Judicial Decisions of the European Court of Human Rights: Looking into the Crystal Ball

Masha Medvedeva[1,2] Michel Vols[2] Martijn Wieling[1]
[1]Center for Language and Cognition Groningen
[2]Department of Legal Methods
University of Groningen, the Netherlands
m.medvedeva@rug.nl m.vols@rug.nl m.b.wieling@rug.nl

Abstract

When courts started publishing judgements, big data analysis (i.e. large-scale statistical analysis 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.

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 existed since the early 90s. Today computers are attempting automatic summarization of legal information and information extraction (e.g., *DecisionExpress*), categorization of legal resources (e.g., *BiblioExpress*), and statistical analysis (e.g., *StatisticExpress*).

Language analysis in the legal domain has also been used for a long time. In criminal justice, for example, text classification is often used in forensics linguistics. While in earlier times, for instance, in the Unabomber case,[2] the analysis

[1] "European legislation on re-use of public sector information — Digital Single Market," n.d.

[2] https://archives.fbi.gov/archives/news/stories/2008/april/unabomber_042408

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, Robles, Edmondson, & Turner, 2011), and even the identity of the author[3] almost flawlessly.

In this study we address the potential of using language analysis and automatic information extraction in order to boost statistical research in the legal domain. In this paper 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 Section 3), we are able to use a computer to perform quantitative analysis of 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 enough, we may subsequently analyse which words made the most impact on this decision and thus identify what influences the judicial decision the most.

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 we want to assess to what extent their decisions are predictable (i.e. transparent).

In the following section, we will discuss the background for this study. In Section 3 we discuss how machine learning can be used for classification of texts within the legal domain. Section 4 is dedicated to describing data we have used for our experiments. In Section 5 we describe three experiments that we have conducted for this study and report the results. In Sections 6 and 7 respectively we discuss the results and draw conclusions.

2 Background

As long as there are judges, there is case law and as long as there is case law, legal researchers have analysed it. 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: 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' (Hutchinson and Duncan, 2012: 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 analyses case law (Epstein & Martin, 2010). Nowadays, however, because of the massive amount of case law that is published, it is physically impos-

[3] `https://github.com/sixhobbits/yelp-dataset-2017`

sible 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 are promising and can help us to make sense of all available case law (Custers & Leeuw, 2017; Derlén & Lindholm, 2018; Goanta, 2017; Šadl & 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, that contains manually collected and expertly-coded data on the US Supreme Court's behaviour of the last two hundred years (Katz, Bommarito II, & Blackman, 2017). A large amount of these studies analyse the relationship between the political background of judges and their decision-making (see Epstein, Landes, and Posner, 2013; Rachlinski and Wistrich, 2017; Frankenreiter, 2018).

In other countries 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, Nixon, and Blandy, 2008 hold: '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' (2008, 79). 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, Zampieri, Malmasi, et al., 2017), Germany (Dyevre, 2015; Bricker, 2017), Israel (Doron, Totry-Jubran, Enosh, & Regev, 2015), Latvia (Bricker, 2017), the Netherlands (Vols, Tassenaar, & Jacobs, 2015; Vols & Jacobs, 2017; van Dijck, 2016; Bruijn, Vols, & Brouwer, 2018), Poland (Bricker, 2017), Slovenia (Bricker, 2017), Spain (Garoupa, Gili, & Gómez-Pomar, 2012) and Sweden (Derlén & 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, Bijleveld, & Smeulers, 2012; Tarissan & Nollez-Goldbach, 2014, 2015), the Court of Justice of the European Union (Lindholm & Derlén, 2012; Derlén & Lindholm, 2014; Tarissan & Nollez-Goldbach, 2016; Derlén & Lindholm, 2017a, 2017b; Frankenreiter, 2017a, 2017b; Zhang, Liu, & Garoupa, 2017) or the European Court of Human Rights (F. J. Bruinsma & De Blois, 1997; J. F. Bruinsma, 2007; White & Boussiakou, 2009; Christensen, Olsen, & Tarissan, 2016; Olsen & Küçüksu, 2017; Madsen, 2017).

