Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Natural language processing in law: Prediction of outcomes in the higher courts of Turkey

Emre Mumcuoğlu [a], Ceyhun E. Öztürk [a], Haldun M. Ozaktas [a], Aykut Koç [a,b,*]

[a] *Electrical and Electronics Engineering Department, Bilkent University, Ankara, Turkey*
[b] *UMRAM, Bilkent University, Ankara, Turkey*

ARTICLE INFO

Keywords:
Natural language processing
Law
Machine learning
AI in law
Legal text mining
Deep learning

ABSTRACT

Natural language processing (NLP) based approaches have recently received attention for legal systems of several countries. It is of interest to study the wide variety of legal systems that have so far not received any attention. In particular, for the legal system of the Republic of Turkey, codified in Turkish, no works have been published. We first review the state-of-the-art of NLP in law, and then study the problem of predicting verdicts for several different courts, using several different algorithms. This study is much broader than earlier studies in the number of different courts and the variety of algorithms it includes. Therefore it provides a reference point and baseline for further studies in this area. We further hope the scope and systematic nature of this study can set a framework that can be applied to the study of other legal systems. We present novel results on predicting the rulings of the Turkish Constitutional Court and Courts of Appeal, using only fact descriptions, and without seeing the actual rulings. The methods that are utilized are based on Decision Trees (DTs), Random Forests (RFs), Support Vector Machines (SVMs) and state-of-the-art deep learning (DL) methods; specifically Gated Recurrent Units (GRUs), Long Short-Term Memory networks (LSTMs) and bidirectional LSTMs (BiLSTMs), with the integration of an attention mechanism for each model. The prediction results for all algorithms are given in a comparative and detailed manner. We demonstrate that outcomes of the courts of Turkish legal system can be predicted with high accuracy, especially with deep learning based methods. The presented results exhibit similar performance to earlier work in the literature for other languages and legal systems.

1. Introduction

Law constantly grows and evolves to meet emerging and changing needs of societies in response to social, political, economic, and technological changes. Both ever-continuing transformations in legislation and the rapid increase in the number of precedent cases encumber an ever-growing burden on law professionals. This naturally leads to the question of whether any machine assistance can be provided in the field. Such a system would facilitate the job of lawyers, prosecutors and judges, as well as other related professionals, and may contribute positively to the greater good of the public by saving time, reducing error, and improving consistency. Computers can quickly scan and analyze vast amounts of legal text.

Natural language processing (NLP) has been used successfully in many information science applications related to the social sciences (Ji et al., 2020a; Junqué de Fortuny et al., 2014; Li et al., 2020; Qian et al., 2019; Schumaker & Chen, 2009; Tuke et al., 2020). Heavily relying on the written word, law is one of the fields that can greatly benefit from NLP, (Aletras et al., 2016; Ikram

[*] Corresponding author.
E-mail address: aykut.koc@bilkent.edu.tr (A. Koç).

https://doi.org/10.1016/j.ipm.2021.102684
Received 8 September 2020; Received in revised form 21 April 2021; Accepted 27 June 2021
Available online 12 July 2021

& Chakir, 2019; Katz et al., 2017; Kowsrihawat et al., 2018; Long et al., 2019; Şulea et al., 2017a, 2017b; Virtucio et al., 2018). A history of AI and law has been superbly presented in the survey given by Bench-Capon et al. (2012). A comprehensive overview of the NLP applications in the legal domain can be found in the survey of Chalkidis and Kampas (2018).

1.1. Research objectives

To the best of our knowledge, the legal system of Turkey has not been the subject of NLP-based approaches. An important problem is to investigate the effectiveness of machine learning models in predicting case outcomes. The aim of the work described in this paper is to predict the rulings of Turkish higher courts by looking only at the fact descriptions that are provided. Through this study, we hope to lay groundwork for future research on the legal system of Turkey and provide a baseline with which future work can be compared to.

The judicial branch of government is one of the three separate powers of the Republic of Turkey, along with the legislative and executive, all of whose authority is based on the Constitution. The Constitutional Court oversees them and ensures conformance to the constitution. With some simplification, and with the exception of the Court of Jurisdictional Disputes, courts can be categorized under two main categories: judicial courts and administrative courts. Both categories include first instance courts and the District Courts of Appeal. Above them, with the highest authority, are the Court of Cassation for judicial courts and the Council of State for administrative courts. Each higher court has the authority to alter or remove the ruling of a lower court in the hierarchy (Ansay & Wallace, 2005). Our focus is on the rulings of higher courts, specifically on the District Courts of Appeal and the Constitutional Court. For the others, either data is not available or verdicts do not include case descriptions (see Section 3 on how we divide and process the data).

The secondary aim of our research is related to the experimentation methodology. There are a number of existing studies in the literature for legal case text classification and predicting case outcomes for different legal systems all around the world (Aletras et al., 2016; Ikram & Chakir, 2019; Katz et al., 2017; Kowsrihawat et al., 2018; Long et al., 2019; Şulea et al., 2017a, 2017b; Virtucio et al., 2018). In these studies, machine learning methods have been deployed to predict the court rulings. However, this literature is quite fragmented in that most of these studies are limited in scope in that they usually consider only one type of court and one machine learning method. Given that the legal systems they study are already different, this makes it difficult to make any comparative studies or derive generalizations. Therefore, in this study we also aimed to cover as many courts as possible (being limited by available data) and compare several machine learning methods. Thus, to the best of our knowledge, we not only provide the first application of these methods to the legal system of Turkey, but also do so in a systematic and comprehensive way.

Our work should not only provide a baseline for further studies of Turkish legal system, but can also provide a framework within which other legal systems can be analyzed and then compared with each other. To facilitate comparative studies that can lead to useful generalizations, collection of data from different types and levels of courts and also applying several learning methods should be performed. To that end, we strive for a framework consisting of (i) systematic separation of the legal system corpus to sub-corpora according to the different types and levels of courts, (ii) a reproducible method of pre-processing data that makes it suitable for further higher-level processing, (iii) performing experiments to characterize the performances of baseline, classical machine learning approaches like SVMs and random forests, and several contemporary deep learning based methods with and without attention mechanism. This will help further research to increase performance in a systematical and comparative manner. While some aspects of our approach are specific to the legal system of Turkey, the general framework is mostly applicable to the study of other legal systems as well.

The NLP application to legal domain, that our proposed methodology focuses on, is to predict the outcomes of cases by looking only at the description of facts written by the court. The courts considered were District Courts of Appeal and the Constitutional Court. This is because the decisions of first instance courts, in addition to being mostly unavailable, cannot be predicted easily due to complicated verdicts and the possibility of many penalties in a single case, whereas decisions of higher courts tend to be binary as in *reject* or *admit*, sometimes with minor corrections to the original decision. This is the general approach in the literature as well. Thus the problem is formulated as a binary classification problem between “reject” or “admit”. Details of how we categorize case texts and the pre-processing thereof are described in detail later in the paper. We utilize various classification methods including prominent ones in the literature: Decision Trees (DTs), Random Forests (RFs), Support Vector Machines (SVMs), Gated Recurrent Units (GRUs), Long Short-Term Memory networks (LSTMs), bidirectional LSTMs (BiLSTMs) and variants of these deep learning models with attention mechanism, and compare their performances. Supervised methods on sentence-level annotated data are not used in this work, as the goal was to develop an end-to-end system for prediction. The results that we have obtained are varying over different courts. However, accuracy values reaching 93%, and more importantly, F1 scores reaching 0.87 are obtained, which are on par with the best results in the literature.

1.2. Organization of the paper

The rest of this paper is organized as follows. A comprehensive overview of the related literature is given in Section 2 and the corpus that was used and the processing thereof is explained in Section 3. The methods that we used and their implementation details are given in Section 4, together with details on data preparation and the metrics that were used to evaluate the performance. The results of our experiments are presented in Section 5. Discussions of results and implications of our research are given in Section 6. Finally, we conclude in Section 7.

2. Related work

2.1. Origins

The use of artificial intelligence (AI) in law has a long history. The conception of the idea that these two fields could be brought together goes back to the 1970s. Buchanan and Headrick (1970) speculated such a relationship, offering a multitude of areas where AI might contribute to law. According to Bench-Capon et al. (2012), however, the birth of an active community of AI and law research is marked by the year 1987 when the first International Conference on AI and Law (ICAIL) was held. Early research focuses on exploring and utilizing logical structures in legal argumentation, also making use of a knowledge base consisting of legal cases. Finding and distinguishing precedents in legal discourse (Ashley, 1989; Sartor, 1993) or factors that, for example, might favor a side (Ashley & Rissland, 1988) attracted research, motivated by *case based reasoning* (CBR) systems (Ashley, 1988). These systems are designed to work based on a knowledge of previous cases, with a variety of possible applications. Extensive research has been done involving the design, improvement and utilization of CBR systems (Ashley, 1988, 1991; Bench-Capon et al., 2012; Hafner & Berman, 2002; Wyner, 2008). By being able to model legal arguments and making it possible to do better indexing and retrieval depending on the structures of cases, these models have many applications (Ashley, 1992). Similar rule-based approaches are used for evaluating cases and predicting court rulings (Aleven, 2003; Ashley & Brüninghaus, 2009). One special advantage is the interpretability of the results they provide. For further details on the topic, one may refer to the overview written by Bench-Capon et al. (2012). While similar research remains active, recent developments in NLP and deep learning have also found their way into the field of law, increasing the number of applications. The analysis of legal documents with NLP involves carrying common NLP methods into the legal domain and possibly further customizing and specializing them to better fit different tasks at hand.

