the past decade than those of the other two articles.

Importantly, while these results show that predicting judgements for future cases is possible, the performance is lower than simply predicting decisions for random cases [such as the approach in our Experiment 1 and the approach employed by Aletras et al. (2016), Sulea et al. (2017b)].

5.3 Experiment 3: judges

5.3.1 Set-up

We also wanted to experiment with a very simple model. Consequently, here we use only the names of the judges that constitute a Chamber, including the President, but not including the Section Registrar and the Vice-Section Registrar when present as they do not decide cases. The surnames are extracted based on the list provided by ECtHR on their website.[16] However, ad hoc judges were not extracted, unless they are on the same list (e.g., from a different Section), due to the unavailability of a full list of ad hoc judges for the whole period of the Court's existence. In our extraction efforts we did not account for any misspellings in the case documents, and consequently only correctly spelled surnames were extracted.

We have set-up our prediction model in line with the previous 2 experiments. As input for the model to learn from we have used the surnames of the judges. In total there were 185 judges representing 47 States at different times. The number of judges per State largely depends on when the State ratified the ECHR. Given that the 9-year terms for judges were established recently, some judges might have been part of the Court for a very long time. Some States, such as Serbia, Andorra, and

[16] https://www.echr.coe.int/Pages/home.aspx?p=court/judges.

Table 11 Tenfold cross-validation and test set results for Experiment 3

	Art 2	Art 3	Art 5	Art 6	Art 8	Art 10	Art 11	Art 13	Art 14	Average
Judges	0.61	0.67	0.67	0.68	0.59	0.56	0.67	0.73	0.66	0.65
Test	0.62	0.64	0.66	0.67	0.55	0.65	0.60	0.79	0.73	0.66

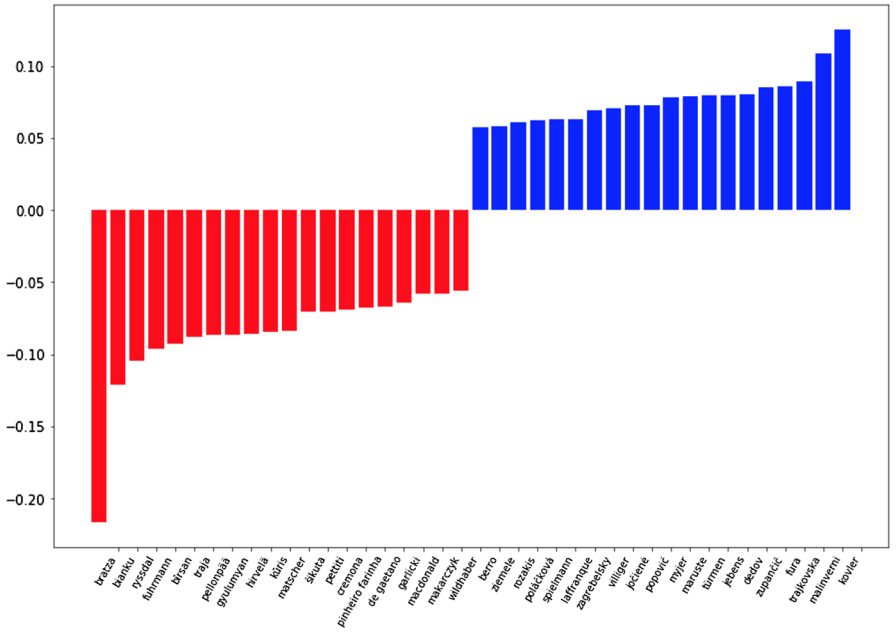

Fig. 4 Coefficients (weights) assigned to different names of the judges for predicting violations of Article 13 of ECHR. Top 20 'violation' predictors (blue on the right) and top 20 'non-violation' predictions (red on the left). (Color figure online)

Azerbaijan only have had 2 judges, while Luxembourg has had 7, and the United Kingdom has had 8. Only a single judge represents the State at any time.

We retained the same set of documents in the dataset as in the Experiment 1, but have only provided the model with the surnames of the judges. However, for this experiment we did not use tf-idf weighing. Instead we represented the features as each judge being either present on the bench or not.