In most research projects, the case law seems to be manually collected and hand-coded. Still, a number of researchers use computerized techniques to collect the case law and automatically generate usable information from the collected case law (see Trompper and Winkels, 2016; Livermore, Riddell, and Rockmore, 2017; Shulayeva, Siddharthan, and Wyner, 2017; Law, 2017) . For example, Dyevre, 2015 discusses the use of automated content analysis tech-

niques such as Wordscores and Wordfish in the legal discipline. He applied these two techniques to analyse a dataset of 16 judgements of the German Federal Constitutional Court on European integration. He found that Wordscore as Wordfish can 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. Aletras, Tsarapatsanis, Preoţiuc-Pietro, and Lampos, 2016 automatically extracted textual features (N-grams and sets of N-grams (topics)) from judgements of the ECtHR. Panagis, Christensen, and Sadl, 2016 used topic modelling techniques to automatically find latent topics in a set of judgements Court of Justice of the European Union 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. F. J. Bruinsma and De Blois, 1997; White and Boussiakou, 2009; De Jaeger, 2017; Madsen, 2017; Vols and Jacobs, 2017). Other studies present results of relatively simple statistical tests such as correlation analysis (e.g. Doron et al., 2015; Evans, McIntosh, Lin, and Cates, 2017; Bruijn et al., 2018). A growing body of papers present results of more sophisticated statistical tests such as 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).

Still, researchers outside the USA have conducted such analyses as well (see for references Garoupa et al., 2012, footnotes 9-11). See also: Holá et al., 2012; Behn and Langford, 2017; Bricker, 2017; van Dijck, 2016; Zhang et al., 2017; Frankenreiter, 2017a, 2018.

Yet, a number of researchers have begun to apply more sophisticated quantitative methods to analyse case law. 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). Other scholars have applied this method to case law from European countries, such as Sweden (Derlén & 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 & Derlén, 2012; Derlén & Lindholm, 2014; Tarissan & Nollez-Goldbach, 2016; Derlén & Lindholm, 2017a, 2017b; Frankenreiter, 2017a, 2017b, 2018). Derlén and Lindholm, 2017a: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 concerned citation network analyses of case law of the ECtHR (Lupu & Voeten, 2012; Christensen et al., 2016; Olsen & Küçüksu, 2017). Olsen and Küçüksu, 2017:19 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 & Olsen, 2017). Other papers used this method to analyse case

law of the International Criminal Court (Tarissan & Nollez-Goldbach, 2014, 2015, 2016). A relatively small number of studies have used machine learning techniques to analyse case law (see Evans, McIntosh, Lin, and Cates, 2007; Custers and Leeuw, 2017). Again, researchers in the United States seem to be the first to use this techniques to predict the courts' decisions or voting behaviour of judges (Katz, 2012; Wongchaisuwat, Klabjan, & McGinnis, 2017). Recently, Katz et al., 2017 developed a prediction model that aims to predict whether the US Supreme Court as a whole affirms or reverse 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 achieves 70.2% accuracy at the case outcome level and 71.9% at the justice vote level.

Outside the United States, only few scholars have used machine learning techniques to predict the courts' decisions. Sulea, Zampieri, Vela, and van Genabith, 2017 used machine learning techniques to analyse case law of the French Court de Cassation. They aimed to predict the law area of a case and the court ruling. Their model achieves over 92% accuracy. 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 achieves 79% accuracy at the case outcome level.

In this paper we build upon on the results achieved by Aletras et al., 2016 for the ECtHR and examine additional approaches.

3 Machine learning for legal text classification

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 (e.g., a verdict) associated to a new element (e.g., a case). For our task we will use 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 judgement. 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 computer, 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 computer uses the information it identified to be important during the training phase.

To illustrate how supervised machine learning works in very simplistic terms let's imagine a non-textual example. Let's say we want to write a programme

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: *cat* or *dog*. Then we show the system those pictures with labels one by one. If we show enough pictures, eventually the programme 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 it *learns* this information, we can show it a picture without a label and it will hopefully guess which *class* the picture belongs to.

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 programme can analyse the text and the style it was written in. Research conducted on social media data showed that when training such models, we can see that men and women on social media generally talk about different things: women use more pronouns than men (Rangel & Rosso, 2013), while men swear more (Schwartz et al., 2013).

For this research we wrote a computer programme that analyses texts of judgements of ECtHR cases available on the HUDOC website[4] and predicts whether any particular article of ECHR was violated.

