Automatic Judgement Forecasting for Pending Applications of the European Court of Human Rights

Masha Medvedeva[1,2], Ahmet Üstun[1], Xiao Xu[1,3], Michel Vols[2] and Martijn Wieling[1]

[1]*Centre for Language and Cognition, University of Groningen, the Netherlands*
[2]*Department of Legal Methods, University of Groningen, the Netherlands*
[3]*Netherlands Interdisciplinary Demographic Institute, The Netherlands*

Abstract

Judicial decision classification using Natural Language Processing and machine learning has received much attention in the last decade. While many studies claim to 'predict judicial decisions', most of them only *classify* already made judgements. Likely due to the lack of data, there have been only a few studies that discuss the data and the methods to forecast *future* judgements of the courts on the basis of data available before the court judgement is known. Besides proposing a more consistent and precise terminology, as classification and forecasting each have different uses and goals, we release a first benchmark dataset consisting of documents of the European Court of Human Rights to address this task. The dataset includes raw data as well as pre-processed text of final judgements, admissibility decisions and communicated cases. The latter are published by the Court for pending applications (generally) many years before the case is judged, allowing one to forecast judgements for pending cases. We establish a baseline for this task and illustrate that it is a much harder task than simply classifying judgements.

Keywords

judicial decisions, machine learning, text classification, datasets, neural networks

1. Introduction

Digital access to case law (i.e. court judgements) provides us with a unique opportunity to process legal data automatically on a large scale using natural language processing techniques. It is, therefore, not surprising that using machine learning for judicial outcome classification has seen a substantial increase in recent years. If we rely on the presumption that legal systems and legal decision-making are consistent and predictable, we should be able to ultimately create a system that would be able to automatically predict judicial decisions correctly. Consequently, such a system could also be used to identify patterns which might be less consistent and perhaps reveal biases in the legal system and judicial decision-making.

At present, much work has been done on classifying the outcomes of final judgements [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Classification of final judgements is in principle a useful task, as it may be used to identify important factors and arguments of the court, and thereby may provide insight into the process of decision-making. Some previous research even suggests that one day such classification systems will be able to provide legal assistance [2] and promote accessibility to justice [5], while others suggest that the courts, such as the European Court of Human Rights (ECtHR), may eventually use it to prioritise violations cases [1, 4]. Additionally, it has been argued that these type of systems will eventually be able reduce human error of the judges [6]. While each of these suggestions can be scrutinised from the legal perspective, it is still clear there are a large number of potential applications for a successful classification system.

While many of the currently proposed systems show promising results with a classification performance of about 80 percent correct, this is an overly optimistic view of their performance. One of the reasons for this is that classification performance is generally evaluated by predicting the outcome for a random subset of cases which were already known but not considered when creating the model. While this may seem fair, an arguably more interesting task is to predict future judgements.[1]

Importantly, however, all of the aforementioned studies claim to 'predict judicial decisions', which suggests these systems are able to predict (future) rulings on the basis of the available information. Unfortunately, classifying future judgements causes performance to suffer [9]. This lower performance may be caused by, for example, changes in the interpretation of the law, or new social

---

[1]It is important to note that predicting court judgements is a very different task from actual decision-making. The machine learning systems which are the focus of this study make pattern-based guesses on the basis of (sequences of) words in the text of a case. We discuss the ethical consideration for making this distinction later in the paper.

phenomena and developments due to changing societies. In addition, almost all classification systems rely on data about the case which is made available when the outcome of the case is known. Having knowledge about the outcome of a case may influence how the facts of the case are described (e.g., irrelevant facts for the outcome may be removed, or facts identified after an investigation and relevant to the outcome may be highlighted), compared to a situation in which the outcome would not have been known. This would mean that systems which use information composed when the outcome was not known may be disadvantaged compared to systems which use information extracted from documents composed when the outcome was known. One goal of this paper is to evaluate whether this indeed is the case.

A further goal of this paper is to propose making a distinction between *forecasting* judgements and *classifying* judgements. For the former, available textual data describing the (facts of the) case is required which was created *before* the decision was reached, so that the input of the forecasting system is not influenced by the outcome. For the latter, available textual data about the (facts of the) case was created *after* the decision was reached. Being explicit about this distinction is important, as many current studies in the field claim to 'predict judicial decisions', which suggests that they are forecasting future judgements, while instead they are classifying previously made judgements. For example, while Medvedeva et al. [9] predict the performance for future cases (by training a model on the basis of data from past cases), this is still a classification task as the input data is (a subset of) textual data which was created after the decision was reached.

Forecasting thus requires data related to a judgement that are published before the actual judgement was delivered. While the courts publish more and more case law every day [11], only little access is provided to documents that are available before the judgements were made. Forecasting future judgements is therefore a task which is impossible for many online available datasets. For this reason, the large majority of machine learning systems for legal data were built to provide a classification of court judgements, as opposed to forecasting judgements.

In this study, however, we concentrate on the ECtHR, as it publishes all of its final judgements online together with many supplementary documents, including admissibility cases, press-releases, summaries of cases, et cetera. Several of these documents were created before the decision was reached, and therefore this specific dataset enables both classification as well as forecasting of the judgements.[2] Besides evaluating whether forecasting is indeed a harder task than classification by evaluating

---

[2]To enable reproducibility, we provide our dataset containing pending ECtHR applications, as well admissibility decisions and final judgements of the court that can be used for a variety of tasks.

both sets of algorithms on the same cases, we aim to compare the relative performance of algorithms previously used for classifying court judgements, both for the task of classification as well as forecasting (using the the information published in the *communicated cases*; see Section 3.2). We do not introduce any new algorithms, as the purpose of this study is to determine the difference in performance for the two types of tasks.