5.3.2 Results

Using the same approach as illustrated in Sect. 5.1, we obtained the results shown in Table 11. In addition, Figs. 4 and 5 show the weights determined by

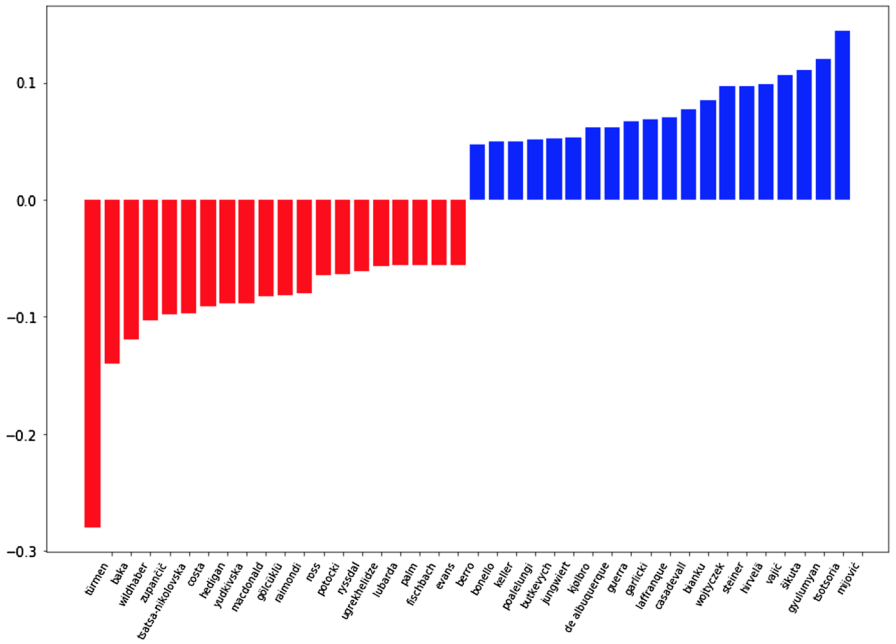

Fig. 5 Coefficients (weights) assigned to different names of the judges for predicting violations of Article 14 of ECHR. Top 20 'violation' predictors (blue on the right) and top 20 'non-violation' predictions (red on the left). (Color figure online)

the machine learning program for the top-20 predictors (i.e. the names of the judges) predicting the violation outcome versus the non-violation outcome.

5.3.3 Discussion

While one may not know the judges that are going to assess a particular case, these results show that the decision is influenced to a large extent by the judges in the Chamber.

In this experiment we did not consider how each judge voted in each case, but only what the final decision on the case was. Consequently, it is important to note that, while some judges may be strongly associated with cases which were judged to be violations (or non-violations), this does not mean that they always rule in favour of a violation when it comes to a particular article of the ECHR. It simply means that this judge is more often in a Chamber which voted for a violation, irrespective of the judge's own opinion.

Importantly, judges have different weights depending on the article that we are considering. For example, Polish judge Lech Garlicki frequently is associated with a 'non-violation' of Article 13, but for Article 14, this judge is more often associated with a 'violation'. This is consistent with the numbers we have in our training data. Garlicki was in a Chamber that voted for a non-violation of Article 14 36 times and

for a violation 34 times. On the other hand, Garlicki was in a Chamber that voted for violation of Article 13 6 times and for a non-violation 38 times.

It is interesting to see the results for the test set in this experiment. While the average results are very similar to cross-validation, scores for particular articles are very high. For instance, when predicting outcomes for Article 13 (*right to an effective remedy*) names of the judges are enough to get a correct judgement for 79% of violation cases. For Article 14 (prohibition of discrimination) the number is also very high—73%.

While the average results are lower than using the n-grams, it is clear that the identity of the judges is still a useful predictor, given that the performance is higher than the (random guess) performance of 50%.

6 Discussion

In this paper we have shown that there is potential in treating case law as quantitative data to predict the outcome of cases. With respect to Aletras et al. (2016), we have increased the amount of articles, as well as the amount of cases we used per article. We have also made different decisions on which parts of the case should be used for machine learning. By excluding the *Law* part of the cases [which Aletras et al. (2016) did not] we have reduced the bias that the model would have when having access to the discussions of the court.

For the 3 articles analysed in Aletras et al. (2016) we have achieved slightly lower scores (0.77 vs. 0.79). However, we believe that our approach is more representative as we make use all of the available data: after balancing the dataset we have 1942 cases for the 3 articles, while Aletras et al. (2016) mention only 584. Furthermore, as they use the *Law* part of the cases, which sometimes also explicitly mentions the verdict, in our opinion their results are likely biased. Thus, we have created a new, reproducible baseline that we (and others) may improve upon in the future.

In this study, we have chosen to build separate models for different articles of the ECHR. When performing the parameter search it was clear that different parameters work better for different articles, and therefore we should not treat them all the same. In all three experiments (using n-grams, predicting the future, or using only the judges' names), we also observed a different performance for the various articles.