As we have mentioned in Section 2, to our knowledge techniques from machine learning have not often been used in the legal domain. Nevertheless, the data that we have is perfect 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. 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 labels from each other with the least amount of error (see Figure 1).

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 programme indicate for every case whether it thinks that it is 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

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

[5]http://diggdata.in/post/94066544971/support-vector-machine-without-tears

Figure 1: Illustration of an SVM dividing data into classes.[5]

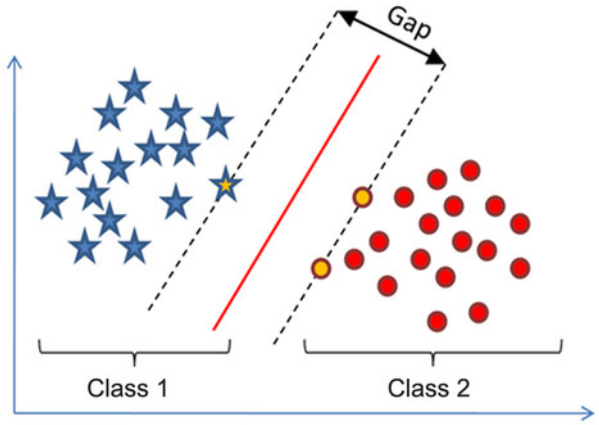
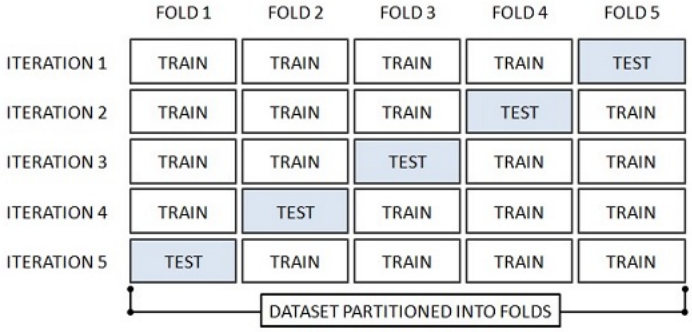


Figure 2: Example of 5-fold cross-validation.[6]



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 5-fold cross-validation and train and test the model 5 times (see Figure 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.

[6]http://www.dummies.com/programming/big-data/data-science/resorting-cross-validation-machine-learning/

4 Data

In this study we use the publicly available data published by the ECtHR. We consider all admissible judgements of both the Chamber and the Grand Chamber, without specifically distinguishing between the two.

In order to create a database which we can use to conduct our experiments on, we had to automatically collect all data online. We therefore created a programme that automatically downloaded all documents in English from the HUDOC website.[7] Our database[8] contains all the 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 and representative 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 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' by the state. The specific distribution, however, depends on the article that is being considered.

4.1 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 different decisions of the Court.

Generally, 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 programme that predicts

[7] "HUDOC - European Court of Human Rights," n.d.

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

Table 1: Initial distribution of cases (in English) obtained on HUDOC website on September 11, 2017.

Art.	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 privatAe 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

whether there is a violation of article 13, and feed it all 170 'non-violation' cases together with all 1230 'non-violation' cases. With such an imbalance in the number of cases per type, it is likely that the programme 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. These cases were randomly selected and removed from the dataset. These missing cases are available online.[9]

The results of this 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

[9] See test20 at https://www.dropbox.com/s/lxpvvqdwby30157/crystal$_b$all$_d$ata.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

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

5.1 Experiment 1: textual analysis

The data we provided to the machine learning programme does not include all 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 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 programme 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*).

Until now we have ignored one important detail, namely how the text of a case is represented for the machine learning programme. For this we need to define features (i.e. an observable characteristic) of each case. Using the cats-and-dogs example above, 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 programme then will determine which features are the most important for classification. In 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 important 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 programme will then determine which of these words or word sequences are most characteristic for either a violation or a non-violation. A 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, 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 are possible to automatically extract, we need to decide what the values are associated with them for each case

description. A very simple approach would be to use 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 as the frequency with which an n-gram occurs in a document occurs. 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 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 this absolute 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*).[10]

In order to identify which feature-set 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 sentence occurs in multiple case descriptions) and therefore are less useful to include. For this reason we limit the maximum word sequence to 4. But there are also other choices to make, such as if all words should be converted to lowercase, or if the capitalisation is important.

