Neural Legal Judgment Prediction in English

Ilias Chalkidis∗ **Ion Androutsopoulos**∗
Nikolaos Aletras∗∗
∗ Department of Informatics, Athens University of Economics and Business, Greece
∗∗ Computer Science Department, University of Sheffield, UK
[ihalk,ion]@aueb.gr, n.aletras@sheffield.ac.uk

Abstract

Legal judgment prediction is the task of automatically predicting the outcome of a court case, given a text describing the case's facts. Previous work on using neural models for this task has focused on Chinese; only feature-based models (e.g., using bags of words and topics) have been considered in English. We release a new English legal judgment prediction dataset, containing cases from the European Court of Human Rights. We evaluate a broad variety of neural models on the new dataset, establishing strong baselines that surpass previous feature-based models in three tasks: (1) binary violation classification; (2) multi-label classification; (3) case importance prediction. We also explore if models are biased towards demographic information via data anonymization. As a side-product, we propose a hierarchical version of BERT, which bypasses BERT's length limitation.

1 Introduction

Legal information is often represented in textual form (e.g., legal cases, contracts, bills). Hence, legal text processing is a growing area in NLP with various applications such as legal topic classification (Nallapati and Manning, 2008; Chalkidis et al., 2019), court opinion generation (Ye et al., 2018) and analysis (Wang et al., 2012), legal information extraction (Chalkidis et al., 2018), and entity recognition (Cardellino et al., 2017; Chalkidis et al., 2017). Here, we focus on *legal judgment prediction*, where given a text describing the facts of a legal case, the goal is to predict the court's outcome (Aletras et al., 2016; Şulea et al., 2017; Luo et al., 2017; Zhong et al., 2018; Hu et al., 2018).

Such models may assist legal practitioners and citizens, while reducing legal costs and improving access to justice (Lawlor, 1963; Katz, 2012; Stevenson and Wagoner, 2015). Lawyers and judges can use them to estimate the likelihood of winning a case and come to more consistent and informed judgments, respectively. Human rights organizations and legal scholars can employ them to scrutinize the fairness of judicial decisions unveiling if they correlate with biases (Doshi-Velez and Kim, 2017; Binns et al., 2018).

This paper contributes a new publicly available English legal judgment prediction dataset of cases from the European Court of Human Rights (ECHR).[1] Unlike Aletras et al. (2016), who provide only features from approx. 600 ECHR cases, our dataset is substantially larger (∼11.5k cases) and provides access to the raw text. As a second contribution, we evaluate several neural models in legal judgment prediction for the first time in English. We consider three tasks: (1) binary classification (i.e., violation of a human rights article or not), the only task considered by Aletras et al. (2016); (2) multi-label classification (type of violation, if any); (3) case importance detection. In all tasks, neural models outperform an SVM with bag-of-words (Aletras et al., 2016; Medvedeva et al., 2018), the only method tested in English legal judgment prediction so far. As a third contribution, we use an approach based on data anonymization to study, for the first time, whether the legal predictive models are biased towards demographic information or factual information relevant to human rights. Finally, as a side-product, we propose a hierarchical version of BERT (Devlin et al., 2019), which bypasses BERT's length limitation and leads to the best results.

2 ECHR Dataset

ECHR hears allegations that a state has breached human rights provisions of the European Conven-

[1] The dataset is submitted at https://archive.org/details/ECHR-ACL2019.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4317–4323
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

tion of Human Rights.[2] Our dataset contains approx. 11.5k cases from ECHR's public database.[3] For each case, the dataset provides a list of *facts* extracted using regular expressions from the case description, as in Aletras et al. (2016)[4] (see Fig. 1). Each case is also mapped to *articles* of the Convention that were violated (if any). An *importance score* is also assigned by ECHR (see Section 3).

The dataset is split into training, development, and test sets (Table 1). The training and development sets contain cases from 1959 through 2013, and the test set from 2014 through 2018. The training and development sets are balanced, i.e., they contain equal numbers of cases with and without violations. We opted to use a balanced training set to make sure that our data and consequently our models are not biased towards a particular class. The test set contains more (66%) cases with violations, which is the approximate ratio of cases with violations in the database. We also note that 45 out of 66 labels are not present in the training set, while another 11 are present in fewer than 50 cases. Hence, the dataset of this paper is also a good testbed for few-shot learning.

