Machine learning for explaining and ranking the most influential matters of law

Max R. S. Marques
Institut Mines-Télécom Atlantique
Brest, France
max.sobrozamarques@imt-atlantique.fr

Tommaso Bianco
Predictice
Paris, France
tommaso.bianco@predictice.com

Maxime Roodnejad
Predictice
Paris, France
maxime.roodnejad@gmail.com

Thomas Baduel
Predictice
Paris, France
thomas.baduel@predictice.com

Claude Berrou
Institut Mines-Télécom Atlantique
Brest, France
claude.berrou@imt-atlantique.fr

ABSTRACT

In this work, we propose a novel method in order to rank the most relevant legal principle citations in law-cases to support a certain motion. The first score relies on feature importance metrics, where each law article is a feature supplied to a classifier for the decision outcome. The second score is based on word embeddings text similarity. As a result, our method outperforms the baseline techniques based on feature importance selection and Information Retrieval methods in the ranking evaluation relevance criteria.

CCS CONCEPTS

• **Applied computing** → **Law**.

KEYWORDS

Machine learning, model explanation, feature selection, argument mining, legal principles, text similarity, word embeddings, NLP

ACM Reference Format:
Max R. S. Marques, Tommaso Bianco, Maxime Roodnejad, Thomas Baduel, and Claude Berrou. 2019. Machine learning for explaining and ranking the most influential matters of law. In *Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19), June 17–21, 2019, Montreal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3322640.3326734

1 INTRODUCTION

In legal decisions, the law professionals cite facts and legal principles with the purpose of supporting their claims or decisions about the case. The increasing availability of legal documents and recent advances in text information retrieval enable us to build new tools for citation analysis (e.g., Shepherds and KeyCite), thus giving legal users the means to better evaluate the importance of facts and law principles within each legal document.

ICAIL '19, June 17–21, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6754-7/19/06…$15.00
https://doi.org/10.1145/3322640.3326734

Legal principles citations are references to law articles or codes, and can be used either by a plaintiff/defendant in order to give foundation to their argument, or by the judge/court, to justify the decision about a request. The outcome of a judgment is based on the interpretation of legal and fact principles. On the particular case of countries or provinces that adopt *the Civil law system*, (e.g., Japan and France), legal principles citation information is very relevant to determine the motions of a legal case. Moreover, several authors (e.g. [7, 11, 13]) underline the benefits of relevant law article extraction systems to improve related legal classification tasks. According to them, specially on *Civil law* use cases, the most challenging problems to consider is the usage of law articles on legal classification tasks are: (1) the scaling problems due to the large number of legal principles, (2) the multi-label nature of legal principles extraction task.

In recent years, there has been considerable interest in Natural Language Processing (NLP) to use Deep Neural Networks in order to extract features from unstructured textual content. In NLP literature, methods that construct distributed word feature vectors from unsupervised learning of raw text were introduced such as *word2vec* [12] and *fastText* [2]. Similarly, this interest is growing in legal analytics [4].

A common issue in the Artificial Intelligence field including NLP is the trade-off between prediction accuracy and interpretability. Simpler models such as the linear model despite of easy interpretability tend to have a poor accuracy prediction. On the other hand, complex models such as Deep Learning are more accurate but less interpretable (e.g. [6]). Some progress has been made in recent works in the field (e.g. [10]). For instance, Lundberg and Lee [10] proposed a method based on game theory called *SHAP* (*SHapley Additive exPlanations*) in order to understand (and interpret) the decisions made by Machine Learning (ML) methods. Despite latest advances in the domain, the lack of interpretability still remains a major problem in predictive legal analytics field. Better understanding of decisions made by ML tools can contribute to improve existing applications in legal analytics and it can also provide guarantees to limit the bias (e.g, genre or social status).

We propose a method to extract and rank the most relevant legal principles to support a specific motion in a law-case. In this case, this method allows us to understand which law principles citations contribute the most to a case. Our method is suited to

legal jurisprudence corpus that are based on *Civil law system* and it scales to large unbalanced datasets containing millions of different law articles and hundreds of different motions. Our algorithm combines two different techniques: document similarity computation based on neural word embeddings and a bagging approach of *SHAP* regression values. Both techniques measure the relevance of a law article citation regarding the motion prediction.

2 DATASET CONSTRUCTION

The dataset used in this work is composed of two types of documents: legal jurisprudence cases and legal principles text.

2.1 Legal jurisprudence cases

The data collection contains about 2 millions legal documents of French court of second instance (all categories of decisions from Cours d'appel) decision judgments and their respective related motions. A motion can be defined as an oral or written request that a party makes to the court for a ruling or an order on a particular point. For example, an employee that claims to have been unfairly dismissed by his employer might refer to the motion for *unfair dismissal*. Each legal decision document can contain one or more motions. The set of possible motions was enumerated by legal experts and contains 610 different motions.