All parameters we have tried are listed in Table 3[11]. 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 3-fold cross-validation for each article. The programme 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 articles 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 programme 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.[12]

For most articles unigrams worked best, 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 *Proce-*

[10] We use the formula defined by `scikit-learn` Python package (Pedregosa et al., 2011): $tfidf(d,t) = tf(t) * idf(d,t)$ where $idf(d,t) = \log(n/df(d,t)) + 1$ where n is the total number of documents and $df(d,t)$ is the document frequency. Document frequency is the number of documents d that contain term t. In our case the terms are n-grams.

[11] 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

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

Table 3: List of values over which the grid-search was performed.

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
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

Table 4: Selected parameters used for the best model.

article	**parts**	**n-grams**	**remove capitalization**	**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	✓	✓

dure 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.*

After investigating which combinations of parameters worked best, we used these parameter settings together with 10-fold cross-validation to ensure that the model performed well in general and was not overly sensitive to the specific set of cases on which is was trained. When performing 10-fold cross-validation instead of 3-fold 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 'n-grams' in Table 5. Note that as we used a balanced dataset, 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.

During the training phase, an SVM assigns different weights to the bits of information it is given (i.e. n-grams). After training the model we may inspect these weights in order to see what information had the most impact on the model's decision to predict a certain ruling. The n-grams that were most important hopefully may yield some insight into what might influence Court's decision making. In Figure 3 we may see the phrases that ranked the highest to identify a case as a 'violation' (blue) or a 'non-violation' (red).

Figure 3: Coefficients (weights) assigned to different n-grams for predicting violations of article 2 of ECHR. Top 20 'violation' predictors (blue) and top 20 'non-violation' predictions (red).

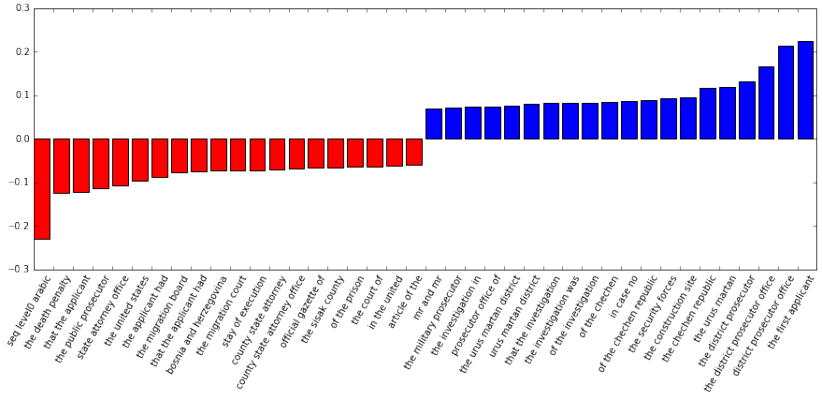


Table 5: Cross-validation (10-fold) results for Experiment 1.

	art 2	art 3	art 5	art 6	art 8	art 10	art 11	art 13	art 14	average
n-grams	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

After tuning and evaluating the results using cross-validation we have tested the system on violation cases (for article 14: non-violation cases) that the programme has never seen before. The results can be found in Table 5 (test). On this set the model performed very similar and sometimes even better.

For multiple articles (6, 8, 10, 11, 13) it performed 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 worse. The opposite happens when the model learns to predict violations better. In that case, the results on this violation-only test set get higher. Note, that the test set for article 14 contains non-violations only, increase in the performance here indicates that the model has probably learnt 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.

5.2 Experiment 2: predicting the future

In the first experiment we were training and predicting using cross-validation for randomly selected data without any restrictions on the cases selected. In this section we will experiment 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 know. 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 show lower, but more realistic performance than the results shown in Table 5. For this reason, we have set up an additional experiment to check whether this is indeed the case and whether our system is sensitive to this change. For this experiment, we have only considered the datasets with a larger amounts of cases (articles 3, 6, and 8) and have split them into smaller groups. Specifically, we evaluate the performance on cases from either 2014-2015 (period 1) or 2016-2017 (period 2) while we use cases up to 2013 for testing. In addition, we evaluate if the performance

on the 2016-2017 test set is improved when we also use the data from 2014-2015 for training. We use cross-validation on training set to choose the best parameters, and then predict judgements for cases from both periods, similarly to how we predicted judgements for the test set in Experiment 1. We have used the same cases as in the previous experiment, and split them according to the year. 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 6.