3 Legal Prediction Tasks

3.1 Binary Violation

Given the facts of a case, we aim to classify it as positive if *any* human rights article or protocol has been violated and negative otherwise.

3.2 Multi-label Violation

Similarly, the second task is to predict which specific human rights articles and/or protocols have been violated (if any). The total number of articles and protocols of the European Convention of Human Rights are 66 up to day. For that purpose, we define a multi-label classification task where no labels are assigned when there is no violation.

3.3 Case Importance

We also predict the importance of a case on a scale from 1 (key case) to 4 (unimportant) in a regression task. These scores, provided by the ECHR, denote a case's contribution in the development of case-law allowing legal practitioners to identify pivotal cases. Overall in the dataset, the scores are: 1 (1096 documents), 2 (904), 3 (2,982) and 4 (6,496), indicating that approx. 10% are landmark cases, while the vast majority (83%) are considered more or less unimportant for further review.

Subset	Cases (C)	Words/C	Facts/C	Articles/C
Train	7,100	2,421	43	0.71
Dev.	1,380	1,931	30	0.96
Test	2,998	2,588	45	0.71

Table 1: Statistics of the ECHR dataset. The size of the label set (ECHR articles) per case (C) is $L = 66$.

4 Neural Models

BiGRU-Att: The fisrt model is a BIGRU with self-attention (Xu et al., 2015) where the facts of a case are concatenated into a word sequence. Words are mapped to embeddings and passed through a stack of BIGRUs. A single case embedding (h) is computed as the sum of the resulting context-aware embeddings ($\sum_i a_i h_i$) weighted by self-attention scores (a_i). The case embedding (h) is passed to the output layer using a sigmoid for binary violation, softmax for multi-label violation, or no activation for case importance regression.

HAN: The Hierarchical Attention Network (Yang et al., 2016) is a state-of-the-art model for text classification. We use a slightly modified version where a BIGRU with self-attention reads the words of each fact, as in BIGRU-ATT, producing fact embeddings. A second-level BIGRU with self-attention reads the fact embeddings, producing a single case embedding that goes through a similar output layer as in BIGRU-ATT.

LWAN: The Label-Wise Attention Network (Mullenbach et al., 2018) has been shown to be robust in multi-label classification. Instead of a single attention mechanism, LWAN employs L attentions, one for each possible label. This produces L case embeddings ($h^{(l)} = \sum_i a_{l,i} h_i$) per case, each one specialized to predict the corresponding label. Each of the case embeddings goes through a separate linear layer (L linear layers in total), each with a sigmoid, to decide if the corresponding label should be assigned. Since this is a multi-label model, we use it only in multi-label violation.

BERT and HIER-BERT: BERT (Devlin et al., 2019) is a language model based on Transformers (Vaswani et al., 2017) pretrained on large corpora.

[2] An up-to-date copy of the European Convention of Human Rights is available at https://www.echr.coe.int/Documents/Convention_ENG.pdf.

[3] See https://hudoc.echr.coe.int. Licensing conditions are compatible with the release of our dataset.

[4] Using regular expressions to segment legal text from ECHR is usually trivial, as the text has a specific structure. See an example from ECHR's Data Repository (http://hudoc.echr.coe.int/eng?i=001-193071).

For a new task, a task-specific layer is added on top of BERT and is trained jointly by fine-tuning on task-specific data. We add a linear layer on top of BERT, with a sigmoid, softmax, or no activation, for binary violation, multi-label violation, and case importance, respectively.[5] BERT can process texts up to 512 wordpieces, whereas our case descriptions are up to 2.6k words, thus we truncate them to BERT's maximum length, which affects its performance. This also highlights an important limitation of BERT in processing long documents, a common characteristic in legal text processing.

To surpass BERT's maximum length limitation, we also propose a hierarchical version of BERT (HIER-BERT). Firstly BERT-BASE reads the words of each fact, producing fact embeddings. Then a self-attention mechanism reads fact embeddings, producing a single case embedding that goes through a similar output layer as in HAN.