We have only used balanced datasets to predict the decisions, however it is still important to remember that the Court rules in favour of a violation much more often than against. That can be partly explained by the filtering out of the non-violation cases during the admissibility stage of the ruling. Many cases with non-violations never make it to the merit stage. Therefore, if we were to teach the model to predict violations better (e.g., when in doubt: predict violation, or giving violation features more weight) the performance would increase. The models we introduced here do not take this distribution into account, hence lowering prediction accuracy in real life. However, our approach does allow us to more clearly identify which features are more important for the system, and therefore lets us make more informed decisions about adapting the model in the future. Moreover, it would be interesting to

experiment with various oversampling techniques (i.e. artificially generating more cases with 'non-violation' verdict), as well as targeted undersampling (i.e. removing only specific cases with 'violation' instead of random sampling) in order to create a better, more representative training set.

It is important to note that while we are trying to develop a system that could predict judicial decisions automatically, we have no intention of creating a system that could replace judges. Rather, in this work, we assess to what extent their decisions are predictable (i.e. transparent).

In this work we have assessed how well a very simple model is able to determine court judgements. This will function as a baseline for future improvements. In future work we are hoping to be able to better predict the court judgements by including using more advanced machine learning techniques, as well as introducing more detailed linguistic information (such as semantics).

In addition to reducing the amount of information that is provided to the model, we would like to be able to take into account the context in which the words occur. For instance an approach using so-called *word embeddings* (see Mikolov et al. 2013) would allow us to have more abstract representations of words and sentences instead of the words themselves. Due to our desire to create models which are intuitive and can be explained, using neural-network-approaches is less suitable as these are often considered black boxes. However, further experiments have to evaluate if it is possible to use neural-network-approaches for some parts of the data processing, while still retaining the ability to analyse the results of the system.

7 Conclusion

In this paper we have conducted several experiments that involved analysing language of the judgements of the European Court of Human Rights to predict if the case was judged to be a violation of one's rights or not. Our results showed that using relatively simple and automatically obtainable information, our models are able to predict decisions correctly in about 75% of the cases, which is much higher than the chance performance of 50%. We have discussed the possibilities of analyzing weights assigned to different phrases by the machine learning algorithm, and how these may be used for identifying patterns within the texts of proceedings. Further research will have to assess how these systems may be improved by using a more sophisticated legal and linguistic analysis.

Acknowledgements We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

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

Basile A, Dwyer G, Medvedeva M, Rawee J, Haagsma H, Nissim M (2017) N-gram: new groningen author-profiling model. CLEF (Working Notes) 2017

Behn D, Langford M (2017) Trumping the environment? An empirical perspective on the legitimacy of investment treaty arbitration. J World Invest Trade 18(1):14–61

Bricker B (2017) Breaking the principle of secrecy: an examination of judicial dissent in the European constitutional courts. Law Policy 39(2):170–191

Bruijn LM, Vols M, Brouwer JG (2018) Home closure as a weapon in the dutch war on drugs: does judicial review function as a safety net? Int J Drug Policy 51:137–147

Bruinsma FJ (2007) The room at the top: separate opinions in the grand chambers of the ECHR (1998–2006). Recht der werkelijkheid 2007(2):7–24

Bruinsma FJ, De Blois M (1997) Rules of law from westport to wladiwostok. Separate opinions in the European Court of Human Rights. Neth Q Hum Rights 15(2):175–186

Chien CV (2011) Predicting patent litigation. Tex Law Rev 90:283

Christensen ML, Olsen HP, Tarissan F (2016) Identification of case content with quantitative network analysis: an example from the ECtHR. Proc JURIX 2016:53–62

Custers B, Leeuw F (2017) Quantitative approaches to empirical legal research. J Empir Leg Stud 34:2449–2456

De Jaeger T (2017) Gerechtelijke achterstand: de piñata van de wetgever. NJW 2017, afl. 361, pp 290–307

Derlén M, Lindholm J (2014) Goodbye *van Gend en Loos*, Hello *Bosman*? Using network analysis to measure the importance of individual CJEU judgments. Eur Law J 20(5):667–687

Derlén M, Lindholm J (2017a) Is it good law? Network analysis and the CJEU's internal market jurisprudence. J Int Econ Law 20(2):257–277

Derlén M, Lindholm J (2017b) Peek-A-Boo, it's a case law system: comparing the European Court of Justice and the United States Supreme Court from a network perspective. Ger Law J 18:647

Derlén M, Lindholm J (2018) Serving two masters: CJEU case law in swedish first instance courts and national courts of precedence as gatekeepers. Available at SSRN: https://ssrn.com/abstract=2952783 or https://doi.org/10.2139/ssrn.2952783