In the following section we will discuss earlier work involving the latest attempts at classifying and forecasting court judgements. Section 3 is dedicated to describing the data we have used for our experiments and the larger dataset we release with this paper. In Section 4, we discuss various methods that can be used for forecasting decisions, their power and limitations. In Section 5, we report the results of the experiments that we have conducted for this study. In Section 6, we discuss the results and make suggestions regarding future work. Finally, in Sections 7 and 8, respectively, we make a note about ethical issues when conducting this type of research, and draw conclusions.

2. Related work

In this paper, we exclusively focus on the closed-class (often binary) tasks of outcome classification and outcome forecasting. These tasks are different from charge prediction, which predicts an open set of outcomes, such as the duration of a prison term [12, 13, 14, 15, 16, 17, 18].

While a growing number of courts share their data online, not many courts publish all of their cases online. Furthermore, for many published cases it may be hard to determine a binary or at least a small set of pre-set outcomes, making it hard to use the data from these courts for the type of machine learning models discussed in this study. The most recent papers that use machine learning approaches for classifying judicial decisions are therefore generally focusing on a limited number of courts, mainly including the US Supreme Court [3, 19, 20], the French Supreme Court [2, 21], and the European Court of Human Rights [5, 6, 7, 9]. A few other courts around the world have also been the focus of this type of analysis, including courts of the UK [22], Canada [8, 23], India [10], and Thailand [24].

There is a tradition of using statistical techniques to analyse the case law of the US Supreme Court (SCOTUS). The advantage of working with the SCOTUS database is that due to the attention it attracts, all trial data has systematically and manually been annotated with hundreds of variables by legal experts, shortly after the case has been tried. Katz et al. [19] used variables which are in principle available before SCOTUS reached its decision in an approach called extremely randomised trees to forecast the decision of the court. Their approach re-

sults in predicting 70% of the cases correctly, which is a somewhat lower performance than achieved by some of the state-of-the-art classification approached applied to data from other courts. However, as Katz et al. [19] performed the task of forecasting court decisions, rather than classifying court decisions, their lower performance may also be indicative of the potentially higher difficulty of forecasting.

Most of the courts in Europe, unfortunately, do not have the advantage of being able to generate such scrupulously annotated datasets, and often provide no access to all case law. For the European Court of Human Right a baseline model for classifying judgements using a so-called Support Vector Machine (SVM) on the basis of n-grams (i.e. sequences of one or more words extracted from the text) has been put forward by Medvedeva et al. [9]. Their model classified court decisions with an average accuracy of 75% for nine articles. Their work extended and corrected some data extraction issues (i.e. arguments of the court referencing the outcome were still included in the input training data, thereby resulting in overly optimistic performance) of an earlier study by Aletras et al. [1]. Additional work on the topic has been conducted by Chalkidis et al. [5], where they tested new methods and additional tasks, such as predicting the importance of a court case and identifying the articles that may (not) have been violated. While Chalkidis et al. [5] also trained the system on cases up to 2013 and tested on 2014-2018 (following the approach of Medvedeva et al. [25]), they extracted their data from the judgements, thereby making their approach a classification task instead of a forecasting task.

To our knowledge, only one study has tried to show that using documents from the early stages of the legal process may not always be as useful and *predictive* as final judgements. Specifically, Branting et al. [26] conducted experiments using statements from attorney misconduct complaints submitted to the Bar Association in the USA. The researchers set up a task of predicting whether the case would be investigated or closed. Using six different machine learning systems the authors showed that the text of the complaints themselves had very low predictive accuracy (maximum weighted f1-score: 0.52), and also adding additional metadata (i.e. extra information filled in during the complaint, attorney history, sentiment score, etc) was not very beneficial (maximum weighted f1-score: 0.55). Only data from later stages in the process, specifically allegation codes assigned by the intake staff substantially improved results (maximum weighted f1-score: 0.70). Nevertheless, these scores are still substantially lower than the scores reported by many studies classifying final decisions (see above). While Branting et al. [26] also deal with legal documents, they are not judicial decisions, but rather disciplinary proceedings conducted by the Bar Association, and therefore are not

directly comparable to the experiments conducted on court judgements.

There are currently only very few studies that focus on forecasting judgements, and most show a lower performance level than studies on judgement classification. Specifically Sharma et al. [20] and Katz et al. [19] forecast court decisions of the US Supreme Court. They reported an accuracy of around 70%. For courts in Europe, only Waltl et al. [27] forecast the outcome of appeal decisions involving German tax law (reporting a relatively low performance, with an average F-score of 0.57). Furthermore, Medvedeva et al. [28] forecast decisions on the basis of data from the ECtHR with their online system JURI (yielding an accuracy of around 70%).[3] The latter study is the approach we follow and extend in this paper. Specifically, we aim to investigate how the more advanced machine learning approaches of Chalkidis et al. [5] and Chalkidis et al. [29] perform when forecasting the ECtHR judgements.

3. Data

3.1. The Court

The European Court of Human Rights was established in 1959 as an international court that deals with individual and State applications claiming violation of various rights laid out in the European Convention on Human Rights (ECHR) [30, 31]. Applications are always brought by an individual/institution or multiple individuals/institutions against a State or multiple States that have ratified the Convention. No applications are considered between individuals, or from a State against an individual. Only five cases of a State against a State have been judged so far in the history of the Court. In 2020 the Court processed 41,700 applications, which were added to already pending applications. A total of 37,289 applications were dismissed based on the admissibility criteria, while the rest were decided by a Chamber or a Grand Chamber (762 cases based on 1,901 applications). From those cases, 880 were found to represent a violation of human rights. The majority of the documents produced by the court during the process are published online by the Court.[4]

3.2. Communicated cases

In order to describe the data that we use for our system it is important to clarify what the application process of the Court entails.