2.2. Overview of NLP in law

A primary problem is to extract features from a legal text. One such task is Named Entity Recognition (NER). NER systems specific to the legal domain have been studied in the literature (de Araujo et al., 2018; Cardellino et al., 2017; Dozier et al., 2010; Leitner et al., 2019). Another study which can set an example of utilizing previous knowledge and improving it on legal texts is the work of Elnaggar et al. (2018). The authors show that a transfer learning approach which incorporates training a Named Entity Linking system first on non-legal data, and then further training it on legal data, improves performance compared to solely training it on legal data. Although there are studies also on Turkish NER (Akkaya & Can, 2020; Güneş & Tantuğ, 2018; Güngör et al., 2019; Tür et al., 2003), there exists no work specific to the legal domain.

Another important task in extracting features from legal texts is detecting word or sentence level law-specific features such as facts, obligations, prohibitions and principles (Ashley & Brüninghaus, 2009; Chalkidis et al., 2018; O'Neill et al., 2017; Shulayeva et al., 2017; Sleimi et al., 2018). For the annotation of legal features, supervised learning is commonly used, where every sentence in a text is manually annotated as belonging to one or more of the given classes (Ashley & Brüninghaus, 2009; Shulayeva et al., 2017; Sleimi et al., 2018). In the work of Ashley and Brüninghaus (2009), case sentences are automatically represented by features called *factors* (parts of text that matter for the result) using a nearest-neighbor algorithm. A hand-coded algorithm decides on the labels of new cases, by comparing them to existing cases, with 92% accuracy in predicting the outcome. Shulayeva et al. (2017) have used a multinomial Bayesian classifier to decide whether a given sentence contains a legal fact/principle or not, allowing the detection of facts and principles with 85% accuracy. Similarly, the work of Sleimi et al. (2018) does not utilize any learning method, but achieves a 0.86 F1 score at classifying texts into one of seventeen classes (such as action, agent, condition, constraint etc.) using hard-coded decision rules that search for certain sequences of part-of-speech (POS) tags and specific words. Extracted features like those listed above can then be used for retrieval or reasoning systems (Aleven, 2003; Sangeetha et al., 2017).

A different problem addressed in the literature is detecting the logical relations between texts, such as deciding whether a given plea is made based on the law. For instance, Nguyen et al. (2018) have devised a deep learning model consisting of cascaded neural structures to break a sentence into its requisite and effectuation parts, and have achieved 0.78 F1 score on annotated test data. A second example of relations between texts that is considered in the literature is logical entailment. Finding such relations is addressed via modeling the task as a classification problem where a classifier decides, for a given pair of sentences, whether one entails the other. In the legal domain, models have been developed to decide whether there is an entailment between a given query and a law article. A direct application of this would be to automatically find and cite supporting law articles (Chalkidis & Kampas, 2018; Do et al., 2017; Kim et al., 2017; Morimoto et al., 2017; Nanda et al., 2017).

2.3. Case outcome prediction with machine learning

We now look at studies that are relevant to the main aim of this paper, namely, predicting legal case outcomes. Our prediction methods, which we comparatively evaluate as a secondary aim of this paper, are adopted from these similar studies. Aleven (2003) and Ashley and Brüninghaus (2009) have developed systems to predict the outcomes of cases by using rule-based algorithms that make their decisions based on the results of similar cases, and compared these algorithms to simple machine learning methods. Similar cases are retrieved by finding the nearest neighbors of a given case in terms of previously extracted legal features such as aforementioned *factors*. They both achieved accuracy scores reaching around 92%. The fact that the decision of a court can be predicted by using only a few parameters was demonstrated in *The Supreme Court Forecasting Project* (Martin et al., 2004; Ruger et al., 2004). In this project, Decision Trees were used whose decision criteria were manually extracted from case descriptions. These trees were trained on a collection consisting of more than 600 cases. The simple and statistically determined decision hierarchy of these trees was able to predict case outcomes with 75% accuracy, and was sometimes able to surpass experts at predicting justice votes. More recently, a number of different legal systems and languages have been studied with the purpose of trying to predict court decisions using machine learning techniques (Aletras et al., 2016; Katz et al., 2017; Kowsrihawat et al., 2018; Long et al., 2019; Şulea et al., 2017b; Virtucio et al., 2018).

Traditional machine learning techniques using language features, usual word and n-gram frequencies, have proven quite useful for the case outcome prediction task. Katz et al. (2017) consider more than 28,000 cases of the US Supreme Court, spanning a period of nearly two centuries, and aims to predict the outcomes both at case level and justice vote level. In their work, they train online-growing Random Forests on categorical variables that were partly obtained from a database and partly engineered. They achieve 70% accuracy at the binary classification task of predicting the outcomes of cases. Support vector machines (SVMs) have also been shown to be successful in a similar task. Aletras et al. (2016) have compiled a corpus consisting of cases of the European Court of Human Rights, using more than 500 cases related to articles 3, 6 and 8 of the European Convention on Human Rights. They attempted to classify cases as 'violation' or 'no violation' using solely textual features. Using most frequent n-grams of up to order four as features, they trained SVMs and achieved 79% average accuracy in this binary classification. Another example is the work of Şulea et al. (2017b) that used SVMs to predict outcomes of French Supreme Court cases. On a collection consisting of over 130,000 documents, they extracted unigrams and bigrams from each case, and assigned a corresponding label in terms of the result. They trained SVMs to predict the outcomes. In addition to their temporal analyses of the results, they report an overall accuracy reaching 97% and an F1 score reaching 0.97 on predicting case outcomes. However, Şulea et al. (2017b) study the problem without the usual binary classification (accept/reject) model but use multi-label classifications with six or eight labels. Experiments have also been performed by Virtucio et al. (2018) to test the performance of Random Forests and SVMs on predicting outcomes of Philippine Supreme Court cases. They were able to reach 59% accuracy with a Random Forest classifier.

2.4. Deep learning in law

A very powerful machine learning framework applied in NLP is deep learning, specifically Gated Recurrent Units (GRUs), Long Short-Term Memory networks (LSTMs) (Hochreiter & Schmidhuber, 1997) and their variants combined with vector representations of words known as word embeddings (Mikolov et al., 2013a; Pennington et al., 2014; Turney & Pantel, 2010). Training word embeddings that are suitable to legal applications is a subject that needs to be addressed on its own. Chalkidis and Kampas (2018) provide the first publicly available legal word embeddings called *law2vec*, which are trained on a large legislation corpus in English consisting of 492M tokens. Chalkidis and Kampas (2018) also provide an overview of the deep learning techniques used in the legal domain, focusing on three issues: text classification, information extraction and information retrieval. As for classification, based on the work of O'Neill et al. (2017) that aims to classify sentence modality, Chalkidis et al. (2018) developed a state-of-the-art modality classifier using several LSTM based methods operating on law-specific word embeddings provided by Chalkidis and Androutsopoulos (2017). In another instance of text classification, Branting et al. (2018) trained a neural model to predict administrative adjudications. Deep learning has also been utilized for information extraction in legal texts. Examples include recognizing parts of sentences that are labeled as requisite and effectuation (Nguyen et al., 2018) and the work of Chalkidis and Androutsopoulos (2017) which focuses on extracting contract elements based on the dataset provided by Chalkidis et al. (2017). Yet another area that deep learning has proven successful is information retrieval in the legal domain. Such applications include finding law articles related to a query (Do et al., 2017; Kim et al., 2017; Morimoto et al., 2017; Nanda et al., 2017), matching cases with law provision (Tang et al., 2016) and finding fact assertions in cases related to a given query (Nejadgholi et al., 2017).

Deployment of deep learning for case outcome prediction has been introduced by Long et al. (2019). For this purpose, they have developed a prediction system called *AutoJudge*. This model consists of three bidirectional Gated Recurrent Units (BiGRUs) that each encode pleas, fact descriptions and relevant laws separately, a pairwise attention mechanism that they themselves have designed to turn these encodings into more meaningful representations, and finally a Convolutional Neural Network (CNN) layer that turns these representations into the final vectors used for classification. They trained this model and other simpler deep learning models on 100,000 divorce proceedings of the Supreme People's Court of the People's Republic of China for predicting whether they are granted divorce or not. They achieved the best results of 82% accuracy and 0.83 F1 score with *AutoJudge*. Kowsrihawat et al. (2018) have utilized a similar system consisting of BiGRU encoders for encoding facts and laws, followed by further attention and hidden layers for case outcome prediction. They evaluated this model on a collection consisting of more than 1000 cases of the Thai Supreme Court for binary classification. The best Macro-F1 score of 0.63 was obtained with this model.

2.5. Relevance to our work

Although sharing a common ground, the aforementioned related works have their differences compared to each other in terms of the data and methods used. Katz et al. (2017) and Ruger et al. (2004), for instance, utilize categorical features and their derivatives, unlike others that use features like word or n-gram counts. Şulea et al. (2017b) work with six or eight class labels whereas all others formulate the task as binary classification. Long et al. (2019) and Kowsrihawat et al. (2018) take texts of relevant laws into account in addition to case descriptions provided by the court. Our work mostly shares the common ground of these works. We aim to predict outcomes of cases by looking solely at case descriptions (and at relevant laws where available). We use textual features and classify cases into one of two possible results, violation or no violation. Although the works mentioned above may have some

Table 1
Summary of previous work (accuracy scores are in percentages). Scores in each work are obtained using a different collection of legal data from mentioned courts.