Table 6: Number of cases

	art 3	art 6	art 8
Train set	356	746	350
2014-2015	72	80	52
2016-2017	140	90	56

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.

In addition, we have tried to predict judgements of cases from the 'whole period' (from 2014 to 2017), but reduced the amount of cases to the same as in 2014-2015 period (e.g. 72 cases for article 3, 80 cases for article 6, etc) in order to be able to compare results. The cases were chosen randomly from 2014-2017 period, but we kept the 50-50 distribution of 'violation'/'non-violation' cases.

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 want to use the same parameters, because they were tailored to predict mixed-year cases. To create an unbiased model, we therefore performed the parameter tuning only on the data up to and including 2013.

As we can see from table 7 training on one period and predicting for another is harder than for a random selection of cases (as shown in Table 5). Note that the lower results of 2014-2015 (compared to those reported in Table 5) are not only worse due to the difference in time, but also due to the smaller amount of training data.

Results for article 3 in 2014-2017 period are lower than when tested on a more particular particular period, while performance on articles 6 and 8 is higher. This might suggest that due to the diversity of issues that the latter articles deal with, the sample we have for a particular period differs in substance from the training data and therefore is harder to predict.

As we mentioned, lower results, compared to mixed-case predictions can also be partly explained by a smaller train set. In order to test how much size of the training data influences the performance, we have also shown the results for

Table 7: 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
2014-2017	0.68	0.65	0.73	0.69
2014-2015*	0.71	0.68	0.71	0.70
2016-2017*	0.75	0.67	0.64	0.69
Experiment 1	0.80	0.80	0.72	0.77

those periods with additional training data. When testing on 2014-2015* period we added 2016-2017 to the train set, when testing on 2016-2017* we added 2014-2015 to the train set. We did not adapt any of the grid-search parameters again. We can see that the performance is a bit better when trained on more data, but still not as well as when training and testing on random cases.

We saw that when adding more training data and training on the closer period of time (e.g. 2016-2017*) the results are higher. This means that as long as we keep continuously updating the dataset with new cases and improving the machine learning model, predictions of future cases should be possible.

5.3 Experiment 3: judges

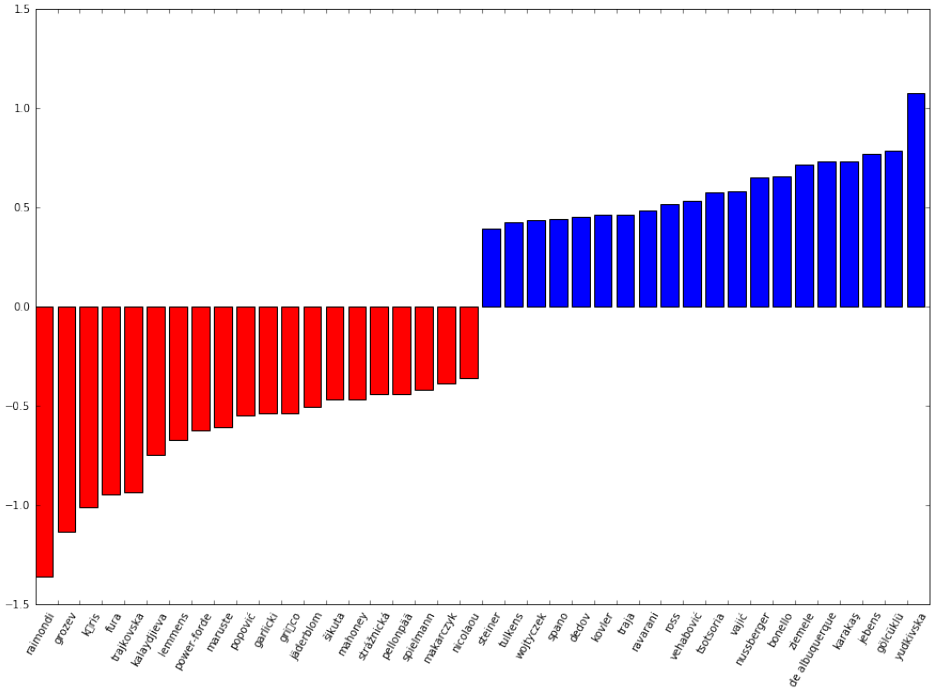
One might rightfully argue that because the summaries were written after the case, t he information that was found irrelevant and was dismissed during the trial might not appear in the text. Therefore it would be not completely fair to use the summary of a case in order to predict the outcome. However the segments of the case that are used in our training do not contain discussion of the cases, but only the very basic information that should be easier to replicate. Another argument against the adequacy of using summaries is that we model textual information and it is likely that the programme often learns particular phrasings rather than specific information of the case. Consequently, we cannot exclude the possibility that our model does not predict the outcome of the case on the basis of objective information, but rather the outcome of the case written down by someone who already had knowledge about the outcome. For this reason, we also wanted to experiment with a very simple model exclusively containing objective information. In this case, we illustrate this approach using only the names of the judges that constitute a Chamber.