A resident of a country that ratified the ECHR can claim a potential violation within a certain time frame. The application is submitted via mail. On arrival, it is

registered by the Court and sent to the legal division that deals with the cases of a particular State, as they are familiar with the legislation of the country. Subsequently, the case is allocated to one of the Court's judicial formations.

Most of the cases are found inadmissible without meriting an investigation, due to not meeting the formal admissibility criteria. For example, often the application is dismissed because the applicant did not file the complaint within the required time frame. A decision regarding these cases is normally rendered by a single judge. If the application was not dismissed directly, the decision on admissibility is taken by a Committee of three judges (in case the Court has dealt with a number of similar cases before) or the Chamber of seven judges. In some cases admissibility decisions may even be made by the Grand Chamber (consisting of seventeen judges). Those usually concern the interpretation of the Convention itself, or if there is a risk of inconsistency compared to the previous judgements of the Court.

When an application is judged to be admissible based on formal parameters, the Chamber will examine its merits. Before doing so, the Court will *communicate* the application to the government that is the potential violator of the rights of the applicant (Rule 60 of the Court – Claims for just satisfaction). This is not done for all applications, but only for a part (approx. 15-20%). Such *communicated cases* contain the summary of the facts of the case, as well as questions to the government pertaining to the applicant's complaint. This document allows the government concerned to submit its observations on the matter of dispute. These documents are often communicated years before the case is judged, which provides a unique opportunity to use them for predicting the judgements of future cases. Moreover, the questions posted to the state often reflect on the Court's legal characterisation of the complaint. See, for instance, a question from a case of Arki against Hungary (application no. 10755/14, communicated on June 6, 2014):

> 1. Have the applicants been subjected to inhuman or degrading treatment on account of their cramped prison conditions, in breach of Article 3 of the Convention?

As a consequence, these documents can potentially be used to identify the facts or even (parts of) arguments related to certain judgements before those judgements are made.

Cases concerning repetitive issues do not merit a communicated case, and not every communicated case corresponds directly to a specific judgement. Multiple applications concerning the same events can be merged into a single case during the communication stage, but may be separated during final decision-making. Similarly, multiple applications can be communicated separately, but eventually judged together. Each year thousands of applications are communicated (i.e. 6,442 in 2019 and 7,681 in 2020). Only communicated cases from the year 2000 and later are available online. The Court decides on the order in which the cases are dealt with, based on the importance and urgency of the issues raised (Rule 41 of the Court – Order of Dealing with Cases)[5]. Therefore, the cases being judged may be mixed up and do not always respect the chronological order of when they were submitted.

For the machine learning systems created in our study, we will only use communicated documents that have judgements or have been found inadmissible based on merit for training and testing.

3.3. Data collection

We collected the data for this study in the following way. We scraped the ECtHR's 'HUDOC' website[6] and downloaded all communicated cases. We did the same for the judgements and admissibility decision documents, such as the admissibility cases from the Chamber and the Committee. We filtered the cases on the website to only download English versions of the documents. As the filter did not always work adequately, we also filtered using Google's language detection (langdetect) library.[7] In addition, we extracted all available metadata, such as the application number, state, importance level, et cetera. We used the application number of each communicated case to link the associated documents to corresponding admissibility decisions and judgements. We then extracted the conclusion of the court proceedings ('violation' or 'no violation'), as well as the facts of the cases from the judgement text. We use these facts in a classification model, so we can compare its performance to the performance of a forecasting model using data from the communicated cases.

While the facts in communicated cases are the summary of the events as described by the applicant, the facts that end up in the final judgement are compiled after the investigation and therefore also include the side of State. We only use the facts of the case from the final judgements since these are most comparable to the communicated cases. Specifically, these have also been argued to potentially be available before the outcome was reached [9] and do not contain references to the outcome [9, 5]. This also mirrors the set up in Chalkidis et al. [5] that we follow.

To enable a fair comparison, the cases (but not the extracted information about these cases) used for training and testing are identical for both models. We assume that cases that were found to be *inadmissible based on*

merit are similar to cases that were judged as having no violation. From a legal point of view, these cases can be characterised as simply more clear 'non-violation' cases. The court has made judgements on similar applications many times before, and hence these do not merit a full judgement. For cases that went though to the final judgement stage, we assign the 'violation' label for all cases that were judged to show a violation of at least one article of the ECHR.

As we mentioned before, individual communicated cases do not always directly correspond to unique cases which received a judgement or admissibility decision, as communicated cases can be split or merged during the process. For the split cases, the assigned label of the associated communicated case depended on whether any of the split cases resulted in a violation of at least one article ('violation' label), or not ('non-violation' label, i.e. none of the split cases exhibits a violation of any article). To ensure the set of cases considered for the classification task and the forecasting task is identical, we randomly selected a single judged case (from the associated split cases) where the assigned label matched that of the assigned label to the communicated case. For judgements associated with multiple merged communicated cases, we randomly chose one of the communicated cases and removed the rest. Finally, duplicate cases and judged cases which did not have (correctly formatted) facts were excluded from the dataset used for both tasks. In this way, the set of cases considered for the classification task and the forecasting task is identical.

Subsequently, we split the data into training and test sets (on average a 77%-23% split). We trained each system three times, with different setups (with a decreasing amount of training data) to assess the robustness of the results. Setup 1 concerns model training with cases that received judgement in years 2000-2019, whereas model testing was conducted with cases that received judgement in the year 2020. Setup 2 uses 2000-2018 data for training and 2019 data for testing. Setup 3 uses 2000-2017 data for training and 2018 data for testing. Each setup is used once for forecasting judgements using data from the communicated cases, and once for classifying judgements using data from the facts extracted from the final judgement. As in each setup the number of violation cases exceeded the number of non-violation cases, we balanced the training set in each setup by removing older violation cases until the same number of documents was present for each label. Table 1 shows the number of documents available for training and testing for each setup.