Authors	Court	Machine learning methods							
		DT		RF		SVM		DL	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Aletras et al. (2016)	European Court of Human Rights					79.0			
Katz et al. (2017)	US Supreme Court			70.2	0.69				
Kowsrihawat et al. (2018)	Thai Supreme Court						0.61		0.63
Long et al. (2019)	Supreme People's Court of People's Republic of China					55.5	0.56	82.2	0.83
Ruger et al. (2004)	US Supreme Court	75.0							
Şulea et al. (2017b)	French Supreme Court					96.9	0.97		
Virtucio et al. (2018)	Philippine Supreme Court			59.0		55.0			

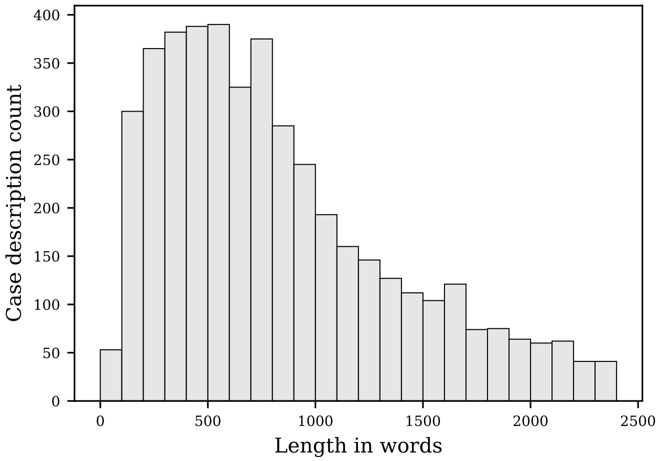


Fig. 1. Histogram demonstrating lengths (in words) of the case descriptions for Constitutional Court cases.

different aspects, together they roughly constitute a criterion of success that we can compare our work to and draw conclusions from.

While these pioneering works provide proof of concepts of the viability of machine-learning-based prediction, most of them are limited in scope, as they usually deal with only one type of court or only one algorithm and thus provide a fragmented picture of what can be achieved. This can be better appreciated by examining Table 1 that summarizes previous results, which we observe to be quite sparse. In the present work, we consider a variety of courts and a variety of algorithms and compare these with earlier studies. By providing a comprehensive study for the legal system of Turkey, we not only provide a baseline for this particular case, but also a framework for the study of other legal systems.

3. Corpus creation

In this section we describe the Turkish legal corpus compiled for this study. We analyze the decisions of the Constitutional Court of the Republic of Turkey (*Anayasa Mahkemesi*), Civil Court of Appeal (*Bölge Adliye Mahkemesi Hukuk Daireleri*), Criminal Court of Appeal (*Bölge Adliye Mahkemesi Ceza Daireleri*), Administrative Court of Appeal (*Bölge İdare Mahkemesi İdare Daireleri*) and Court of Appeal on Taxation (*Bölge İdare Mahkemesi Vergi Daireleri*). The Courts of Appeal that are mentioned consist of many regional subdivisions. The corpus we have constructed spans all the local subdivisions, and is not restricted to one region or court.

The texts we have compiled are obtained from official data made available online. The structures of the court cases differ for each court. The details will be given below along with how they have been preprocessed and labeled.

3.1. The constitutional court of the Republic of Turkey

The major duty of the Constitutional Court is a posteriori constitutional review for newly made legislation. However, since 2010, individual applications regarding human rights violations have also been allowed. This has provided abundant court case data for our work. Furthermore, for all individual applications, in addition to the text of the court case, a convenient *case overview table* including information on the topic, the verdict and relevant law are also provided.

In our comprehensive study, all 6,485 court cases for individual applications to the Constitutional Court available at the time of our study were used. On the Constitutional Court website, these court cases are provided with case overview tables. These tables

Table 2
Number of "violation"/"no violation" Constitutional Court cases by constitutional rights discussed in the cases.

Constitutional right	Violation	No violation
Right to Equitable Trial	138	27
Freedom of Expression	155	32
Right to Trial within a Reasonable Time	987	49
Property Right	371	102
Right to Respect for Private and Family Life	192	70
Right to Access to Courts	203	43
Right to Personal Freedom and Security	196	108

include the constitutional rights at stake, verdicts on whether there is a violation or not for each right, and the law that is relevant to the case, together with a full textual explanation. Relevant law articles are later used as features for classification (see Section 4.2). The texts of the cases are processed as described in Section 4.2, and are also utilized for classification. These texts consist of five parts: a brief overview of the topic, details of the application, a description of facts, the review and the result. The first, second and third parts that constitute the case description are extracted from each text, and the rest (the decisions) are discarded not to reveal the result to our learning algorithms. Documents that could not be divided as such are omitted. Fig. 1 shows the distribution of extracted case description texts according to their lengths in word counts. These case descriptions are used in training. The verdicts are also provided in the case overview tables, removing any need to extract them from the text. This information is then used to label each court case as "violation" or "no violation". Cases that resulted in the irrelevancy or ill-foundedness of the application or were filed after the statute of limitations deadline and therefore excluded from further consideration by the court are not used in our study.

Cases of the Constitutional Court can be subdivided into categories according to the constitutional right whose violation is being brought into question (it should be noted that a single case might appear in more than one of these categories if multiple rights are in question). When this categorization is done, a total of 41 categories emerged, 21 of which contain less than a hundred cases, and 7 which contain less than ten. With the further removal of cases that did not result in either "violation" or "no violation" from these categories, only 7 categories remained with a sufficient number of cases so that even when further partitions that are later described in this section are carried out, there exist samples from both "violation" and "non-violation" cases in each set. The number of cases in these seven categories adds up to a total of 2,673 (counting duplicates if a case appears in more than one category). The numbers are low for some categories, but still enough to apply simple classification methods. Table 2 shows the number of court cases in each category.

A unified corpus was also created from the Constitutional Court cases to allow for further experiments by providing a larger collection. In this unified corpus, all cases that resolve in "violation" or "no violation" were brought together regardless of the constitutional right in question. Cases that brought multiple rights into question where each resulted in a different verdict were discarded (if a case in Table 2 appears in more than one category with different results, it is considered *mixed*). These procedures led to a lower number of cases than the total number in Table 2. A collection consisting of 1,290 cases is obtained as a result of this compilation, 149 with no violation and 1,141 with violation. Table 3 gives a numerical overview of this collection.

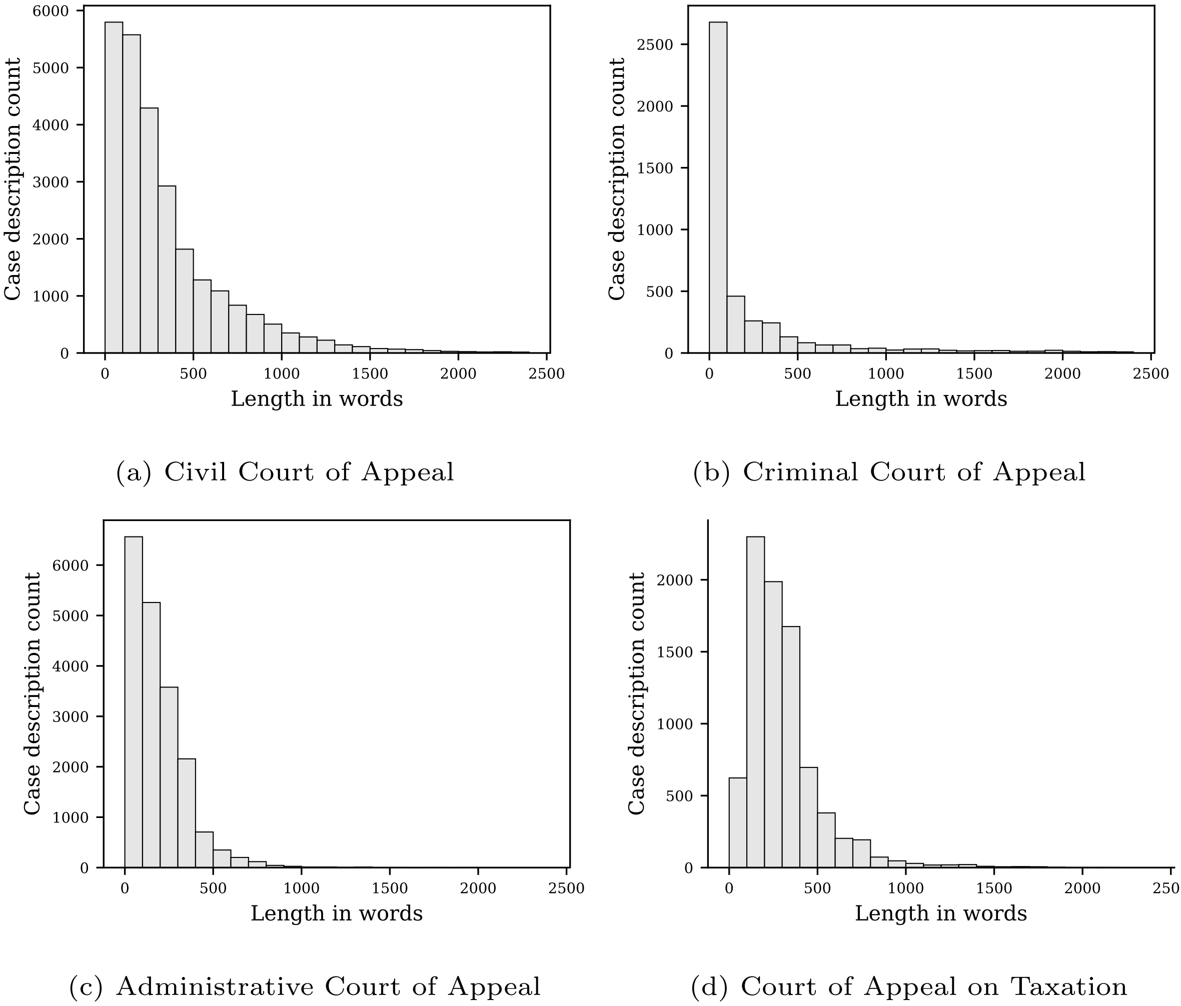
3.2. Civil court of appeal

The cases of the Courts of Appeal do not contain readily extracted features such as an overview table or a list of relevant laws, which was the case with the Constitutional Court. Their documents also do not follow a strict pattern as there are many regional divisions. Thus, working on these cases is more complicated and harder than working with those of the Constitutional Court. However, there are still certain common keywords (or keyphrases), usually written in all capitals or as separate titles, that mark where the description of the facts ends and the justification for the decision begins. We search for these keywords and divide the document into two from the first line where one of these keywords occur. The part before the keyword is used as the case description (used for training), and the part after the keyword is used as the case decision (used for label extraction). There is no exact, non-changing set of such keywords, and the keywords we use are of our choice. The set of used keywords might change with respect to legal corpus or time. If none of the keywords are found in a document, the document is deemed unsplittable and is discarded. From those that are successfully split, a label is extracted from the case decision part with another keyword search. If label extraction fails, those documents are discarded as well. We choose the splitting and labeling set of keywords from fundamental expressions denoting an assessment or verdict. Our choice of keywords is such that most of our case documents can be split and labeled successfully (see Table 3).