Table 8: 10-fold cross-validation results for Experiment 3.

	art 2	art 3	art 5	art 6	art 8	art 10	art 11	art 13	art 14	average
judges	0.66	0.69	0.67	0.64	0.56	0.64	0.69	0.73	0.69	0.67

Using the same approach as illustrated in Section 5.1, we obtained the results shown in Table 8. In addition, Figure 4 shows the weights determined by the machine learning programme for the top-20 predictors (i.e. the names of the judges) predicting the violation outcome versus the non-violation outcome.

Figure 4: Coefficients (weights) assigned to different names of the judges for predicting violations of article 2 of ECHR. Top 20 'violation' predictors (blue) and top 20 'non-violation' predictions (red).

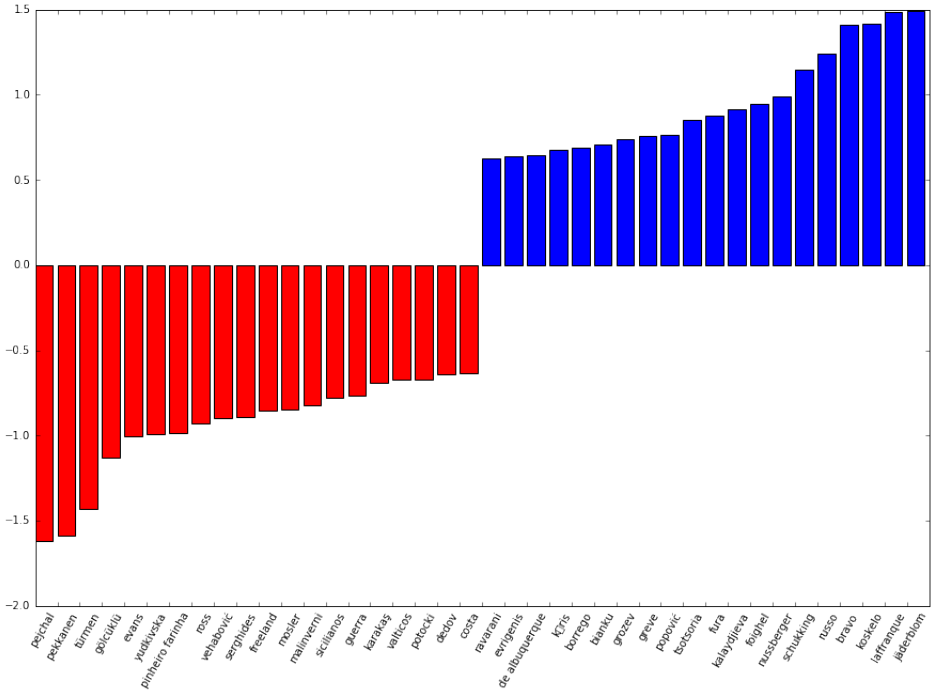


While one may not know the judges that are going to assess a particular case, these results show that the decision is influenced to a large extent by the judges in the Chamber.

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 they 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. In this experiment we did not consider how each judge voted in each case, but only what the final decision on the case was.

Importantly, judges have different weights depending on the article that we are testing for (see also Figure 5). For example, Ganna Yudkivska, a judge from Ukraine, frequently is associated with a 'violation' of article 2, but a non-violation for article 14.