5 Experiments

5.1 Experimental Setup

Hyper-parameters: We use pre-trained GLOVE (Pennington et al., 2014) embeddings ($d = 200$) for all experiments. Hyper-parameters are tuned by random sampling 50 combinations and selecting the values with the best development loss in each task.[6] Given the best hyper-parameters, we perform five runs for each model reporting mean scores and standard deviations. We use categorical cross-entropy loss for the classification tasks and mean absolute error for the regression task, Glorot initialization (Glorot and Bengio, 2010), Adam (Kingma and Ba, 2015) with default learning rate 0.001, and early stopping on the development loss.

Baselines: A majority-class (MAJORITY) classifier is used in binary violation and case importance. A second baseline (COIN-TOSS) randomly predicts violation or not in binary violation task. We also compare our methods against a linear SVM with bag-of-words features (most frequent [1, 5]-grams across all training cases weighted by TF-IDF), dubbed BOW-SVM, similar to Aletras et al. (2016) and Medvedeva et al. (2018) for the binary task; multiple one-vs-rest classifiers for the multi-label task; and Support Vector Regression (BOW-SVR) for the case importance prediction.[7]

	P	R	F1
MAJORITY	32.9 ± 0.0	50.0 ± 0.0	39.7 ± 0.0
COIN-TOSS	50.4 ± 0.7	50.5 ± 0.8	49.1 ± 0.7
Non-Anonymized			
BOW-SVM	71.5 ± 0.0	72.0 ± 0.0	71.8 ± 0.0
BIGRU-ATT	87.1 ± 1.0	77.2 ± 3.4	79.5 ± 2.7
HAN	88.2 ± 0.4	78.0 ± 0.2	80.5 ± 0.2
BERT	24.0 ± 0.2	50.0 ± 0.0	17.0 ± 0.5
HIER-BERT	**90.4** ± 0.3	**79.3** ± 0.9	**82.0** ± 0.9
Anonymized			
BOW-SVM	71.6 ± 0.0	70.5 ± 0.0	70.9 ± 0.0
BIGRU-ATT	**87.0** ± 1.0	76.6 ± 1.9	78.9 ± 1.9
HAN	85.2 ± 4.9	**78.3** ± 2.0	**80.2** ± 2.7
BERT	17.0 ± 3.0	50.0 ± 0.0	25.4 ± 0.4
HIER-BERT	85.2 ± 0.3	78.1 ± 1.3	80.1 ± 1.1

Table 2: Macro precision (P), recall (R), F1 for the **binary violation** prediction task (± std. dev.).

5.2 Binary Violation Results

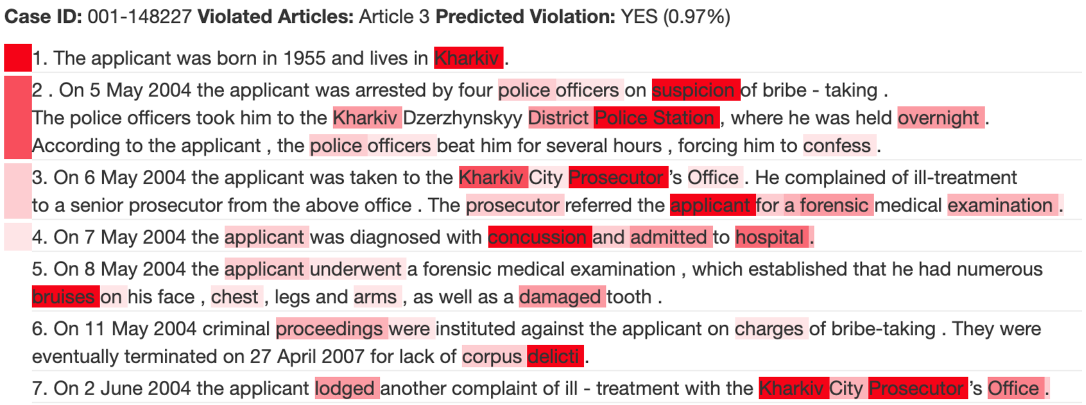
Table 2 (upper part) shows the results for binary violation. We evaluate models using macro-averaged precision (P), recall (P), F1. The weak baselines (MAJORITY, COIN-TOSS) are widely outperformed by the rest of the methods. BIGRU-ATT outperforms in F1 (79.5 vs. 71.8) the previous best performing method (Aletras et al., 2016) in English judicial prediction. This is aligned with results in Chinese (Luo et al., 2017; Zhong et al., 2018; Hu et al., 2018). HAN slightly improves over BIGRU-ATT (80.5 vs. 79.5), while being more robust across runs (0.2% vs. 2.7% std. dev.). BERT's poor performance is due to the truncation of case descriptions, while HIER-BERT that uses the full case leads to the best results. We omit BERT from the following tables, since it performs poorly.