For the Civil Court of Appeal, all 47,796 available case documents were used. The documents were split from the first occurrence of one of the following keywords: *'GEREKÇE'* (justification), *'KARAR'* (decision), *'DEĞERLENDİRİLMESİ / DEĞERLENDİRME'* (assessment), *'HÜKÜM'* (verdict), such that the rest will be kept hidden from the machine learning models. Fig. 2 presents the lengths of case description texts extracted from the Courts of Appeal cases. Then, if the remainder of the text (the decision) contains one of the following keywords, the case is labeled as such: *'REDDİ'* (rejection), *'KABULÜ / KALDIRILMASI'* (admission). Cases with mixed/partial decisions are discarded (in our circumstance those that include both of these keywords). If a document fails to meet these structural patterns (such as not containing the keywords) they are omitted as well. The resulting number of documents are listed in Table 3 with their corresponding labels.

Table 3
Number of successfully split and labeled documents and their labels for all courts. Only rejected or admitted cases are considered in our study. For the Constitutional Court, rulings of “violation“ are listed under admitted and of “no violation” are listed under rejected. (843 cases of the Constitutional Court that are not listed here have other results such as “Irrelevancy”. Some of the cases that are listed as mixed here may appear in constitutional rights-based categories).

Court	Input number of cases	Split & labeled	Rejected	Admitted	Mixed
Constitutional Court	6,485	4,973	149	1,141	2,840
Civil Court of Appeal	47,796	26,327	12,519	8,702	5,106
Criminal Court of Appeal	9,241	4,385	1,891	420	2,074
Administrative Court of Appeal	20,948	19,046	6,851	966	11,229
Court of Appeal on Taxation	8,870	8,302	3,276	559	4,467



(a) Civil Court of Appeal

(b) Criminal Court of Appeal

(c) Administrative Court of Appeal

(d) Court of Appeal on Taxation

Fig. 2. Histograms demonstrating lengths (in words) of case description texts for (a) Civil Court of Appeal, (b) Criminal Court of Appeal, (c) Administrative Court of Appeal and (d) Court of Appeal on Taxation.

3.3. Criminal court of appeal

All 9,241 case documents from the Criminal Court of Appeal were used. The same keywords were used for the split with the addition of *‘GEREĞİ DÜŞÜNÜLDÜ’* (a verdict has been reached). When labeling, documents containing *‘YER OLMADIĞINA’* (unjustifiable) were also considered rejected. Cases that could not be split or had partial decisions were omitted. The resulting number of documents can be seen in Table 3.

Table 4
Overview of created corpora. For the Constitutional Court, rulings of "violation" are listed under admitted and of "no violation" are listed under rejected.

Corpus	Training	Validation	Test	Rejected	Admitted	Total
Right to Equitable Trial	99	33	33	27	138	165
Right to Freedom of Expression	112	37	38	32	155	187
Right to Trial within a Reasonable Time	621	207	208	49	987	1,036
Property Right	283	94	96	102	371	473
Right to Respect for Private and Family Life	157	52	53	70	192	262
Right to Access to Courts	147	49	50	43	203	246
Right to Personal Freedom and Security	182	60	62	108	196	304
Constitutional Court (unified)	902	194	194	149	1,141	1,290
Civil Court of Appeal	14,854	3,183	3,184	12,519	8,702	21,221
Criminal Court of Appeal	1,617	347	347	1,891	420	2,311
Administrative Court of Appeal	5,471	1,173	1,173	6,851	966	7,817
Court of Appeal on Taxation	2,684	575	576	3,276	559	3,835

3.4. *Administrative court of appeal*

All 20,948 case documents from the Administrative Court of Appeal were processed. These documents follow a different structure. They begin with a description of facts, followed by the phrase *'TÜRK MİLLETİ ADINA'* (in the name of the Turkish People) and the decision and the reason thereof. These documents were split from this point, and the preceding parts (case description) were used for training. For the labels, the same keywords as above are sought in the remaining portion of the text. The documents that could not be split or labeled were omitted. The resulting number of case documents can be seen in Table 3.

3.5. *Court of appeal on taxation*

All 8,870 case documents from the Court of Appeal on Taxation were used. The splitting and labeling process is the same as the one for administrative courts. The resulting number of documents can be seen in Table 3.

After all the steps described above were carried out for each court, the resulting corpora were finally split into training, validation and test sets. We use 60% training, 20% validation, and 20% test data splits for the constitutional rights-based corpora (to avoid having too few examples of a class in the validation and test sets), and 70%, 15%, 15% splits for other corpora (unified Constitutional Court corpus and Courts of Appeal corpora). An overview of the statistics of the created corpora after all the splitting, labeling and partitioning can be seen in Table 4.

4. Methods

In this section, prediction methods that we have utilized, data preparation, and evaluation metrics will be described.

4.1. *Classification methods*

The most prominent methods in the literature for predicting case outcomes are Decision Trees, Random Forests and Support Vector Machines (SVMs) (Aletras et al., 2016; Katz et al., 2017; Şulea et al., 2017b; Tan et al., 2005).

Decision Trees can be used as a classification method, and can be grown from training data using appropriate methods. Decision Trees in such tasks are usually built top-down heuristically by splitting the data at decision nodes usually according to either an impurity criterion or information gain (Rokach & Maimon, 2005). Decision Trees make it possible to use the most important features of the data, which are ordered according to the increase in purity or information gain. Ruger et al. (2004) have shown that Decision Trees with even a few nodes can be effective at predicting decisions of the US Supreme Court. However, unlike our work where automatically extracted features are used (see Section 4.2), they worked on manually crafted features and achieved 75% accuracy. We have therefore incorporated Decision Tree as our first learning algorithm in our work. We tune the parameter of minimum allowed samples per leaf based on the validation set, effectively pruning the tree if necessary.

Random Forest is an ensemble learning technique where the compound of many decision trees vote on the result and the number of trees is an additional parameter (Breiman, 2001). It is more robust against noise compared to Decision Trees because of the random splits at each classifier. Random Forests are widespread in NLP (Follett et al., 2019; Haneczok & Piskorski, 2020; Kaufhold et al., 2020), and the reason thereof is that they, like Decision Trees, offer a way to classify texts considering only the most important features.

As mentioned earlier, Random Forests are used to predict the outcomes of the US Supreme Court in the work of Katz et al. (2017) and were proven to be useful with an accuracy of 70%. We therefore incorporate them as one of our methods. The hyperparameters that are specifically tuned in our work are the minimum number of samples at each leaf of a tree, and the number of trees.

SVM is a very successful classification method that is widely used for both general text classification and in the legal domain (Haneczok & Piskorski, 2020; Kumar et al., 2020). For multiclass classification, a separate SVM is trained for each class against the rest, and the one with the highest score is chosen as a label. The virtual high dimensional feature space is achieved by the choice of an appropriate kernel function (Cortes & Vapnik, 1995).

SVMs are also studied in the legal text processing domain in the works of Aletras et al. (2016) and Şulea et al. (2017b). The former achieved 79% accuracy on the European Court of Human Rights, and the latter achieved 97% accuracy on the French Supreme Court. In our work, the hyperparameters of SVM that are optimized are the regularization parameter and the choice of the kernel function.

Decision Trees, Random Forests and SVMs are trained on principle component projections of word count vectors (see Section 4.2 for all details). Hyperparameters are tuned by inspecting validation accuracy of the models. All of them were trained with balanced class weights to prevent models from biasing towards the majority class.

Deep Learning based methods are also widely used for text classification where there are sufficient data (Elnagar et al., 2020; Goodfellow et al., 2016; Haneczok & Piskorski, 2020). The usual approach is to replace words with their word embeddings, and then use a deep neural network, usually recurrent models, to process this information (Ji et al., 2020b; Long et al., 2019). Then a simple neural network classifies texts based on extracted features. Long Short-Term Memory Networks (LSTMs) (Hochreiter & Schmidhuber, 1997) or Gated Recurrent Units (GRUs) (Cho et al., 2014), which are simpler alternatives to LSTMs, and their bidirectional variants are the most well-established approaches to the processing of text. These gated models, especially LSTMs, are well suited for the processing of long texts due to their internal memory mechanisms that help keep important information throughout the process (Goodfellow et al., 2016). However, for particularly long texts such as the ones in the legal text corpora we collected in this work, a state-of-the-art improvement is the use of attention mechanisms (Bahdanau et al., 2015). With this approach, either words themselves or the internal states of the neural units are weighted with an attention score and summed to obtain a final representation of the text (the case document in our situation). Attention mechanisms, although increasing the parameter count, allow distinguishing important parts of text, as well as ensuring better gradient flow during training.

Long et al. (2019) have utilized deep neural networks with attention mechanism to predict case outcomes of the Supreme People's Court of People's Republic of China. They have achieved 82% accuracy and a 0.83 F1 score.

In our work, we apply GRUs, LSTMs and bidirectional LSTMs (BiLSTMs) as representative deep learning methods. Two alternatives for each method are considered, with and without the attention mechanism. Models are used to encode a given text into a feature vector. This encoding structure is then followed by a dense classification layer with softmax activation. The outputs of this layer correspond to probabilities of the two classes: "admit" and "reject". For models with attention mechanism, the attention output is used as the feature vector.