Dhami MK, Belton I (2016) Statistical analyses of court decisions: an example of multilevel models of sentencing. Law Method 10:247–266

Doron II, Totry-Jubran M, Enosh G, Regev T (2015) An american friend in an Israeli Court: an empirical perspective. Isr Law Rev 48(2):145–164

Dyevre A (2015) The promise and pitfalls of automated text-scaling techniques for the analysis of judicial opinions. Available at SSRN: https://ssrn.com/abstract=2626370 or https://doi.org/10.2139/ssrn.2626370

Epstein L, Martin A (2010) Quantitative approaches to empirical legal research. In: The Oxford handbook of empirical legal research. Oxford University Press

Epstein L, Landes WM, Posner RA (2013) The behavior of federal judges: a theoretical and empirical study of rational choice. Harvard University Press, Harvard

Evans M, McIntosh W, Lin J, Cates C (2007) Recounting the courts? Applying automated content analysis to enhance empirical legal research. J Empir Leg Stud 4(4):1007–1039

Evans M, McIntosh W, Lin J, Cates C (2017) Kruisbestuiving tussen straf- en bestuursrecht: de ontwikkeling van de verwijtbaarheid in het bestuursrecht. Nederlands Tijdschrift voor Bestuursrecht 10:351–357

Frankenreiter J (2017a) Network analysis and the use of precedent in the case law of the CJEU—a reply to Derlen and Lindholm. Ger Law J 18:687

Frankenreiter J (2017b) The politics of citations at the ECJ—policy preferences of EU member state governments and the citation behavior of judges at the European Court of Justice. J Empir Leg Stud 14(4):813–857

Frankenreiter J (2018) Are advocates general political? Policy preferences of eu member state governments and the voting behavior of members of the European Court of Justice. Rev Law Econ, De Gruyter 14(1):1–43
Garoupa N, Gili M, Gómez-Pomar F (2012) Political influence and career judges: an empirical analysis of administrative review by the Spanish Supreme Court. J Empir Leg Stud 9(4):795–826
Goanta C (2017) Big law, big data. Available at SSRN: https://ssrn.com/abstract=3166290. Accessed 24 June 2019
Golbeck J, Robles C, Edmondson M, Turner K (2011) Predicting personality from twitter. In: Privacy, security, risk and trust (passat) and 2011 IEEE third inernational conference on social computing (SocialCom). IEEE, pp 149–156
Grabmair M (2017) Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs. In: Proceedings of the 16th edition of the international conference on artificial intelligence and law. ACM, pp 89–98
Holá B, Bijleveld C, Smeulers A (2012) Consistency of international sentencing: ICTY and ICTR case study. Eur J Criminol 9(5):539–552
Hunter C, Nixon J, Blandy S (2008) Researching the judiciary: exploring the invisible in judicial decision making. J Law Soc 35(s1):76–90
Hutchinson T, Duncan N (2012) Defining and describing what we do: Doctrinal legal research. Deakin Law Rev 17:83
Joachims T (2002) Learning to classify text using support vector machines: methods, theory and algorithms, vol 186. Kluwer Academic Publishers, Norwell
Katz DM (2012) Quantitative legal prediction-or-how i learned to stop worrying and start preparing for the data-driven future of the legal services industry. Emory Law J 62:909
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
Law DS (2017) The global language of human rights: a computational linguistic analysis. Available at SSRN: https://ssrn.com/abstract=3049625. Accessed 24 June 2019
Lindholm J, Derlén M (2012) The Court of Justice and the Ankara agreement: exploring the empirical approach. Europarättslig tidskrift 3:462–481
Livermore MA, Riddell AB, Rockmore DN (2017) The Supreme Court and the judicial genre. Ariz Law Rev 59:837
Lupu Y, Voeten E (2012) Precedent in international courts: a network analysis of case citations by the European Court of Human Rights. Br J Polit Sci 42(2):413–439
Madsen MR (2018) Rebalancing European human rights: has the Brighton Declaration engendered a new deal on human rights in Europe? J Int Dispute Settl 9:199–222
Matthews AA (2017) Connected courts: the diffusion of precedent across state supreme courts. PhD thesis, The University of Iowa
Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems 26 (NIPS 2013), pp 3111–3119
Mochales R, Moens MF (2008) Study on the structure of argumentation in case law. In: Proceedings of the 2008 conference on legal knowledge and information systems, pp 11–20
Olsen HP, Küçüksu A (2017) Finding hidden patterns in ECtHR's case law: On how citation network analysis can improve our knowledge of ECtHR's Article 14 practice. Int J Discrim Law 17(1):4–22
op Vollenbroek MB, Carlotto T, Kreutz T, Medvedeva M, Pool C, Bjerva J, Haagsma H, Nissim M (2016) Gronup: Groningen user profiling. CLEF (Working Notes) 2016:846–857
Panagis Y, Christensen ML, Sadl U (2016) On top of topics: leveraging topic modeling to study the dynamic case-law of international courts. Proc JURIX 2016:161–166
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, VanderPlas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
Rachlinski JJ, Wistrich AJ (2017) Judging the judiciary by the numbers: empirical research on judges. Annu Rev Law Soc Sci 13:203–229
Rangel F, Rosso P (2013) Use of language and author profiling: identification of gender and age. Nat Lang Process Cogn Sci 177:117–186