Figure 5: Coefficients (weights) assigned to different names of the judges for predicting violations of article 14 of ECHR. Top 20 'violation' predictors (blue) and top 20 'non-violation' predictions (red).



6 Discussion

In this paper we have shown that automatic analysis of court cases for predictions has a lot of potential. With respect to Aletras et al., 2016, we have increased the amount of articles and cases 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 we have reduced the bias that the model would have when having access to the discussions of the court.

As an improvement over Aletras et al., 2016, we now have an opportunity to analyse the weight that the algorithm assigns to different phrases (or judges). We may use this, for example, by simplifying our model to only consider these important phrases.

In this work we were trying to see what the model can do with the bare minimum of looking at different phrases in text, and therefore created a baseline for future improvements. In future work we are hoping to be able to achieve higher performance by conducting experiments using more linguistically-oriented approaches, such as semantic analysis as well as more advanced machine learning techniques, including neural networks.

It would also be interesting to analyse the votes of each judge and dissenting opinions of the ones that disagree in order to be able to predict the vote of each judge.

7 Conclusion

For 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 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%. Further research will assess how these systems can conduct a more sophisticated legal and linguistic analysis in order to improve performance.

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 Computer Science*, *2*, e93.

Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., & Nissim, M. (2017). N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*.

Behn, D. & Langford, M. (2017). Trumping the environment? an empirical perspective on the legitimacy of investment treaty arbitration. *The Journal of World Investment & 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, L. M., Vols, M., & Brouwer, J. G. (2018). Home closure as a weapon in the dutch war on drugs: Does judicial review function as a safety net? *International Journal of Drug Policy*, *51*, 137–147.

Bruinsma, F. J. & De Blois, M. (1997). Rules of law from westport to wladiwostok. separate opinions in the european court of human rights. *Netherlands Quarterly of Human Rights*, *15*(2), 175–186.

Bruinsma, J. F. (2007). The room at the top: Separate opinions in the grand chambers of the echr (1998-2006). *Recht der werkelijkheid*, *2007*(2), 7–24.

Chien, C. V. (2011). Predicting patent litigation. *Tex. L. Rev. 90*, 283.

Christensen, M. L., Olsen, H. P., & Tarissan, F. (2016). Identification of case content with quantitative network analysis: An example from the ecthr. In *Jurix* (pp. 53–62).

Custers, B. & Leeuw, F. (2017). Quantitative approaches to empirical legal research. *Journal of Empirical Legal Studies*, *34*, 2449–2456.

De Jaeger, T. (2017). Gerechtelijke achterstand: De piñata van de wetgever. *NJW*, 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. *European Law Journal*, *20*(5), 667–687.

Derlén, M. & Lindholm, J. (2017a). Is it good law? network analysis and the cjeu's internal market jurisprudence. *Journal of International Economic 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. *German LJ*, *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.

Dhami, M. K. & Belton, I. (2016). Statistical analyses of court decisions: An example of multilevel models of sentencing. *Law and Method*, *10*, 247–266.

Doron, I. I., Totry-Jubran, M., Enosh, G., & Regev, T. (2015). An american friend in an israeli court: An empirical perspective. *Israel Law Review*, *48*(2), 145–164.

Dyevre, A. (2015). The promise and pitfalls of automated text-scaling techniques for the analysis of judicial opinions.

Epstein, L., Landes, W. M., & Posner, R. A. (2013). *The behavior of federal judges: A theoretical and empirical study of rational choice*. Harvard University Press.

Epstein, L. & Martin, A. D. (2010). Quantitative approaches to empirical legal research.

European legislation on re-use of public sector information — Digital Single Market. (n.d.). https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information. (Accessed on 01/10/2018).

Evans, M., McIntosh, W., Lin, J., & Cates, C. (2007). Recounting the courts? applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, *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. *German LJ*, *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. *Journal of Empirical Legal Studies*, *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. *Browser Download This Paper*.

Garoupa, N., Gili, M., & Gómez-Pomar, F. (2012). Political influence and career judges: An empirical analysis of administrative review by the spanish supreme court. *Journal of Empirical Legal Studies*, *9*(4), 795–826.

Goanta, C. (2017). Big law, big data. *7*(3), 1–20.