4.2. Data preparation

Traditional machine learning methods demand a feature vector extracted from each document. For this purpose, a vector representing word (unigram) frequencies was used. To create these vectors, first, each case text was tokenized and stemmed using the Turkish NLP tool *Zemberek* (Akın & Akın, 2013). This stemming method uses a hand-crafted algorithm to stem the words as opposed to methods such as that in the work of Tursun et al. (2016) which are also suitable for agglutinative languages such as Turkish. Numbers, dates etc. are dropped so that only words remain. A vocabulary was created from these words, and very rare words whose numbers of occurrence throughout the corpus lie below a threshold of 50 are removed. This threshold is chosen by inspection so that the words that lie below this threshold are mostly proper nouns. For each case document, a vector of word frequencies was created whose size corresponds to the vocabulary, and entries to the number of occurrences of a word. The vocabulary size for each corpus can be seen in Table 5. On the Constitutional Court website, cases are provided together with a list of relevant laws in *case overview tables*. Relevant articles of law are extracted from these tables, and a one-hot encoded vector of binary features is created for every case, with one component for each law article. The texts of the law articles were not used. A total of 3,528 law articles are mentioned in all cases of the Constitutional Court. The number of relevant law articles for each constitutional right can be seen in Table 5, resulting in vectors of that size. These binary vectors were appended at the end of the vocabulary vectors when training on the constitutional rights-based sets. This procedure is not followed in the unified Constitutional Court corpus to leave it as a pure NLP task and allow comparison with Courts of Appeal. When creating the feature vectors, since the vocabulary size is large, the procedure resulted in a very high dimensional feature space. However, most of the variance in the data is correlated or uniform across cases. Therefore, this information can be represented using a reduced number of dimensions. To reduce the dimensions of the feature space, Principle Component Analysis (PCA) was performed. The PCA dimensions are chosen for each court such that 95% of the variance in the data is to be preserved. This indeed yielded significant dimensionality reduction. Table 5 shows original vocabulary and law vector sizes, and the resulting dimensions after PCA is applied.

For deep learning, most modern NLP applications take advantage of distributional representations of words, also known as *word embeddings* (Mikolov et al., 2013a; Pennington et al., 2014; Turney & Pantel, 2010). Word embeddings are trained on large corpora to place word vectors in a semantic space. The underlying assumption (named the distributional hypothesis of linguistics) is that the meaning of a word, to some degree, can be inferred from its statistical co-occurrence with other words. They have performed well on simple low-level evaluation tasks such as semantic similarity tasks (Mikolov et al., 2013b; Turney & Pantel, 2010).

To make use of word embeddings, we performed tokenization and removal of non-words. Stemming, however, was not performed although it can decrease the size of the vocabulary. Unlike our previous methods, the methods that use word embeddings do not suffer from increased dimensionality of the model as the vocabulary size increases. More importantly, in agglutinative languages like Turkish, stemming a word may lead to significant loss of meaning. Therefore, not stemming words ensures better quality of word representations without increasing complexity.

After tokenization, each word is replaced by 400-dimensional word2vec vectors (Mikolov et al., 2013a) which were pre-trained on Wikipedia articles in Turkish (Köksal, 2018).

Table 5
Dimensions before and after PCA is applied.

Corpus	Vocabulary size	Law vector size	Dim. after PCA
Right to Equitable Trial	4,252	245	38
Right to Freedom of Expression	7,159	143	24
Right to Trial within a Reasonable Time	7,300	1,051	120
Property Right	5,407	685	61
Right to Respect for Private and Family Life	5,012	318	29
Right to Access to Courts	4,295	384	73
Right to Personal Freedom and Security	6,566	198	18
Constitutional Court (unified)	10,606	n/a	334
Civil Court of Appeal	12,946	n/a	587
Criminal Court of Appeal	6,041	n/a	119
Administrative Court of Appeal	8,123	n/a	607
Court of Appeal on Taxation	4,795	n/a	250

For most of the court case corpora we created, the number of texts are relatively small when considering the large number of parameters that deep learning models bring about. To address this issue, we enhanced the training sets by breaking each text into 100-word chunks, and including these new chunks of text as training samples for our models, in addition to samples from the original training splits mentioned in Section 3. This data augmentation procedure increased the number of samples massively and slowed the training down. By simulating a larger collection of texts, although slowing down the training, it is aimed that models would be forced to predict outcomes by looking at different parts of texts, instead of possibly inclining toward one irrelevant word or phrase in each text, which is likely given that our corpora are not very large. It is thus intended that in a way, this would prevent the models from *overfitting* the data. A more detailed discussion of the effects of this operation is given in Section 5.

4.3. Evaluation metrics

We use the most common evaluation metrics in the literature, the first one being accuracy (Aletras et al., 2016; Katz et al., 2017; Long et al., 2019; Şulea et al., 2017b), defined as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP and FN denote the numbers of true positives, true negatives, false positives and false negatives, respectively. However, most of our data is heavily imbalanced as can be seen in Table 4, and in some cases, achieving 80% accuracy is trivial. We therefore used two more metrics that take imbalance into account. The first one is balanced accuracy (BACC), which is the average of per-class accuracies defined as

$$BACC = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$$

where P and N denote the total numbers of positives and negatives. The other is the F1 score, which is commonly used in the literature (Chalkidis et al., 2018; Elnaggar et al., 2018; Long et al., 2019; Nanda et al., 2017; Sleimi et al., 2018; Şulea et al., 2017b), defined as

$$F1 = 2\,\frac{Precision \cdot Recall}{Precision + Recall}.$$

However, F1 score assumes a positive and a negative class, and it does not make sense in our case to arbitrarily call a class positive, as it would depend on from which one of the parties' perspective one looks. We therefore used macro-averaged F1 score (MA-F1) as in the literature (Kowsrihawat et al., 2018), which is the mean of per-class F1 scores that are calculated by using both of the alternatives.

5. Results of experiments[1]

In this section, the results of our experiments are presented. Reported scores were obtained from test runs over unseen data. Scores were compared to a baseline that only makes random guesses by using class label weights based on their class frequency in the training set.

The first experiments for deep learning were done using data where the augmentation procedure described in Section 4.2 is not carried out. In that case, validation scores were at the vicinity of other methods, and the training that took orders of magnitude longer did not show any significant improvement in return. Especially for models without attention, gradients vanishing after a few hundreds of words would prevent getting the best out of training. Therefore, data augmentation was done by breaking each case text into 100-word chunks as described in Section 4.2. All of the reported results are for experiments on augmented data.

[1] Data and codes are available at: https://github.com/koc-lab/law-turk.

Table 6
Constitutional rights-based classification results of Constitutional Court cases.

Constitutional right	Method	ACC(%)	BACC(%)	MA-F1
Right to Equitable Trial	Baseline	72.0	72.0	0.42
	DT	78.9	72.9	0.73
	RF	78.9	60.0	0.60
	SVM	89.5	80.0	0.84
Freedom of Expression	Baseline	62.0	37.5	0.38
	DT	63.0	50.0	0.46
	RF	81.5	60.4	0.59
	SVM	81.5	60.4	0.59
Right to Trial within a Reasonable Time	Baseline	87.8	47.2	0.47
	DT	84.0	62.0	0.55
	RF	90.4	47.6	0.47
	SVM	92.9	49.0	0.48
Property Right	Baseline	68.1	47.6	0.48
	DT	50.7	55.0	0.40
	RF	50.7	36.6	0.36
	SVM	89.6	57.6	0.58
Right to Respect for Private and Family Life	Baseline	55.0	49.0	0.46
	DT	42.4	36.5	0.37
	RF	69.7	67.3	0.64
	SVM	63.6	46.3	0.46
Right to Access to Courts	Baseline	65.8	35.7	0.40
	DT	75.0	60.3	0.58
	RF	72.2	50.3	0.50
	SVM	88.9	60.0	0.64
Right to Personal Freedom and Security	Baseline	52.2	48.7	0.49
	DT	38.5	40.6	0.38
	RF	38.5	44.2	0.32
	SVM	53.8	56.4	0.54

In the first experiments using the Constitutional Court corpus, cases for each constitutional right were considered separately. Since the number of cases in each category is not enough to train neural networks, only Decision Trees, Random Forests and SVMs were used for prediction and thus a comparison to deep learning models is not available. For these models, validation macro-F1 score was monitored as a success criterion. Parameters of models were tuned according to macro-F1 score to ensure a fair comparison. The test results are shown in Table 6. Since the number of instances in each category is very low, the results are not balanced and no one classification algorithm stands out clearly. Also, the prevalence of balanced accuracy scores close to 50% despite high accuracy scores may indicate that the classifiers are biased towards the majority class, and cannot detect instances of the smaller classes, which is “no violation”. The high accuracy scores, therefore, do not say much about the selectivity of the classifier. This can be seen more clearly if one compares these to the numbers in Table 2 where the accuracies are almost consistent with the majority class ratio. The balanced accuracy and F1 scores, however, indicate at least some success. SVM has proven, in general, to be more useful than Decision Tree and Random Forest, since it can handle small data better. The inadequacy in scores is an indication of the insufficiency of the size of each dataset, where a prediction is being made with very few samples that have high dimensional features even after dimensionality reduction.

Observing the inadequacy of using a very small corpus, as a next step, we conduct experiments on the unified Constitutional Court corpus. All classifiers are trained on this unified corpus. Furthermore, relevant law articles were not used as features in this experiment to allow for comparison with the Courts of Appeal and to allow training of the deep learning based method. This makes the task purely text-based unlike in the previous case, where we used one-hot encoded feature vectors corresponding to law articles in addition to textual features. Then, the same data preparation procedure that was performed for the Courts of Appeal is applied. The results of prediction on the test set for all methods are shown in Table 7. The highest accuracy of 91.8% and F1 score of 0.67 is obtained with the LSTM model with attention. Models with attention have outperformed their counterparts without attention. This is expected because cases of the Constitutional Court are quite long.