The motions and law articles citations were extracted from the text of jurisprudence cases by using regular expression matching. The same methodology of annotation was applied in other studies [17] also based on *the Civil law system*. This process build a dataset composed pairs of *legal articles citations* and *motions* for each document. This dataset is highly unbalanced since some motions appear in most documents such as *repayment of legal expenses in a legal case* while others are present only in a few number of decisions (see Fig. 1).

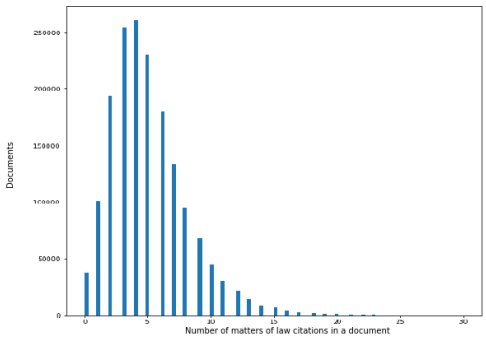

Figure 1: Histogram of number of articles law citations per document of annotated legal decision documents.

The dataset is highly unbalanced since more than 300 motions (of a total of 610) contain less than 400 examples of documents and for example, the motion *Dismissal without real and serious reason* contains more than 75K examples of documents. For these reasons, we increased the correlation of cited legal principles features by using an external source of information. This external source contains the different versions of a law article. We exploited this source of information by using information of relevant legal principles articles graph extracted from *LEGIFRANCE* database.

2.2 Legal principles database

Legal principles were collected from the French government database website of legal principles *LEGIFRANCE*[1]. This open source database contains references and texts of more than 1.5 millions of French law articles from 75 different Law Codes (e.g., *Code du Travail*, *Code Civil*, ...). The database of French law articles is frequently updated. New laws are being continuously created while others are amended, invalidated, cease to exist or just modified. This issue was also referenced in [7] work.

3 OUR METHOD

In Fig. 2, we pictured the general pipeline architecture of our method. The final result of the relevance of law article is obtained by combining the scores of *SHAP* values predictions and law article text similarity. In this section, each one of these steps is explained in more details.

3.1 Legal citation feature extraction based on graph analysis

Algorithm 1 is applied to obtain the set of relevant law successors from cited legal principles of a given document j. The intention of applying the method is to encode in a simple manner the information of several temporal related versions of the same cited legal principles. The problem of temporal normative change is addressed in [1].

Algorithm 1 Generate features of the legal document from the graph of law abrogations $\mathcal{G} = (V, E)$, where V is a set of vertex (legal principles), $E \subseteq \{(u, v); u, v \in V\}$ the set of oriented edges (relations of law abrogations) and T_j is the set of legal principles cited in the document d_j.

Require: $\mathcal{G} = (V, E)$; $\{t_1, t_2, ..., t_{n_j}\} = T_j$
1: $features := T_j$
2: **for** each $t_i \in T_j$ **do**
3: $\quad features := features \cup extractLawSuccessors(t_i, \mathcal{G})$
4: **end for**
5: **return** $features$

The recursive $extractLawSuccessors(t_i, \mathcal{G})$ applies the strategy DFS (Depth-first search) in the graph $\mathcal{G}$ to obtain all children nodes of t_i including lead nodes that represent not repealed legal principles. Each component $x_a^{(j,m)}$ of the vector $\boldsymbol{x}^{(j,m)} = [x_1^{(j,m)}, x_2^{(j,m)}, ..., x_a^{(j,m)}, ..., x_{n_A}^{(j,m)}]$ encodes the existence of a citation of the law article a in the document j. The vector $\boldsymbol{x}^{(j,m)}$ is obtained using the output of Algorithm 1. The information of the label (or motion) of each document j is encoded by the multi-label binary vector $\boldsymbol{y}^{(j)}$ where $y_m^{(j)} = 1$ if the motion m is applicable to the legal document j and $y_m^{(j)} = 0$ otherwise.

[1] https://www.legifrance.gouv.fr/

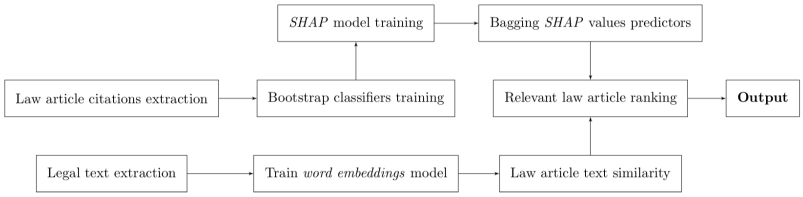

Figure 2: Pipeline architecture of our method

3.2 Explaining motion predictions

The *SHAP* framework [10] provides a post-hoc model agnostic tool to explain individual predictions performed by a trained classifier. In other words, the method provides for users a technique to understand the rationale of a specific prediction made by the the ML model. Models with this property are called locally interpretable models. In contrast, globally interpretable models are used to extract general feature interpretations based on the entire dataset.