Fig. 1 shows the attention scores over words and facts of HAN for a case that ECHR found to violate Article 3, which prohibits torture and 'inhuman or degrading treatment or punishment'. Although fact-level attention wrongly assigns high attention to the first fact, which seems irrelevant, it then successfully focuses on facts 2–4, which report that police officers beat the applicant for several hours, that the applicant complained, was referred for forensic examination, diagnosed with concussion etc. Word attention also successfully focuses on words like 'concussion', 'bruises', 'damaged', but it also highlights entities like 'Kharkiv', its 'District Police Station' and 'City Prosecutor's office', which may be indications of bias.

[5] The extra linear layer is fed with the 'classification' token of the BERT-BASE version of Devlin et al. (2019).

[6] Ranges: GRU hidden units {200, 300, 400}, number of stacked BIGRU layers {1, 2}, batch size {8, 12, 16}, dropout rate {0.1, 0.2, 0.3, 0.4}, word dropout rate {0.0, 0.01, 0.02}.

[7] We tune the hyper-parameters of BOW-SVM/SVR and select kernel (RBF, linear) with a grid search on the dev. set.



Case ID: 001-148227 **Violated Articles:** Article 3 **Predicted Violation:** YES (0.97%)

1. The applicant was born in 1955 and lives in Kharkiv .

2 . On 5 May 2004 the applicant was arrested by four police officers on suspicion of bribe - taking .
The police officers took him to the Kharkiv Dzerzhynskyy District Police Station , where he was held overnight .
According to the applicant , the police officers beat him for several hours , forcing him to confess .

3. On 6 May 2004 the applicant was taken to the Kharkiv City Prosecutor 's Office . He complained of ill-treatment
to a senior prosecutor from the above office . The prosecutor referred the applicant for a forensic medical examination .

4. On 7 May 2004 the applicant was diagnosed with concussion and admitted to hospital .

5. On 8 May 2004 the applicant underwent a forensic medical examination , which established that he had numerous
bruises on his face , chest , legs and arms , as well as a damaged tooth .

6. On 11 May 2004 criminal proceedings were instituted against the applicant on charges of bribe-taking . They were
eventually terminated on 27 April 2007 for lack of corpus delicti .

7. On 2 June 2004 the applicant lodged another complaint of ill - treatment with the Kharkiv City Prosecutor 's Office .

Figure 1: Attention over words (colored words) and facts (vertical heat bars) as produced by HAN.

Models Biases: We next investigate how sensitive our models are to demographic information appearing in the facts of a case. Our assumption is that an unbiased model should not rely on information about nationality, gender, age, etc. To test the sensitivity of our models to such information, we train and evaluate them in an anonymized version of the dataset. The data is anonymized by using SPACY's (`https://spacy.io`) Named Enity Recognizer, replacing all recognized entities with type tags (e.g., 'Kharkiv' → LOCATION).

While neural methods seem to exploit named entities among other information, as in Figure 1, the results in Table 2 indicate that performance is comparable even when this information is masked, with the exception of HIER-BERT that has quite worse results (2%) compared to using non-anonymized data, suggesting model bias. We speculate that HIER-BERT is more prone to overfitting compared to the other neural methods that rely on frozen GLOVE embeddings, because the embeddings of BERT's wordpieces are trainable and thus can freely adjust to the vocabulary of the training documents including demographic information.

5.3 Multi-label Violation Results

Table 3 reports micro-averaged precision (P), recall (R), and F1 results for all methods, now including LWAN, in multi-label violation prediction. The results are also grouped by label frequency for all (OVERALL), FREQUENT, and FEW labels (articles), counting frequencies on the training subset.