The data used for the two different tasks differs somewhat. For the communicated cases, we used all data available (i.e. the facts and the questions as they were presented in the text), whereas for the judgements, we only used data from the facts section. In general, the

average number of words associated with the extracted facts from each judgement are not much higher (i.e. 2000 words) than the number of words of the associated communicated case (i.e. 1800 words).

Table 1
Distribution of training and testing data for different setups.

	setup 1: 2020	setup 2: 2019	setup 3: 2018
train (balanced)	2264	1806	1386
test (no violation)	167	229	210
test (violation)	342	311	309

3.4. Dataset

In addition to the data used in this study, we have extracted data for a large set of additional cases, which were not taken into account in our analysis. This dataset is released together with this paper.[8] Specifically, this dataset contains all of the communicated cases, admissibility cases and final judgements of the Court published between 1960 and 2020. We provide raw text, the metadata (e.g., date, court-assigned importance, parties, and section) as well as the preprocessed text of communicated cases (split into facts and questions), admissibility decisions (extracted facts) and final judgements (split into sections: Procedure, Facts, Relevant domestic law, Law - including arguments of the court, Outcome, and Dissenting opinions) in order to facilitate further research in ECtHR judgement forecasting and classification. In addition, the case numbers are linked throughout each stage of the court proceedings (where applicable). This dataset may be suitably used for a number of classification tasks in legal analysis, including judgement classification based on facts (using Facts and possibly Procedure sections) and/or arguments (using the Law sections).

4. Methodology

As we mentioned before, the approach most relevant work for our study is that of Chalkidis et al. [5]. Specifically, in one of their tasks they focused on classifying the court judgements depending on whether at least one article of the ECHR was violated or not.[9] In addition, they experimented with using anonymized vs. non-anonymized

---

[8]https://drive.google.com/drive/folders/1ltIpHlcqcRlT_JDebHsyLgvgoa4Vbxo8?usp=sharing

[9]The purpose of their Chalkidis et al. [5] second task was to identify all of the violated articles for a single court document (i.e. multi-label classification). However, as the involved articles are known as soon as the application is submitted, it is not clear what the practical use is of predicting the list of articles potentially violated. A realistic scenario for the ECtHR would only involve deciding whether or not a *given* article was violated.

data. While we perform the same task as Chalkidis et al. [5], enabling us to benefit from more data than when we would predict (non-)violation per article separately, we only use non-anonymized data. For the anonymized setup, Chalkidis et al. [5] have removed *named entities* (such as names or locations) from the text to make sure the model was not biased towards demographic information. While removing this potential bias is understandable when building a decision-making system, forecasting or classifying judgements is different. Specifically, given that locations may offer relevant information about the case (i.e. some countries are notorious violators of specific rights), models used for forecasting or classification benefit from keeping this information (also known to judges) in.

In our study, we implement three systems used by Chalkidis et al. [5] and compare their performance on the classification and forecasting task. Specifically, we include the SVM model, the Hierarchical-BERT (H-BERT) model and the LEGAL-BERT model (see below for more details). All models were re-created on the basis of the description provided by Chalkidis et al. [5] and Chalkidis et al. [29]. As not all settings and (hyper)parameters were specified in their paper, our reproduction of their models may be slightly different. However, we believe these differences to be minor. Our goal is to see how some of the state-of-the-art models which have been shown to perform very well when applied to final judgements of the ECtHR perform when they are only being provided with data from the applicants to the ECtHR (i.e. victims of a alleged human right violation).

Our SVM classifier is a Linear SVC model including 1-5 n-grams. For a detailed explanation about text classification using machine learning (including Linear SVC), see Medvedeva et al. [9].

BERT or Bidirectional Encoder Representations from Transformers [32] is a popular pre-trained transformer-based [33] machine-learning technique resulting in a so-called language model. The method also allows fine-tuning the language model for a specific task, i.e. adapting the pre-trained model to the target task, in our case classifying and forecasting ECtHR judgements.

To use BERT on long case documents without having a maximum text length restriction, H-BERT [5] processes each fact separately and combines them by using a self-attention layer to generate an embedding for a case. This resulting embedding is then used for classification and forecasting.[10] Instead of the standard BERT model (which [5] reported to have sub-par performance), we used LEGAL-BERT [29] in our experiments. LEGAL-BERT is a BERT model which was pre-trained on legal

texts from different sources.

BERT and many of its variations, including H-BERT, have shown to result in substantial improvements compared to the state-of-the-art in a large variety of text classification tasks. Specifically, Chalkidis et al. [5] have shown that using H-BERT resulted in a very high performance (macro F-score of 0.82) for the binary task (violation of at least 1 article of ECHR vs. no violation), and an even higher macro F-score of 0.83 for LEGAL-BERT on the same dataset [29].

In the following, we report the results per class for each model. Our main evaluation metric is the macro F-score. This measure can be described as a mean of the average precision and recall across all classes (i.e. 'violation' and 'no violation'). *Precision* is the percentage of cases given a certain label (i.e., 'violation' or 'no violation') that was correct. *Recall* is the percentage of cases having a certain (correct) label, that were identified as such.[11]

5. Results

We started our experiments with setup 1, by testing on all data from 2020. To our surprise, results for classifying the final judgements were very low compared to Chalkidis et al. [5] (see tables 2 and 3 for the performance per class). In contrast to our expectations, forecasting final judgements on the basis of communicated cases instead of on the basis of the facts of the final judgements yielded better results when using H-BERT. Compared to Chalkidis et al. [5], however, since not all cases are communicated by the court, our training set was much smaller (2264 cases vs. 7100 cases, respectively).