Then, the same experiments were performed for all Courts of Appeal. The results are reported in Table 7. For the Civil Court of Appeal, 69% accuracy and a 0.68 F1 score are observed. In this largest corpus we have, the additional parameters that BiLSTMs bring seem to be useful since there are enough data to train on. On Criminal Court of Appeal cases, 85.6% accuracy and a 0.77 F1 score are obtained. These highest results are achieved with the help of deep learning. Random Forest has overall performed better than SVM for these courts. On Administrative Court of Appeal cases, 91.1% accuracy is obtained, together with an F1 score of 0.77. Again, deep learning models have mostly outperformed other methods in these scores. Finally, for cases of the Court of Appeal on Taxation, for which the highest scores are obtained, an accuracy of 93.2% and an F1 score of 0.87 are achieved at best.

Table 7
Classification results on unified Constitutional Court corpus and Court of Appeal corpora.

Court	Method	ACC(%)	BACC(%)	MA-F1
Constitutional Court (unified corpus)	Baseline	79.4	50.7	0.51
	DT	85.1	60.7	0.62
	RF	87.6	56.3	0.57
	SVM	83.5	59.9	0.61
	GRU	87.6	56.3	0.57
	GRU + attention	89.2	60.0	0.61
	LSTM	83.0	61.3	0.62
	LSTM + attention	91.8	64.2	0.67
	BiLSTM	89.7	54.6	0.56
	BiLSTM + attention	90.2	57.7	0.59
Civil Court of Appeal	Baseline	50.8	49.2	0.49
	DT	61.5	60.3	0.60
	RF	68.7	67.3	0.67
	SVM	64.7	64.1	0.64
	GRU	66.8	63.5	0.64
	GRU + attention	66.7	66.1	0.66
	LSTM	66.7	65.7	0.66
	LSTM + attention	65.9	64.5	0.65
	BiLSTM	69.0	67.7	0.68
	BiLSTM + attention	67.3	65.0	0.65
Criminal Court of Appeal	Baseline	69.5	48.6	0.49
	DT	82.4	75.0	0.73
	RF	81.8	74.6	0.73
	SVM	80.1	71.2	0.70
	GRU	82.4	75.6	0.74
	GRU + attention	82.7	75.8	0.75
	LSTM	85.0	77.8	0.77
	LSTM + attention	85.6	76.6	0.77
	BiLSTM	82.4	71.0	0.72
	BiLSTM + attention	82.4	74.0	0.74
Administrative Court of Appeal	Baseline	78.9	51.0	0.51
	DT	86.3	73.0	0.72
	RF	86.7	74.3	0.73
	SVM	83.2	78.7	0.72
	GRU	90.0	71.3	0.74
	GRU + attention	89.6	69.2	0.72
	LSTM	90.1	75.2	0.77
	LSTM + attention	90.8	70.9	0.75
	BiLSTM	89.3	69.1	0.72
	BiLSTM + attention	91.1	72.8	0.76
Court of Appeal on Taxation	Baseline	76.7	53.8	0.54
	DT	89.9	83.6	0.81
	RF	89.4	83.3	0.80
	SVM	92.4	90.0	0.86
	GRU	91.8	81.1	0.84
	GRU + attention	92.0	79.8	0.83
	LSTM	92.9	89.3	0.87
	LSTM + attention	92.9	84.3	0.86
	BiLSTM	91.7	78.8	0.82
	BiLSTM + attention	93.2	80.6	0.85

6. Discussion of results and implications

In order to facilitate cross-comparison and provide an overall view of our results, the performance over all courts are tabulated in Table 8. On the deep learning column, we report the best out of all deep learning models. When an average is taken over the best results for each court, an average accuracy of 86.1%, average balanced accuracy of 75.7%, and average F1 score of 0.75 are obtained. These average results are on par with results obtained by other works in the literature reviewed in Section 2. The best scores we have obtained (accuracy 93.2%, balanced accuracy 90%, F1 score 0.87) surpass most of the scores reported in earlier works.

Among all the experiments, the ones that utilize deep learning methods have overall proven to be more useful. It seems that these models can perform at least as well as others because they can capture dependencies between words in addition to information embedded in each word. When we look at previous work, it is difficult to reach a definitive conclusion as to whether there is a similar pattern. Long et al. (2019) have obtained higher scores with their proposed deep learning model when compared to the other methods that they use. The difference is not as clear when compared to other works. Also, it can be seen that the results Kowsrihawat

Table 8
Overall results (accuracy scores are in percentages, F1 scores are macro averaged).

Court	Machine learning methods									
	Baseline		DT		RF		SVM		DL	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Constitutional Court	79.4	0.51	85.1	0.62	87.6	0.57	83.5	0.61	**91.8**	**0.67**
Civil Court of Appeal	50.8	0.49	61.5	0.60	68.7	0.67	64.7	0.64	**69.0**	**0.68**
Criminal Court of Appeal	68.5	0.49	82.4	0.73	81.8	0.73	80.1	0.70	**85.6**	**0.77**
Administrative Court of Appeal	78.9	0.51	86.3	0.72	86.7	0.73	83.2	0.72	**91.1**	**0.77**
Court of Appeal on Taxation	76.7	0.54	89.9	0.81	89.4	0.80	92.4	0.86	**93.2**	**0.87**

et al. (2018) have obtained with deep learning are lower compared to other works. It is likely that there are other factors at play here. Works using high-level features or features specifically designed to capture word relations might perform better even with simpler classification methods. We argue though, by performing many experiments on cases of different courts, and by comparing methods trained on simple word features, that deep learning is overall more reliable in the processing of legal text for case outcome prediction.

Comparing deep learning models, we find that GRUs, being the simplest ones, generally performed worse. It is known that LSTMs are better for processing longer texts (Goodfellow et al., 2016) and they are the ones that performed the best in our experiments. Although BiLSTMs, being more complicated, require the training of almost twice the parameters, they are the best performers in large corpora. For instance, in the Civil Court of Appeal, which is our largest corpus, the BiLSTM performed the best as there were lots of data to train on. In other courts, however, where the corpora are smaller, BiLSTMs seem to be excessively complex to do reliable training. The use of such complex models can only be justified with the availability of large collections of data. This is the case for the work of Long et al. (2019) where they use three bidirectional GRUs, but still obtain very good results, probably due to the size and quality of their data. Another observation that needs to be made from our results is that the attention mechanism almost always improves the performance. Although attention itself requires the training of more parameters too, it can be said confidently that such a mechanism is essential for processing very long texts such as ours, as they provide a way of retaining information throughout and allow smoother training with propagating gradients where even LSTMs are not enough.

Other important factors that affect the results other than the actual methods used are statistics such as the lengths of case texts or number of samples in the corpora. In the literature, works that utilize larger corpora have achieved better results. Examples are the works Long et al. (2019) and Şulea et al. (2017b), which do training on around 130,000 and 100,000 case documents, respectively, (see Table 1). One should be aware, however, that it may not be appropriate to make direct comparisons between works dealing with different countries and languages. In our work, interestingly enough, the scores for the courts with shorter case texts and with less numbers of cases are higher. Scores for Constitutional Court cases, which are the longest, and cases of the Civil Court of Appeal, which are the largest in number, are lower. Even though there is larger corpus to train on, these results might be an indication that to capture information in these longer texts, even more training data is required. It can be seen clearly that the performances of models suffer from working on very long texts, and while improvements such as the attention mechanism or bidirectional variants try to remedy this issue, an even larger corpus, if it was available, would perhaps be able to further elevate the performance of these models.

The numerical results that we have obtained have also implications beyond a simple comparison of algorithms. They tell us not only which algorithm performs better, but also which set of cases are easier or more difficult to predict. This, in turn, can give insight into the nature of cases in that set and the structure of the court they belong to. The performance for the Civil Court of Appeal is considerably lower compared to the other courts of appeal. This partly might have to do with the more balanced number of class labels (see Table 4), as the percent accuracy is naturally lower on a more balanced dataset, compared to a set where a single class dominates and high accuracy is trivial to achieve. However, this explanation does not apply to the other scores which are designed to be less affected by class imbalance. Despite a larger corpus being available and used for the Civil Court of Appeal, which, all else being equal should lead to better prediction performance, the opposite is observed. What could be the reason for this unexpected result? One factor that should not be overlooked is that the texts of the Civil Court of Appeal cases are longer compared to the cases of other courts of appeal (Fig. 2). However, there may also be reasons related to the nature of the law and its enforcement. By nature, the Civil Courts of Appeal deal with civil cases that are very general and indefinite in terms of their content whereas the Criminal and Administrative Courts of Appeals deal with more structured laws and cases. For example, the universal principle known as *the principle of no punishment without law* says that an act needs to be explicitly defined as a *crime* in the law for it to be punishable. Thus, the number of crimes are definite and their nature is well defined. On the other hand, an indefinite number of civil cases can be built on an unpredictable variety of interactions among people, institutions and entities. While the number of civil law articles and principles to be applied indirectly to them are finite, the potential situations to be judged are far from being predefined. Administrative law, while not being as structured as criminal law, nevertheless deals with a narrower scope of issues than civil law and thus may be argued to be more predictable. While these comments are preliminary and would require further work to substantiate, the fact that certain structural and content differences in different courts and the cases they deal with are reflected in prediction performance is meaningful. It highly suggests that the success rates of the algorithms are not merely about processing some texts regardless of content, but that the algorithms are actually doing something that is related to and that reflects the content of the texts, and perhaps more significantly the nature of the cases and the structure of the legal system.