We observe that predicting specific articles that have been violated is a much more difficult task than predicting if *any* article has been violated in a binary setup (cf. Table 2). Overall, HIER-BERT outperforms BIGRU-ATT and LWAN (60.0 vs. 57.6 micro-F1), which is tailored for multi-labeling tasks, while being comparable with HAN (60.0 vs. 59.9 micro-F1). All models under-perform in labels with FEW training examples, demonstrating the difficulty of few-shot learning in ECHR legal judgment prediction. The main reason is that labels in the FEW group, 11 in total, are extremely rare and have been assigned in 1.25% of the documents across all datasets, while the most frequent 4 labels overall (Articles 3, 5, 6 and 13) have been assigned in approx. 42% of the documents.

OVERALL (all labels)			
	P	**R**	**F1**
BOW-SVM	56.3 ± 0.0	45.5 ± 0.0	50.4 ± 0.0
BIGRU-ATT	62.6 ± 1.2	50.9 ± 1.5	56.2 ± 1.3
HAN	65.0 ± 0.4	**55.5** ± 0.7	59.9 ± 0.5
LWAN	62.5 ± 1.0	53.5 ± 1.1	57.6 ± 1.0
HIER-BERT	**65.9** ± 1.4	55.1 ± 3.2	**60.0** ± 1.3
FREQUENT (≥50)			
BOW-SVM	56.3 ± 0.0	45.6 ± 0.0	50.4 ± 0.0
BIGRU-ATT	62.7 ± 1.2	52.2 ± 1.6	57.0 ± 1.4
HAN	65.1 ± 0.3	**57.0** ± 0.8	**60.8** ± 1.3
LWAN	62.8 ± 1.2	54.7 ± 1.2	58.5 ± 1.0
HIER-BERT	**66.0** ± 1.4	56.5 ± 3.3	**60.8** ± 1.3
FEW ([1,50))			
BOW-SVM	-	-	-
BIGRU-ATT	36.3 ± 13.8	03.2 ± 23.1	05.6 ± 03.8
HAN	30.2 ± 35.1	01.6 ± 01.2	02.8 ± 01.9
LWAN	24.9 ± 06.3	**07.0** ± 04.1	**10.6** ± 05.2
HIER-BERT	**43.6** ± 14.5	05.0 ± 02.8	08.9 ± 04.9

Table 3: Micro precision, recall, F1 in **multi-label violation** for all, frequent, and few training instances.

5.4 Case Importance Results

Table 4 shows the mean absolute error (MAE) obtained when predicting case importance. Surprisingly, MAJORITY outperforms the rest of the methods. As already noted in Section 3, the distribution of importance scores is highly skewed in favour of the majority class, thus MAJORITY can correctly predict the score in most cases with zero mean absolute error (MAE). BOW-SVR performs worse

	MAE	SPEARMAN's ρ
MAJORITY	**.369** ± .000	*N/A**
BOW-SVR	.585 ± .000	.370 ± .000
BIGRU-ATT	.539 ± .073	.459 ± .034
HAN	.524 ± .049	.437 ± .018
HIER-BERT	.437 ± .018	**.527** ± .024

Table 4: Mean Absolute Error and Spearman's ρ for **case importance**. Importance ranges from 1 (most important) to 4 (least). * Not Applicable.

than BIGRU-ATT, while HAN is 10% and 3% better, respectively. HIER-BERT further improves the results, outperforming HAN by 17%.

While MAJORITY has the lowest mean absolute error, it cannot distinguish important from unimportant cases, thus it is practically useless. To evaluate the methods on that matter, we measure the correlation between the gold scores and each method's predictions with SPEARMAN's ρ. HIER-BERT has the best ρ (.527), indicating a moderate positive correlation (> 0.5), which is not the case for the rest of the methods. The overall results indicate that a case's importance cannot be predicted solely by the case facts and possibly also relies on background knowledge (e.g., judges' experience, court's history, rarity of article's violation).