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), 2011 ieee third international conference on* (pp. 149–156). IEEE.

Holá, B., Bijleveld, C., & Smeulers, A. (2012). Consistency of international sentencing: Icty and ictr case study. *European journal of criminology*, *9*(5), 539–552.

HUDOC - European Court of Human Rights. (n.d.). https://hudoc.echr.coe.int/eng. (Accessed on 01/10/2018).

Hunter, C., Nixon, J., & Blandy, S. (2008). Researching the judiciary: Exploring the invisible in judicial decision making. *Journal of Law and Society*, *35*(s1), 76–90.

Hutchinson, T. & Duncan, N. (2012). Defining and describing what we do: Doctrinal legal research. *Deakin L. Rev. 17*, 83.

Katz, D. M. (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 LJ*, *62*, 909.

Katz, D. M., Bommarito II, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, *12*(4), e0174698.

Law, D. S. (2017). The global language of human rights: A computational linguistic analysis.

Lindholm, J. & Derlén, M. (2012). The court of justice and the ankara agreement: Exploring the empirical approach. *Europarättslig tidskrift*, (3), 462–481.

Livermore, M. A., Riddell, A. B., & Rockmore, D. N. (2017). The supreme court and the judicial genre. *Ariz. L. 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. *British Journal of Political Science*, *42*(2), 413–439.

Madsen, M. R. (2017). Rebalancing european human rights: Has the brighton declaration engendered a new deal on human rights in europe? *Journal of International Dispute Settlement*.

Matthews, A. A. (2017). *Connected courts: The diffusion of precedent across state supreme courts* (Doctoral dissertation, The University of Iowa).

Olsen, H. P. & 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. *International Journal of Discrimination and the Law*, *17*(1), 4–22.

op Vollenbroek, M. B., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., ... Nissim, M. (2016). Gronup: Groningen user profiling.

Panagis, Y., Christensen, M. L., & Sadl, U. (2016). On top of topics: Leveraging topic modeling to study the dynamic case-law of international courts. In *Jurix* (pp. 161–166).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rachlinski, J. J. & Wistrich, A. J. (2017). Judging the judiciary by the numbers: Empirical research on judges. *Annual Review of Law and Social Science*, *13*, 203–229.

Rangel, F. & Rosso, P. (2013). Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science*, *177*.

Šadl, U. & Olsen, H. P. (2017). Can quantitative methods complement doctrinal legal studies? using citation network and corpus linguistic analysis to understand international courts. *Leiden Journal of International Law*, *30*(2), 327–349.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... Seligman, M. E., 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. *Artificial Intelligence and Law*, *25*(1), 107–126.

Sulea, O.-M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P., & van Genabith, J. (2017). Exploring the use of text classification in the legal domain. *arXiv preprint arXiv:1710.09306*.

Sulea, O.-M., Zampieri, M., Vela, M., & van Genabith, J. (2017). Predicting the law area and decisions of french supreme court cases. *arXiv preprint arXiv:1708.01681*.

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* (pp. 255–264). Springer.

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. *Journal of Complex Networks*, *4*(4), 616–634.

Trompper, M. & Winkels, R. (2016). Automatic assignment of section structure to texts of dutch court judgments.

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

van Dijck, G. (2016). Victim oriented tort law in action: An empirical examination of catholic church sexual abuse cases. *Browser Download This Paper*.

Vols, M. & Jacobs, J. (2017). Juristen als rekenmeesters: Over de kwantitatieve analyse van jurisprudentie. *VAN DEN BERG, P A J and G. MOLIER, eds, In dienst van het recht*, 89–104.

Vols, M., Tassenaar, P., & Jacobs, J. (2015). Anti-social behaviour and european protection against eviction. *International Journal of Law in the Built Environment*, *7*(2), 148–161.

Whalen, R. (2016). Legal networks: The promises and challenges of legal network analysis. *Mich. St. L. Rev.* 539.

White, R. C. & Boussiakou, I. (2009). Separate opinions in the european court of human rights. *Human Rights Law Review*, *9*(1), 37–60.

Wongchaisuwat, P., Klabjan, D., & McGinnis, J. O. (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* (pp. 257–260). ACM.

Zhang, A. H., Liu, J., & Garoupa, N. (2017). Judging in europe: Do legal traditions matter?