Predicting the outcome of cases based on their descriptions can have a number of applications, some of which are bound to lead to ethical considerations and controversies. Attorneys may use this predictive tool to "weigh" the case in order to get a sense of how similar cases have been decided previously. Prosecutors may use it to judge whether a case is worth pursuing; thereby concentrating their efforts on more promising cases. The use of an automated prediction tool by judges would probably be the most controversial one. Should they use it as an aid in making decisions, or should they not even see the result of such predictions as it may bias them towards a particular decision? Should such tools be limited to obtaining aggregate statistics for evaluative purposes, rather than making decisions in individual cases? Apart from the application to the practice of law, these algorithms and results can shed light on the working of the legal system, the extent to which it is consistent, and to understand whether it is aligned with our sense of fairness and justice. These are open problems for future research.

7. Conclusion

Legal case outcome prediction is a machine learning and natural language processing application in law which has not received attention in the context of the legal system of Turkey. We have systematically studied the problem of predicting outcomes for court rulings in Turkey. Almost all possible court types have been studied in detail. Whereas almost all earlier work had used a single machine learning method, we have reported the results of several methods comparatively for several courts. Thanks to this breadth, we believe it will provide a reference point and baseline for further studies in this area. We further hope the scope and systematic nature of this study can set a framework that can be applied to the study of other legal systems, where (i) the legal system corpus is systematically separated to sub-corpora according to the different types and levels of courts within the hierarchy of the legal system at hand, (ii) a reproducible method of pre-processing data that makes it suitable for further higher-level processing is provided, (iii) experiments to characterize the performances of baseline, classical machine learning approaches like SVMs and random forests and several contemporary deep learning based methods with and without attention mechanisms are performed.

The results show that higher court rulings in Turkey can be predicted with good accuracy, as shown by considering several alternative measures. Direct comparison with previous work is not possible because earlier systems were developed for other languages and very different legal systems. Yet, this work should contribute to setting general baselines. Among the various methods considered, deep learning has overall yielded the highest prediction scores.

There are several technical issues that will constitute the content of future work, such as more detailed feature extraction and sentence-level supervision for systems that are not end-to-end. Further experiments and research on the retrieval of leading cases can also be addressed.

Our results have implications beyond comparing algorithms and demonstrating their predictive power. There is a variation in results obtained for different courts, which has interesting potential interpretations. More work is needed to uncover the meaning of this difference but we hypothesize that it is related to the different content of the cases and different structure of the different types of courts. One possibility is that certain courts have more predictable results because of the nature of the data or bookkeeping that is not substantially related to the content of the proceedings or the structure of the law. However, the results point to possible interpretations that predictability may be related to the actual content and structure. This in turn suggests that what the algorithms are doing goes beyond merely processing symbols or text, but is about the content of the case, and perhaps even about the structure of the legal system.

We also discussed some of the practical applications, and the legal and ethical implications of the use of such machine-based predictive systems.

CRediT authorship contribution statement

Emre Mumcuoğlu: Data curation, Software, Validation, Formal analysis, Investigation, Visualization, Resources, Writing - original draft, Writing - review & editing. **Ceyhun E. Öztürk:** Data curation, Software. **Haldun M. Ozaktas:** Writing - review & editing. **Aykut Koç:** Conceptualization, Methodology, Formal analysis, Supervision, Resources, Writing - original draft, Writing - review & editing, Funding acquisition.

Acknowledgments

This work was supported in part by TUBITAK 1001 grant (120E346) and in part by TUBITAK BIDEB 2210. H. M. Ozaktas acknowledges partial support of the Turkish Academy of Sciences.

References

Akın, A. A., & Akın, M. D. (2013). Zemberek, an open source NLP framework for Turkic Languages. https://github.com/ahmetaa/zemberek-nlp, GitHub Repository.

Akkaya, E. K., & Can, B. (2020). Transfer learning for Turkish named entity recognition on noisy text. *Natural Language Engineering*, *27*(1), 35–64. http://dx.doi.org/10.1017/S1351324919000627.

Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science*, *2*, Article e93. http://dx.doi.org/10.7717/peerj-cs.93.

Aleven, V. (2003). Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artificial Intelligence*, *150*, 183–237. http://dx.doi.org/10.1016/S0004-3702(03)00105-X.

Ansay, T., & Wallace, D. (2005). *Introduction to Turkish law*. Kluwer Law International, http://dx.doi.org/10.1093/iclqaj/37.1.241.

de Araujo, P. H. L., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., & Bermejo, P. (2018). LeNER-Br: A dataset for named entity recognition in Brazilian legal text. In *International conference on the computational processing of Portuguese* (pp. 313–323). Springer, http://dx.doi.org/10.1007/978-3-319-99722-3_32.

Ashley, K. D. (1988). *Modelling legal argument: Reasoning with cases and hypotheticals* (Ph.D. thesis), USA: University of Massachusetts, Order No: GAX88-13198.

Ashley, K. D. (1989). Toward a computational theory of arguing with precedents. In *Proceedings of the 2nd international conference on artificial intelligence and law* (pp. 93–102). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/74014.74028.

Ashley, K. D. (1991). Reasoning with cases and hypotheticals in HYPO. *International Journal of Man-Machine Studies*, *34*(6), 753–796.

Ashley, K. D. (1992). Case-based reasoning and its implications for legal expert systems. *Artificial Intelligence and Law*, *1*, 113–208. http://dx.doi.org/10.1007/BF00114920.

Ashley, K. D., & Brüninghaus, S. (2009). Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law*, *17*(2), 125–165. http://dx.doi.org/10.1007/s10506-009-9077-9.

Ashley, K. D., & Rissland, E. L. (1988). A case-based approach to modeling legal expertise. *IEEE Expert: Intelligent Systems and their Applications*, *3*(3), 70–77. http://dx.doi.org/10.1109/64.21892.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations*.

Bench-Capon, T., Araszkiewicz, A. M., Ashley, A. K., Atkinson, K., Bex, F., Borges, F., Bourcier, D., Bourgine, P., Conrad, J. G., Francesconi, E., Gordon, T. F., Governatori, G., Leidner, J. L., Lewis, D. D., Loui, R. P., McCarty, L. T., Prakken, H., Schilder, F., Schweighofer, E., ... Wyner, A. Z. (2012). A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law*, *20*, 215–319. http://dx.doi.org/10.1007/s10506-012-9131-x.

Branting, K. L., Yeh, A., Weiss, B., Merkhofer, E., & Brown, B. (2018). Inducing predictive models for decision support in administrative adjudication. In U. Pagallo, M. Palmirani, P. Casanovas, G. Sartor, & S. Villata (Eds.), *AI approaches to the complexity of legal systems* (pp. 465–477). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-00178-0_32.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Buchanan, B. G., & Headrick, T. E. (1970). Some speculation about artificial intelligence and legal reasoning. *Stanford Law Review*, *23*, 40–62. http://dx.doi.org/10.2307/1227753.

Cardellino, C., Teruel, M., Alemany, L. A., & Villata, S. (2017). A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th edition of the international conference on articial intelligence and law* (pp. 9–18). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3086512.3086514.

Chalkidis, I., & Androutsopoulos, I. (2017). A deep learning approach to contract element extraction. In A. Z. Wyner, & G. Casini (Eds.), *Frontiers in artificial intelligence and applications*: *vol. 302*, *Legal knowledge and information systems* (pp. 155–164). IOS Press, http://dx.doi.org/10.3233/978-1-61499-838-9-155.

Chalkidis, I., Androutsopoulos, I., & Michos, A. (2017). Extracting contract elements. In *Proceedings of the 16th edition of the international conference on articial intelligence and law* (pp. 19–28). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3086512.3086515.

Chalkidis, I., Androutsopoulos, I., & Michos, A. (2018). Obligation and prohibition extraction using hierarchical RNNs. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 254–259). Melbourne, Australia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-2041.

Chalkidis, I., & Kampas, D. (2018). Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, *27*, 1–28. http://dx.doi.org/10.1007/s10506-018-9238-9.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1179.

Cortes, C., & Vapnik, V. N. (1995). Support vector networks. *Machine Learning*, *20*, 273–297. http://dx.doi.org/10.1007/BF00994018.

Do, P., Nguyen, H., Tran, C., Nguyen, M., & Nguyen, M. (2017). Legal question answering using ranking SVM and deep convolutional neural network. http://dx.doi.org/10.13140/RG.2.2.21583.69282, arXiv preprint arXiv:1703.05320 [Cs.IT].

Dozier, C., Kondadadi, R., Light, M., Vachher, A., Veeramachaneni, S., & Wudali, R. (2010). Named entity recognition and resolution in legal text. In *Semantic processing of legal texts: Where the language of law meets the law of language* (pp. 27–43). Berlin, Heidelberg: Springer-Verlag, http://dx.doi.org/10.1007/978-3-642-12837-0_2.

Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic text classification using deep learning models. *Information Processing & Management*, *57*(1), Article 102121. http://dx.doi.org/10.1016/j.ipm.2019.102121.

Elnaggar, A., Otto, R., & Matthes, F. (2018). Deep learning for named-entity linking with transfer learning for legal documents. In *Proceedings of the 2018 artificial intelligence and cloud computing conference* (pp. 23–28). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3299819.3299846.

Follett, L., Geletta, S., & Laugerman, M. (2019). Quantifying risk associated with clinical trial termination: A text mining approach. *Information Processing & Management*, *56*(3), 516–525. http://dx.doi.org/10.1016/j.ipm.2018.11.009.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Güneş, A., & Tantuğ, A. C. (2018). Turkish named entity recognition with deep learning. In *2018 26th IEEE signal processing and communications applications conference* (pp. 1–4). http://dx.doi.org/10.1109/SIU.2018.8404500.

Güngör, O., Güngör, T., & Üsküdarlı, S. (2019). The effect of morphology in named entity recognition with sequence tagging. *Natural Language Engineering*, *25*, 147–169. http://dx.doi.org/10.1017/S1351324918000281.

Hafner, C. D., & Berman, D. H. (2002). The role of context in case-based legal reasoning: Teleological, temporal, and procedural. *Artificial Intelligence and Law*, *10*(1–3), 19–64. http://dx.doi.org/10.1023/A:1019516031847.