Table 2
Performance (precision, recall, f1-score and accuracy) for Linear SVC, H-BERT, and LEGAL-BERT models per class for final judgement classification, trained on cases between 2000 and 2019 and tested on cases decided in 2020

2020 - Final judgements		P	R	F1	#
SVM	no viol.	0.46	0.93	0.62	167
	violation	0.93	0.46	0.62	342
	macro avg.	**0.70**	**0.70**	**0.62**	509
	accuracy			0.62	509
H-BERT	no viol.	0.42	0.92	0.58	167
	violation	0.91	0.38	0.53	342
	macro avg.	**0.66**	**0.65**	**0.56**	509
	accuracy			0.56	509
LEGAL-BERT	no viol.	0.42	0.90	0.58	167
	violation	0.89	0.40	0.55	342
	macro avg.	**0.66**	**0.65**	**0.57**	509
	accuracy			0.57	509

---

[10]While BERT can process each case by including up to 512 tokens (i.e. meaningful word parts), our H-BERT implementation can use up to 1024 tokens (i.e. 128 tokens for each of the the first eight facts).

[11]Exact definition of the F-score can be found here: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Table 3

Performance (precision, recall, f1-score and accuracy) for Linear SVC, H-BERT, and LEGAL-BERT models per class for forecasting judgements, trained on communicated cases between 2000 and 2019 and tested on communicated cases that received a judgement in 2020.

2020 - Communicated cases		P	R	F1	#
SVM	no viol.	0.47	0.51	0.49	167
	violation	0.75	0.72	0.73	342
	macro avg.	**0.61**	**0.61**	**0.61**	509
	accuracy			0.65	509
H-BERT	no viol.	0.45	0.61	0.52	167
	violation	0.77	0.63	0.69	342
	macro avg.	**0.61**	**0.62**	**0.60**	509
	accuracy			0.62	509
LEGAL-BERT	no viol.	0.42	0.54	0.47	167
	violation	0.74	0.63	0.68	342
	macro avg.	**0.58**	**0.58**	**0.57**	509
	accuracy			0.60	509

Table 5

Performance (precision, recall, f1-score and accuracy) for Linear SVC, H-BERT, and LEGAL-BERT models per class for forecasting judgements, trained on communicated cases between 2000 and 2018 and tested on communicated cases that received a judgement in 2019.

2019 - Communicated cases		P	R	F1	#
SVM	no viol.	0.62	0.53	0.57	229
	violation	0.69	0.77	0.73	311
	macro avg.	**0.66**	**0.65**	**0.65**	540
	accuracy			0.67	540
H-BERT	no viol.	0.57	0.67	0.61	229
	violation	0.72	0.63	0.67	311
	macro avg.	**0.64**	**0.65**	**0.64**	540
	accuracy			0.65	540
LEGAL-BERT	no viol.	0.55	0.50	0.52	229
	violation	0.66	0.70	0.68	311
	macro avg.	**0.60**	**0.60**	**0.60**	540
	accuracy			0.61	540

However, when trying setup 2, where we trained using less data (i.e. until 2018) and tested on all data of 2019, results were as expected. Specifically, macro F-scores ranged between 0.79 and 0.92 for the classification task (see Table 4), and performance was much lower for the forecasting task with macro F-scores ranging from 0.60 to 0.65 (see Table 5).

Table 6

Performance (precision, recall, f1-score and accuracy) for Linear SVC, H-BERT, and LEGAL-BERT models per class for final judgement classification, trained on cases between 2000 and 2017 and tested on cases decided in 2018.

2018 - Final judgements		P	R	F1	#
SVM	no viol.	0.67	0.91	0.77	210
	violation	0.92	0.70	0.79	309
	macro avg.	**0.79**	**0.80**	**0.78**	519
	accuracy			0.78	519
H-BERT	no viol.	0.86	0.72	0.78	210
	violation	0.83	0.92	0.87	309
	macro avg.	**0.84**	**0.82**	**0.83**	519
	accuracy			0.84	519
LEGAL-BERT	no viol.	0.88	0.78	0.83	210
	violation	0.86	0.93	0.89	309
	macro avg.	**0.87**	**0.85**	**0.86**	519
	accuracy			0.87	519

Table 4

Performance (precision, recall, f1-score and accuracy) for Linear SVC, H-BERT, and LEGAL-BERT models per class for final judgement classification, trained on cases between 2000 and 2018 and tested on cases decided in 2019.

2019 - Final judgements		P	R	F1	#
SVM	no viol.	0.69	0.95	0.80	229
	violation	0.95	0.68	0.79	311
	macro avg.	**0.82**	**0.81**	**0.79**	540
	accuracy			0.79	540
H-BERT	no viol.	0.90	0.92	0.91	229
	violation	0.94	0.93	0.93	311
	macro avg.	**0.92**	**0.92**	**0.92**	540
	accuracy			0.92	540
LEGAL-BERT	no viol.	0.87	0.90	0.88	229
	violation	0.92	0.90	0.91	311
	macro avg.	**0.90**	**0.91**	**0.90**	540
	accuracy			0.90	540

To determine which of the two setups resulted in representative results, we conducted a final experiment (setup 3), by training with even less data (i.e. until 2017) and testing on all data of 2018. The results showed a similar pattern (with slightly better performance, despite the reduced amount of training data) as the results of setup 2 (2019). See Tables 6 and 7 for an overview of these results.