The *SHAP* values is an estimator that measures the importance of each feature for the overall model prediction of a single sample. In the case where all features are binary positive *SHAP* values mean that features contribute positively to the class prediction while inverse relation for negative *SHAP* values is true. We denote the *SHAP* values applied over the input vector $\boldsymbol{x}^{(j,m)}$ of a trained model to predict the motion m as $\boldsymbol{\phi}^{(j,m)} \in \mathbb{R}^{n_A}$.

3.3 Motion prediction model

Our method is based on multi-target classification. We investigate the effect of each feature (citation of a law article) to determine the presence of a motion in a legal case. The *XGBoost*[5] (or *Extremely Gradient Boosting*) model is a tree-based ensemble classifier that provides that deals more efficiently with data sparseness. The *SHAP* method is also adapted to this classification algorithm in terms of scalability and execution time. According to Liu et al. [9], the accuracy of feature selection methods based on decision trees classifiers can be drastically compromised due to the data imbalance and bagging approaches can decrease this effect. For this reason, we propose a multi-target *one-vs-rest* bootstrapping technique adapted to the *SHAP* framework. For a given motion m, the dataset to train is the set of pairs $\mathcal{D}_m = \{(\boldsymbol{x}^{(j,m)}, y_m^{(j)}) | 1 \leq j \leq n_D\}$. The bootstrap technique applied over *SHAP* values consists in four steps applied for each motion:

(1) Train n *XGBoost* classifiers with the same proportion of positive and negative samples. The number of training samples of each classifier is $2 \min\{|\mathcal{D}_m^+|, |\mathcal{D}_m^-|\}$. These samples are uniformly sampled from $\mathcal{D}_m^+$ a and $\mathcal{D}_m^-$ with replacement.
(2) Train a *SHAP* estimator for each classifier using the same subsampled dataset of previous step. The prediction performed for each classifier of *SHAP* regression estimator values of positive samples $\mathcal{D}_m^+$.
(3) Averaging the *SHAP* values of the same positive samples obtained using n different classifiers

$$\boldsymbol{\phi}_B^{(j,m)} \leftarrow \frac{\boldsymbol{\phi}^{(j,m)} - \min(\boldsymbol{\phi}^{(j,m)})}{\max(\boldsymbol{\phi}^{(j,m)}) - \min(\boldsymbol{\phi}^{(j,m)})}; y_m^{(j)} = 1$$

where $\mathcal{D}_m^+ = \{(\boldsymbol{x}^{(j,m)}, y_m^{(j)}) | y_m^{(j)} = 1\}$, $\mathcal{D}_m^- = \mathcal{D}_m - \mathcal{D}_m^+$ and $|.|$ is the operator of cardinality of a set.

3.4 Law article text similarity based on neural word embeddings approach

We trained *word embeddings* using *FastText skip-gram* method on text of both legal databases using the same hyper-parameters used in the *FastText* paper [2].

We applied a *Bag-of-words* (average of vectors) method in order to obtain sentence vectors of motion and law article sentences. Then, for a given law article and motion m, we find the maximum cosinus similarity score between all sentence vectors of the law article text and the sentence vector of the motion m. Thus, the score of cosinus similarity is normalized considering all law articles citations of a given legal document j and motion m. This normalized vector of similarity scores is $\boldsymbol{s}^{(j,m)}$. The final ranking score of our method is given by weighting the two normalized scores:

$$\alpha \boldsymbol{\phi}_B^{(j,m)} + (1-\alpha)\boldsymbol{s}^{(j,m)} \odot (\boldsymbol{\phi}_B^{(j,m)}. > 0); y_m^{(j)} = 1 \quad (1)$$

where $\odot$ is a element-wise multiplication operator and $. >$ applies a mask over all non-negative vector components. As a result, only law articles citations with positive values of the final ranking score are selected because positive *SHAP* values contribute positively to the class prediction.

4 BASELINES

To the best of our knowledge, there is no method in the literature that uses local interpretations of legal principles features to predict motions in a legal case. In this subsection, we introduce methods of feature selection that provide a framework for understanding the global impact of certain features. Feature selection is the ensemble of techniques commonly employed to remove input features to be used by a classifier. For the interpretation of methods based on ML or statistical, it is important to identify all relevant variables, including those carrying redundant information.

4.1 Information Retrieval methods

4.1.1 Mutual Information (MI). This method uses the Information Entropy of each feature for the classification[15]. The MI term

calculates the contribution (presence or absence) of a legal principle to the correct classification estimated from the empirical frequency of events.