Ruppert E, Hartung D, Sittig P, Gschwander T, Rönneburg L, Killing T, Biemann C (2018) Lawstats—large-scale German Court decision evaluation using web service classifiers. In: International cross-domain conference for machine learning and knowledge extraction. Springer, pp 212–222

Šadl U, Olsen HP (2017) Can quantitative methods complement doctrinal legal studies? Using citation network and corpus linguistic analysis to understand international courts. Leiden J Int Law 30(2):327–349

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, Shah A, Kosinski M, Stillwell D, Seligman ME et al (2013) Personality, gender, and age in the language of social media: the open-vocabulary approach. PloS ONE 8(9):e73791

Shulayeva O, Siddharthan A, Wyner A (2017) Recognizing cited facts and principles in legal judgements. Artif Intell Law 25(1):107–126

Sulea OM, Zampieri M, Malmasi S, Vela M, Dinu LP, van Genabith J (2017a) Exploring the use of text classification in the legal domain. arXiv preprint arXiv:171009306

Sulea OM, Zampieri M, Vela M, van Genabith J (2017b) Predicting the law area and decisions of French Supreme Court cases. arXiv preprint arXiv:170801681

Tarissan F, Nollez-Goldbach R (2014) The network of the international criminal court decisions as a complex system. In: Iscs 2013: Interdisciplinary symposium on complex systems. Springer, pp 255–264

Tarissan F, Nollez-Goldbach R (2015) Temporal properties of legal decision networks: a case study from the international criminal court. In: 28th International conference on legal knowledge and information systems (JURIX'2015)

Tarissan F, Nollez-Goldbach R (2016) Analysing the first case of the international criminal court from a network-science perspective. J Complex Netw 4(4):616–634

Trompper M, Winkels R (2016) Automatic assignment of section structure to texts of Dutch court judgments. In: Legal knowledge and information systems

van Dijck G (2018) Victim-oriented tort law in action: an empirical examination of catholic church sexual abuse cases. J Empir Leg Stud 15(1):126–164

Van Hoecke M (2011) Foreword in 'methodologies of legal research'. European Academy of Legal Theory Series. Hart Publishing, Oxford, pp I–IX

Vols M, Jacobs J (2017) Juristen als rekenmeesters: Over de kwantitatieve analyse van jurisprudentie. In: van den Berg PAJ, Molier G (eds) In dienst van het recht: Opstellen aangeboden aan prof. mr. J.G. Brouwer ter gelegenheid van zijn afscheid als hoogleraar Algemene Rechtswetenschap aan de Rijksuniversiteit Groningen (Brouwer bundel). Boom Juridisch, Den Haag, pp 89–104

Vols M, Tassenaar P, Jacobs J (2015) Anti-social behaviour and European protection against eviction. Int J Law Built Environ 7(2):148–161

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

Whalen R (2016) Legal networks: the promises and challenges of legal network analysis. Mich State Law Rev 2016(539)

White RC, Boussiakou I (2009) Separate opinions in the European Court of Human Rights. Hum Rights Law Rev 9(1):37–60

Wongchaisuwat P, Klabjan D, McGinnis JO (2017) Predicting litigation likelihood and time to litigation for patents. In: Proceedings of the 16th edition of the international conference on articial intelligence and law. ACM, pp 257–260

Wyner A, Mochales-Palau R, Moens MF, Milward D (2010) Approaches to text mining arguments from legal cases. In: Francesconi E, Montemagni S, Peters W, Tiscornia D (eds) Semantic processing of legal texts. Lecture notes in computer science, vol 6036. Springer, Berlin, pp 60–79

Zhang AH, Liu J, Garoupa N (2017) Judging in Europe: do legal traditions matter? Available at SSRN: https://ssrn.com/abstract=3082854. Accessed 24 June 2019