When running the same experiments using successively smaller datasets (i.e. testing on data from 2017, and 2016), the same pattern is visible as for setups 2 and 3. That is, performance when classifying final judgements is much higher than when forecasting final judgements. Table 8 shows the macro F-scores for both tasks for all years (of the test set) ranging from 2016 to 2020 and all three algorithms. Besides showing that classification performance is generally (except for 2020) higher than forecasting performance, these results also show that while H-BERT and LEGEL-BERT generally outperforms SVM in classification (except for 2020), they do not improve over SVM in forecasting.

Table 7
Performance (precision, recall, f1-score and accuracy) for Linear SVC, H-BERT, and LEGAL-BERT models per class for forecasting judgements, trained on communicated cases between 2000 and 2017 and tested on communicated cases that received a judgement in 2018.

2018 - Communicated cases		P	R	F1	#
SVM	no viol.	0.62	0.55	0.58	210
	violation	0.72	0.77	0.74	309
	macro avg.	**0.67**	**0.66**	**0.66**	519
	accuracy			0.68	519
H-BERT	no viol.	0.60	0.63	0.61	210
	violation	0.73	0.71	0.72	309
	macro avg.	**0.67**	**0.67**	**0.67**	519
	accuracy			0.68	519
LEGAL-BERT	no viol.	0.59	0.52	0.55	210
	violation	0.69	0.75	0.72	309
	macro avg.	**0.64**	**0.63**	**0.64**	519
	accuracy			0.66	519

6. Discussion

Our results clearly show that our intuition regarding the increased difficulty of the task of forecasting judgements as opposed to classifying judgements is confirmed. However, the tasks are conceptually very different, and therefore comparing them in terms of accuracy may not be entirely fair. Nevertheless, both fall under 'predicting court decisions' in the existing literature. Our results illustrate that predicting court decisions which have not been made yet is a much harder task than current academic research may suggest.

One potential explanation for the higher performance of the classification approach compared to the forecasting approach may be the higher amount of data (i.e. an average of 2000 words for the facts part of the judgement versus 1800 words for the communicated case). Since LEGAL-BERT and H-BERT have a limited input length of up to 512 or 1024 tokens (respectively), this difference will not play a role for these models. However, this is different for the SVM which does not have such a limit. Consequently, we evaluated an SVM on the 'shortened' facts of the final judgements. Specifically, we removed the facts from the middle of the text (under the presumption that the most important information is present at the beginning and at the end) until the text was approximately the same length as the text of the the corresponding communicated case. This change, however, did not affect the performance, as the SVM on this trimmed data yielded macro F-scores of 0.61, 0.83 and 0.77 for 2020, 2019, and 2018, respectively (compared to 0.62, 0.79 and 0.78). This suggests that the facts are formulated in a way that is affected by the final ruling, rather than that there is a tangible benefit of the higher amount of data.

The SVM model allows us to inspect the top coefficients (weights) of n-grams assigned by the system. We observe that for final judgements the system often prioritises longer n-grams (the average length for the 100 top features is 2.475), while for communicated cases it prioritises unigrams and common collocations consisting of two words such as public prosecutor or minor offences (the average length for the 100 top features is 1.405).

We should also take into account that the communicated case is a summary of an applicant's complaint. As a result, it only reflects this party's side of the events, and may be subjective and incomplete. After sending the communicated case to the State involved, the Court conducts the investigation and inspects the side of the State as well. Consequently, the final judgement contains a more thorough and objective description of the facts that takes the sides of both parties into account. This explains why the facts available in communicated cases can differ considerably from the set of facts presented in the final judgement.

This bias towards a violation of human rights can also be observed in the results. For the forecasting task, all models show a higher performance when predicting the 'violation' label than when predicting the 'non-violation' label (see Tables 3, 5, and 7). In contrast, the gap in performance when predicting the two labels for the classification task is considerably smaller (see Tables 2, 4, and 6), which confirms the intuition that the description of the facts in final judgements are a better representation of the events and therefore better predictors of the outcome. Nevertheless, for the 2018 and 2019 data, the performance predicting the 'violation' label using the communicated cases data (i.e. the forecasting task) is still lower than the overall performance (or the 'violation' label performance) using the extracted facts from the final judgements (i.e. the classification task).

The only case when forecasting judgements shows a higher performance than classifying judgements is on the 2020 test data. However, this is caused by the much lower than usual classification performance. Unfortunately, we have no explanation for this pattern, despite the effort we spent on trying to investigate whether the 2020 data showed deviating patterns compared to the data from earlier years. For example, the average length of the 2020 cases, and overall vocabulary is consistent with the previous years, as well as the distribution of cases between different States and therefore different Chambers. The court has judged only slightly (4%) fewer cases in 2020 than in 2019, and did not adopt any new policies compared to the previous years. There is no indication that the court used a different selection approach for the cases it ruled on. Since the cases originated in the years before 2020, it is also unlikely that this pattern has any relationship with human rights violations related to COVID-19. Finally, the format of case law has also

Table 8

Macro F-scores for Linear SVC, H-BERT, and LEGAL-BERT models for both tasks between 2016 and 2020, including size of training and testing sets.

F-score (macro)					
	2020	2019	2018	2017	2016
Training set size	2264	1806	1386	976	640
Test set size	509	540	519	503	447
SVM (forecasting)	0.61	0.65	0.66	0.65	0.64
H-BERT (forecasting)	0.60	0.64	0.67	0.66	0.66
LEGAL-BERT (forecasting)	0.57	0.60	0.64	0.64	0.58
SVM (classification)	0.62	0.79	0.78	0.78	0.75
H-BERT (classification)	0.56	0.92	0.83	0.84	0.82
LEGAL-BERT (classification)	0.57	0.90	0.86	0.84	0.82

remained the same. For now, we are therefore forced to consider performance on 2020 data (as so many other things in 2020) an anomaly. Whether this deviating pattern will continue in 2021 remains to be evaluated.