5.5 Discussion

We can only speculate that HAN's fact embeddings distill importance-related features from each fact, allowing its second-level GRU to operate on a sequence of fact embeddings that are being exploited by the fact-level attention mechanism and provide a more concise view of the entire case. The same applies to HIER-BERT, which relies on BERT's fact embeddings and the same fact-level attention mechanism. By contrast, BIGRU-ATT operates on a single long sequence of concatenated facts, making it more difficult for its BIGRU to combine information from multiple, especially distant, facts. This may explain the good performance of HAN and HIER-BERT across all tasks.

6 Related Work

Previous work on legal judgment prediction in English used linear models with features based on bags of words and topics to represent legal textual information extracted from cases (Aletras et al., 2016; Medvedeva et al., 2018).

More sophisticated neural models have been considered only in Chinese. Luo et al. (2017) use HANs to encode the facts of a case and a subset of predicted relevant law articles to predict criminal charges that have been manually annotated. In their experiments, the importance of few-shot learning is not taken into account since the criminal charges that appear fewer than 80 times are filtered out. However in reality, a court is able to judge even under rare conditions. Hu et al. (2018) focused on few-shot charges prediction using a multi-task learning scenario, predicting in parallel a set of discriminative attributes as an auxiliary task. Both the selection and annotation of these attributes are manually crafted and dependent to the court. Zhong et al. (2018) decompose the problem of charge prediction into different subtasks that are tailored to the Chinese criminal court using multitask learning.

7 Limitations and Future Work

The neural models we considered outperform previous feature-based models, but provide no justification for their predictions. Attention scores (Fig. 1) provide some indications of which parts of the texts affect the predictions most, but are far from being justifications that legal practitioners could trust; see also Jain and Wallace (2019). Providing valid justifications is an important priority for future work and an emerging topic in the NLP community.[8] In this direction, we plan to expand the scope of this study by exploring the automated analysis of additional resources (e.g., relevant case law, dockets, prior judgments) that could be then utilized in a multi-input fashion to further improve performance and justify system decisions. We also plan to apply neural methods to data from other courts, e.g., the European Court of Justice, the US Supreme Court, and multiple languages, to gain a broader perspective of their potential in legal justice prediction. Finally, we plan to adapt bespoke models proposed for the Chinese Criminal Court (Luo et al., 2017; Zhong et al., 2018; Hu et al., 2018) to data from other courts and explore multitask learning.

Acknowledgements

This work was partly supported by the Research Center of the Athens University of Economics and Business.

[8] http://aclweb.org/anthology/W18-5400

References

Nikolaos Aletras et al. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93.

Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *CHI*, pages 377:1–377:14.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. Legal NERC with ontologies, Wikipedia and curriculum learning. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 254–259.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law (ICAIL)*, pages 19–28, London, UK.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and prohibition extraction using hierarchical RNNs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 254–259.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop (NLLP) of NAACL-HLT*, pages 78–87.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Finale Doshi-Velez and Ben Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 487–498.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Daniel Martin Katz. 2012. Quantitative legal prediction-or-how I learned to stop worrying and start preparing for the data-driven future of the legal services industry. *Emory Law Journal*, 62:909.

Diederik P. Kingma and Jim Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Reed C Lawlor. 1963. What computers can do: Analysis and prediction of judicial decisions. *American Bar Association Journal*, pages 337–344.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Conference on Empirical Methods in Natural Language Processing (EMNLP))*, pages 2727–2736.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the European Court of Human Rights: Looking into the crystal ball. In *Proceedings of the Conference on Empirical Legal Studies*.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 1101–1111.

Ramesh Nallapati and Christopher D. Manning. 2008. Legal docket classification: Where machine learning stumbles. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Conference on Empirical Methods in Natural Language Processing (EMNLP))*, pages 438–446.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Dru Stevenson and Nicholas J Wagoner. 2015. Bargaining in the shadow of big data. *Florida Law Review*, 67:1337.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the law area and decisions of French Supreme Court cases. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP)*, pages 716–722.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

William Yang Wang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. 2012. Historical analysis of legal opinions with a sparse mixed-effects latent variable model. In *ACL*, pages 740–749.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2048–2057.

Zichao Yang et al. 2016. Hierarchical attention networks for document classification. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 1480–1489.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Conference on Empirical Methods in Natural Language Processing (EMNLP))*, pages 1854–1864.

Haoxi Zhong, Guo Zhipeng, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3540–3549.