Haneczok, J., & Piskorski, J. (2020). Shallow and deep learning for event relatedness classification. *Information Processing & Management*, *57*(6), Article 102371. http://dx.doi.org/10.1016/j.ipm.2020.102371.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Ikram, A. Y., & Chakir, L. (2019). Arabic text classification in the legal domain. In *2019 3rd international conference on intelligent computing in data sciences* (pp. 1–6). http://dx.doi.org/10.1109/ICDS47004.2019.8942343.

Ji, D., Gao, J., Fei, H., Teng, C., & Ren, Y. (2020). A deep neural network model for speakers coreference resolution in legal texts. *Information Processing & Management*, *57*(6), Article 102365. http://dx.doi.org/10.1016/j.ipm.2020.102365.

Ji, D., Tao, P., Fei, H., & Ren, Y. (2020). An end-to-end joint model for evidence information extraction from court record document. *Information Processing & Management*, *57*(6), Article 102305. http://dx.doi.org/10.1016/j.ipm.2020.102305.

Junqué de Fortuny, E., De Smedt, T., Martens, D., & Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. *Information Processing & Management*, *50*(2), 426–441. http://dx.doi.org/10.1016/j.ipm.2013.12.002.

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), 1–18. http://dx.doi.org/10.1371/journal.pone.0174698.

Kaufhold, M.-A., Bayer, M., & Reuter, C. (2020). Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing & Management*, *57*(1), Article 102132. http://dx.doi.org/10.1016/j.ipm.2019.102132.

Kim, M.-Y., Xu, Y., & Goebel, R. (2017). Applying a convolutional neural network to legal question answering. In M. Otake, S. Kurahashi, Y. Ota, K. Satoh, & D. Bekki (Eds.), *New frontiers in artificial intelligence* (pp. 282–294). Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-50953-2_20.

Köksal, A. (2018). Turkish pre-trained Word2Vec model. https://github.com/akoksal/Turkish-Word2Vec, GitHub Repository.

Kowsrihawat, K., Vateekul, P., & Boonkwan, P. (2018). Predicting judicial decisions of criminal cases from Thai Supreme Court using bi-directional GRU with attention mechanism. In *2018 5th Asian conference on defense technology* (pp. 50–55). http://dx.doi.org/10.1109/ACDT.2018.8592948.

Kumar, A., Srinivasan, K., Cheng, W.-H., & Zomaya, A. Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management*, *57*(1), Article 102141. http://dx.doi.org/10.1016/j.ipm.2019.102141.

Leitner, E., Rehm, G., & Moreno-Schneider, J. (2019). Fine-grained named entity recognition in legal documents. In M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, & Y. Sure-Vetter (Eds.), *Semantic systems. The power of AI and knowledge graphs* (pp. 272–287). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-33220-4_20.

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management*, *57*(5), Article 102212. http://dx.doi.org/10.1016/j.ipm.2020.102212.

Long, S., Tu, C., Liu, Z., & Sun, M. (2019). Automatic judgment prediction via legal reading comprehension. In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), *Chinese computational linguistics* (pp. 558–572). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-32381-3_45.

Martin, A. D., Quinn, K. M., Ruger, T. W., & Kim, P. T. (2004). Competing approaches to predicting supreme court decision making. *Perspectives on Politics*, *2*(4), 761–767. http://dx.doi.org/10.1017/S1537592704040502.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th international conference on neural information processing systems - volume 2* (pp. 3111–3119). Red Hook, NY, USA: Curran Associates Inc.

Morimoto, A., Kubo, D., Sato, M., Shindo, H., & Matsumoto, Y. (2017). Legal question answering system using neural attention. In K. Satoh, M. Kim, Y. Kano, R. Goebel, & T. Oliveira (Eds.), *EPiC series in computing*: *vol. 47*, *4th competition on legal information extraction and entailment, held in conjunction with the 16th international conference on artificial intelligence and law in king's college London, UK* (pp. 79–89). EasyChair, http://dx.doi.org/10.29007/4l2q.

Nanda, R., John, A. K., Caro, L. D., Boella, G., & Robaldo, L. (2017). Legal information retrieval using topic clustering and neural networks. In K. Satoh, M.-Y. Kim, Y. Kano, R. Goebel, & T. Oliveira (Eds.), *EPiC series in computing*: *vol. 47*, *4th competition on legal information extraction and entailment* (pp. 68–78). EasyChair, http://dx.doi.org/10.29007/psgx.

Nejadgholi, I., Bougueng, R., & Witherspoon, S. (2017). A semi-supervised training method for semantic search of legal facts in Canadian immigration cases. In A. Z. Wyner, & G. Casini (Eds.), *Frontiers in artificial intelligence and applications*: *vol. 302*, *Legal knowledge and information systems* (pp. 125–134). IOS Press, http://dx.doi.org/10.3233/978-1-61499-838-9-125.

Nguyen, T.-S., Nguyen, L.-M., Tojo, S., Satoh, K., & Shimazu, A. (2018). Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. *Artificial Intelligence and Law*, *26*(2), 169–199. http://dx.doi.org/10.1007/s10506-018-9225-1.

O'Neill, J., Buitelaar, P., Robin, C., & O'Brien, L. (2017). Classifying sentential modality in legal language: A use case in financial regulations, acts and directives. In *Proceedings of the 16th edition of the international conference on artificial intelligence and law* (pp. 159–168). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3086512.3086528.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1162.

Qian, Y., Deng, X., Ye, Q., Ma, B., & Yuan, H. (2019). On detecting business event from the headlines and leads of massive online news articles. *Information Processing & Management*, *56*(6), Article 102086. http://dx.doi.org/10.1016/j.ipm.2019.102086.

Rokach, L., & Maimon, O. (2005). Decision trees. In O. Maimon, & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 165–192). Boston, MA: Springer US, http://dx.doi.org/10.1007/0-387-25465-X_9.

Ruger, T., Kim, P., Martin, A., & Quinn, K. (2004). The supreme court forecasting project: Legal and political science approaches to predicting supreme court decisionmaking. *Columbia Law Review*, *104*, 1150–1210. http://dx.doi.org/10.2307/4099370.

Sangeetha, D., Kavyashri, R., Swetha, S., & Vignesh, S. (2017). Information retrieval system for laws. In *2016 8th international conference on advanced computing* (pp. 212–217). IEEE, http://dx.doi.org/10.1109/ICoAC.2017.7951772.

Sartor, G. (1993). A simple computational model for nonmonotonic and adversarial legal reasoning. In *Proceedings of the 4th international conference on artificial intelligence and law* (pp. 192–201). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/158976.159001.

Schumaker, R. P., & Chen, H. (2009). A quantitative stock prediction system based on financial news. *Information Processing & Management*, *45*(5), 571–583. http://dx.doi.org/10.1016/j.ipm.2009.05.001.

Shulayeva, O., Siddharthan, A., & Wyner, A. (2017). Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law*, *25*(1), 107–126. http://dx.doi.org/10.1007/s10506-017-9197-6, Open access via Springer Compact Agreement.

Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., & Dann, J. (2018). Automated extraction of semantic legal metadata using natural language processing. In *2018 IEEE 26th international requirements engineering conference* (pp. 124–135). IEEE, http://dx.doi.org/10.1109/RE.2018.00022.

Şulea, O., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P., & van Genabith, J. (2017). Exploring the use of text classification in the legal domain. In K. D. Ashley, K. Atkinson, L. K. Branting, E. Francesconi, M. Grabmair, M. Lauritsen, V. R. Walker, & A. Z. Wyner (Eds.), *CEUR workshop proceedings*: *vol. 2143*, *Proceedings of the second workshop on automated semantic analysis of information in legal texts co-located with the 16th international conference on artificial intelligence and law*. CEUR-WS.org.

Şulea, O.-M., Zampieri, M., Vela, M., & van Genabith, J. (2017). Predicting the law area and decisions of french supreme court cases. In *Proceedings of the international conference recent advances in natural language processing* (pp. 716–722). Varna, Bulgaria: INCOMA Ltd., http://dx.doi.org/10.26615/978-954-452-049-6_092.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining* (1st ed.). USA: Addison-Wesley Longman Publishing Co., Inc.

Tang, G., Guo, H., Guo, Z., & Xu, S. (2016). *Matching law cases and reference law provision with a neural attention model*. Beijing: IBM China Research.

Tuke, J., Nguyen, A., Nasim, M., Mellor, D., Wickramasinghe, A., Bean, N., & Mitchell, L. (2020). Pachinko prediction: A Bayesian method for event prediction from social media data. *Information Processing & Management*, *57*(2), Article 102147. http://dx.doi.org/10.1016/j.ipm.2019.102147.

Tür, G., Hakkani-Tür, D., & Oflazer, K. (2003). A statistical information extraction system for Turkish. *Natural Language Engineering*, *9*(2), 181–210. http://dx.doi.org/10.1017/S135132490200284X.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141–188. http://dx.doi.org/10.1613/jair.2934.

Tursun, E., Ganguly, D., Osman, T., Yang, Y.-T., Abdukerim, G., Zhou, J.-L., & Liu, Q. (2016). A semisupervised tag-transition-based Markovian model for Uyghur morphology analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *16*(2), 8:1–23. http://dx.doi.org/10.1145/2968410.

Virtucio, M. B. L., Aborot, J. A., Abonita, J. K. C., Aviñante, R. S., Copino, R. J. B., Neverida, M. P., Osiana, V. O., Peramo, E. C., Syjuco, J. G., & Tan, G. B. A. (2018). Predicting decisions of the Philippine supreme court using natural language processing and machine learning. In *2018 IEEE 42nd annual computer software and applications conference* (pp. 130–135). IEEE, http://dx.doi.org/10.1109/COMPSAC.2018.10348.

Wyner, A. (2008). An ontology in OWL for legal case-based reasoning. *Artificial Intelligence and Law*, *16*(361), http://dx.doi.org/10.1007/s10506-008-9070-8.