6.1. Future work

We have discussed a range of approaches to forecast outcomes of pending applications. Each of these could be improved through more careful tuning, pre-processing, data selection, feature selection, et cetera. Furthermore, additional classification or forecasting algorithms could be used as well. However, this was not the goal of the present paper. By releasing our dataset, together with a number of baselines reported in this paper, we hope to have provided a new starting point for the task of *forecasting* ECtHR judgements.

Regarding future research, it would be interesting to assess whether selecting the last tokens, or tokens from specifically chosen facts would be beneficial for BERT-like models. For example, these models might yield better results as initial facts generally are about the procedure and the applicant themselves, while facts from the end of the document often are more closely related to the events relating to the alleged violation of human rights. Due to limited available data, we have only investigated whether or not a case violated *any* article of the ECHR. However, it would be interesting to assess how the difference in performance between forecasting and classification would be affected when individual articles are investigated.

While we can forecast pending applications using data from communicated cases, this does not allow us to forecast the judgements for any future cases as this data may not always be available (e.g., not all cases are communicated to the State). Forecasting using other data available before the judgement is known (i.e. from other sources) may likely be even harder, as the uniform documents created by the court for the communicated cases are likely beneficial.

While predicting judgements is an interesting task in itself, it is beneficial to also gain insight into how the system reaches a certain outcome, and therefore to take a step toward explainable AI [34, 35] and large-scale automatic legal analysis. This requires, for example, understanding which facts lead to which judgement. Particularly for the classification task, where determining a judgement of the court which is already known is of no practical use, determining the basis of the classification is important.

Several methods that are often used in classification tasks allow determining the classification basis (to some extent). Linear SVC, for example, allows the inspection of its coefficients to evaluate which words and phrases are more characteristic for assigning one class than another. Medvedeva et al. [28] also suggest evaluating such a system at the sentence level to identify and highlight the sentences that have the highest probability of belonging to a specific class. Furthermore, the architecture of H-BERT, for example, allows one to assess which of the eight included facts (or questions) had the largest impact on classification on the basis of so-called attention [33]. Unfortunately LEGAL-BERT by itself cannot be used for this. While it often produces very high scores, especially for final judgement classification, and it may function as a good reference point for high classification performance, one cannot see within the black box.

7. Ethical considerations

We believe it is important to emphasise that our goal with this work is only to (try to) forecast and classify court judgements. Our interest is scientific and is focused on assessing whether Natural Language Processing systems are able to identify certain patterns in legal judgements. We do not think that any of the models described in this paper can or should be used for *making decisions* in courts, especially those where human rights are at stake (which concerns the majority of the courts around the world). Moreover, we are opposed to the use of such models in other high-stakes situations, due to the inability of these

types of models to deal with new legal developments and interpretations, previously unobserved issues [36, 37], lacking transparency [38, 39, 40], and cybersecurity concerns [41].

8. Conclusion

In this paper we have proposed to make a distinction between forecasting court judgements and classifying judgements. Forecasting judgements is based on data which is available before the outcome is known (such as the communicated cases of the ECtHR), whereas classifying judgements is based on (a subset of) data compiled when the outcome was known (such as the facts from the ECtHR ruling). Making this distinction is important, as earlier research [26], and the experiments conducted in this paper show that performance seems to be substantially lower when forecasting future judgements compared to classifying decisions which were already made by the court, and the terminology of current papers (i.e. 'predicting court judgements') suggests a forecasting task whereas it actually most often is a task of classifying final judgements. Classification performance should therefore not be used as an indication of how well these types of systems are able to forecast judgements of the court. Interestingly, while more sophisticated models appeared to be beneficial for the simpler classification task, this was not the case for the harder forecasting task.

References

[1] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, V. Lampos, Predicting judicial decisions of the european court of human rights: A natural language processing perspective, PeerJ Computer Science 2 (2016).

[2] O.-M. Şulea, M. Zampieri, M. Vela, J. van Genabith, Predicting the law area and decisions of French Supreme Court cases, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, INCOMA Ltd., Varna, Bulgaria, 2017, pp. 716–722. URL: https://doi.org/10.26615/978-954-452-049-6_092. doi:10.26615/978-954-452-049-6_092.

[3] A. Kaufman, P. Kraft, M. Sen, Machine learning, text data, and supreme court forecasting, Project Report, Harvard University (2017).

[4] C. O'Sullivan, J. Beel, Predicting the outcome of judicial decisions made by the european court of human rights, in: AICS 2019 - 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science., 2019.

[5] I. Chalkidis, I. Androutsopoulos, N. Aletras, Neural legal judgment prediction in English, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4317–4323. URL: https://www.aclweb.org/anthology/P19-1424. doi:10.18653/v1/P19-1424.

[6] A. Kaur, B. Bozic, Convolutional neural network-based automatic prediction of judgments of the european court of human rights, in: AICS, 2019.

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

[8] O. Salaün, P. Langlais, A. Lou, H. Westermann, K. Benyekhlef, Analysis and multilabel classification of quebec court decisions in the domain of housing law, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2020, pp. 135–143.

[9] M. Medvedeva, M. Vols, M. Wieling, Using machine learning to predict decisions of the european court of human rights, Artificial Intelligence and Law 28 (2020) 237–266.

[10] R. A. Shaikh, T. P. Sahu, V. Anand, Predicting outcomes of legal cases based on legal factors using classifiers, Procedia Computer Science 167 (2020) 2393–2402.

[11] M. Marković, S. Gostojić, Open judicial data: A comparative analysis, Social Science Computer Review (2018).

[12] B. Luo, Y. Feng, J. Xu, X. Zhang, D. Zhao, Learning to predict charges for criminal cases with legal basis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2727–2736. URL: https://www.aclweb.org/anthology/D17-1289. doi:10.18653/v1/D17-1289.

[13] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, M. Sun, Legal judgment prediction via topological learning, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3540–3549. URL: https://www.aclweb.org/anthology/D18-1390. doi:10.18653/v1/D18-1390.

[14] X. Jiang, H. Ye, Z. Luo, W. Chao, W. Ma, Interpretable rationale augmented charge prediction system, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Santa Fe, New Mexico, 2018, pp. 146–151. URL: https://www.aclweb.org/anthology/C18-2032.

[15] Y. Li, T. He, G. Yan, S. Zhang, H. Wang, Using case facts to predict penalty with deep learning, in: International Conference of Pioneering Computer

Scientists, Engineers and Educators, Springer, 2019, pp. 610–617.

[16] H. Ye, X. Jiang, Z. Luo, W. Chao, Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1854–1864. URL: https://www.aclweb.org/anthology/N18-1168. doi:10.18653/v1/N18-1168.

[17] H. Chen, D. Cai, W. Dai, Z. Dai, Y. Ding, Charge-based prison term prediction with deep gating network, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6361–6366. URL: https://www.aclweb.org/anthology/D19-1667. doi:10.18653/v1/D19-1667.

[18] W. Chao, X. Jiang, Z. Luo, Y. Hu, W. Ma, Interpretable charge prediction for criminal cases with dynamic rationale attention, Journal of Artificial Intelligence Research 66 (2019) 743–764.

[19] D. M. Katz, M. J. Bommarito II, J. Blackman, A general approach for predicting the behavior of the supreme court of the united states, PloS one 12 (2017).

[20] R. D. Sharma, S. Mittal, S. Tripathi, S. Acharya, Using modern neural networks to predict the decisions of supreme court of the united states with state-of-the-art accuracy, in: International Conference on Neural Information Processing, Springer, 2015, pp. 475–483.

[21] O.-M. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, J. Van Genabith, Exploring the use of text classification in the legal domain, in: In Proceedings of the 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL 2017), 2017.

[22] B. Strickson, B. De La Iglesia, Legal judgement prediction for uk courts, in: Proceedings of the 2020 The 3rd International Conference on Information Science and System, 2020, pp. 204–209.

[23] H. Westermann, V. R. Walker, K. D. Ashley, K. Benyekhlef, Using factors to predict and analyze landlord-tenant decisions to increase access to justice, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 133–142. URL: https://doi.org/10.1145/3322640.3326732. doi:10.1145/3322640.3326732.

[24] K. Kowsrihawat, P. Vateekul, P. Boonkwan, Predicting judicial decisions of criminal cases from thai supreme court using bi-directional gru with attention mechanism, in: 2018 5th Asian Conference on Defense Technology (ACDT), IEEE, 2018, pp. 50–55.

[25] M. Medvedeva, M. Vols, M. Wieling, Judicial decisions of the european court of human rights: Looking into the crystal ball, in: Proceedings of the Conference on Empirical Legal Studies, 2018.

[26] K. Branting, C. Balhana, C. Pfeifer, J. Aberdeen, B. BROWN, Judges are from mars, pro se litigants are from venus: Predicting decisions from lay text, in: Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020, volume 334, IOS Press, 2020, p. 215.

[27] B. Waltl, G. Bonczek, E. Scepankova, J. Landthaler, F. Matthes, Predicting the outcome of appeal decisions in germany's tax law, in: International Conference on Electronic Participation, Springer, 2017, pp. 89–99.

[28] M. Medvedeva, X. Xu, M. Wieling, M. Vols, Juri says: Prediction system for the european court of human rights, in: Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020, volume 334, IOS Press, 2020, p. 277.

[29] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.261. doi:10.18653/v1/2020.findings-emnlp.261.

[30] S. Greer, J. Gerards, R. Slowe, Human Rights in the Council of Europe and the European Union: Achievements, Trends and Challenges, Cambridge Studies in European Law and Policy, Cambridge University Press, 2018. doi:10.1017/9781139179041.

[31] D. J. Harris, M. O'Boyle, E. Bates, C. Buckley, Harris, O'Boyle & Warbrick: Law of the European convention on human rights, Oxford University Press, USA, 2014.

[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkor-

eit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[34] T. Bench-Capon, The need for good old fashioned ai and law, International Trends in Legal Informatics: A Festschrift for Erich Schweighofer. Editions Weblaw, Bern (2020) 23–36.

[35] J. Collenette, K. Atkinson, T. Bench-Capon, An explainable approach to deducing outcomes in european court of human rights cases using adfs, Frontiers in Artificial Intelligence and Applications 326 (2020) 21–32.

[36] R. W. Campbell, Artificial intelligence in the courtroom: The delivery of justice in the age of machine learning, Colo. Tech. LJ 18 (2020) 323.

[37] R. Berk, Berk, Drougas, Machine learning risk assessments in criminal justice settings, Springer, 2019.

[38] A. Završnik, Criminal justice, artificial intelligence systems, and human rights, in: ERA Forum, volume 20, Springer, 2020, pp. 567–583.

[39] F. Thomsen, Iudicium ex machinae – the ethical challenges of automated decision-making in criminal sentencing, in: J. Roberts, J. Ryberg (Eds.), Principled Sentencing and Artificial Intelligence, Oxford: Oxford University Press, forthcoming.

[40] A. Deeks, N. Lubell, D. Murray, Machine learning, artificial intelligence, and the use of force by states, J. Nat'l Sec. L. & Pol'y 10 (2019) 1.

[41] P. M. Nichols, Bribing the machine: Protecting the integrity of algorithms as the revolution begins, American Business Law Journal 56 (2019) 771–814.