4.2 Methods based on Feature Importance

4.2.1 Tree-based methods. The feature importance is a family of methods that calculates the gain of splitting each tree branch based on entropy or Gini criterion of tree-based ensembles models such as *Random forest*[3] and *XGBoost* in a classification task. We tested this framework training *one-versus-rest XGBoost* [5] classifiers of motions. We used this approach to obtain scores of the importance of features (law articles citations) for a specific motion.

4.2.2 Recursive Feature Elimination (RFE). The *Recursive Feature Elimination* algorithm is a method of automatic selection of features. It starts with a decision tree classifier built on all variables. The RFE eliminates one or a few variables at a time which contribute to a heavy workload on the CPU and memory usage. We could not compare this approach in our experiments because of this limitation.

4.2.3 Boruta. The *Boruta* method initially proposed in [8] is a method to the automatic selection features of number of features.*Boruta* method is computationally less costly than RFE. Nevertheless, statistical methods require a considerably high number of measures (or classifier results) in order to decrease the uncertainty of results. For this reason, this method can be considered computationally costly. A faster implementation of initial *Boruta* algorithm (called *BoostARoota*) is available[2].

5 EXPERIMENTS

We designed the following experiments with the aim of evaluate the *cognitive relevance*[14] of the *citation pertinence* generated automatically by our system.

5.1 Ranking metric

All methods were evaluated using the Normalized Discounted Cumulative Gain (NDCG) [16], in order to compare the ranking of most relevant *legal principle* citations for motion classification. This evaluation metric is a measure of ranking quality used for IR systems. The DGC is very used in search engine systems where there are a large number of choices of most relevant documents and it is important to give a rank based on relevance. The DCG measures the gain of *legal principles* sorted from the most relevant one to the least one with regards to the motion. The gain is accumulated from the top of the result list to the bottom, with the gain of each result discounted at lower ranks. The DCG can be computed by the following equation:

$$p@DCG_m = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{2}$$

where $rel_i \in \{0, 1, ..., l\}$, i is the position in the ranking and l is the maximum value of relevance scale. In the case of most relevant legal principles identification, we considered $l = 4$ and $p \in [1, 2, 3, 4, 5, 6]$.

[2]https://github.com/chasedehan/BoostARoota

The optimum value of $p@NDCG_m$ is equal to 1. The criteria established to decide the classification of legal principles into one of these categories were determined by getting the majority vote of three different legal experts. In the ranking evaluation score, we consider to penalize the relevance of old law articles over recent pertinent legal principles. The final metric $p@NDCG$ is obtained by averaging the score $p@NDCG_m$. of all samples that regard the motion m.

5.2 Relevance rate metric

We consider a second metric to evaluate the automatic selection of most relevant legal principles because the NDCG metric is not suited to verify the pertinence of law articles selection. The relevance rate metric is defined as the number of relevant law articles among the number of selected law articles for the method:

$$k@RR_m = \frac{|\{i|rel_i > 2, 1 \leq i \leq k\}|}{k} = \frac{\#relevant\ law\ articles}{\#selected\ law\ articles} \tag{3}$$

The final metric $k@RR$ is obtained by averaging the scores of the same motion.

6 RESULTS

6.1 Global interpretations

All baselines use the binary input vectors extracted from Algorithm 1. Extracted features contain the information about law articles successors of cited legal principles. Techniques of regularization such as pruning the deep of trees were also tested for method based on *Boruta* as recommended in [8]. We applied a stratified sampling on the the number of samples per motion in order to obtain a test set with diversity regarding the distribution of samples. Legal experts ranked the ten first results of each method based on their relevance to support one of chosen motions. In Fig. 5 and Fig. 6, we pictured respectively the results of ranking metric and relevance rate.

Three different scenarios of our method were considered in our experiments in order to test the influence of applied techniques on global ranking results. The first scenario uses only the score obtained from bootstrap training of *SHAP* estimator of *XGBoost* classifiers without considering the text similarity score ($\alpha = 1.0$). The second scenario uses *pre-trained* word vectors on French Wikipedia corpus to text similarity part. The third scenario uses *pre-trained* FastText word embedding vectors on French legal domain corpus. According to the results, the text similarity of law article text globally improves the ranking results. The improvement from text similarity part is more perceptible in relevance rate results than in NDCG ranking.

In the most part of scenarios, our method outperforms the existing techniques of Feature Selection in both metrics (ranking and relevance rate). Other techniques based on *one-versus-rest* decision trees classifiers demonstrated not be suited to to this problem in view of decorrelation between human ranking evaluation and produced results. Most part of produced errors using these methods is due to the fact that selected *law articles* are pertinent to specific motions of the *rest* classification instead of being relevant to the desired motion m.

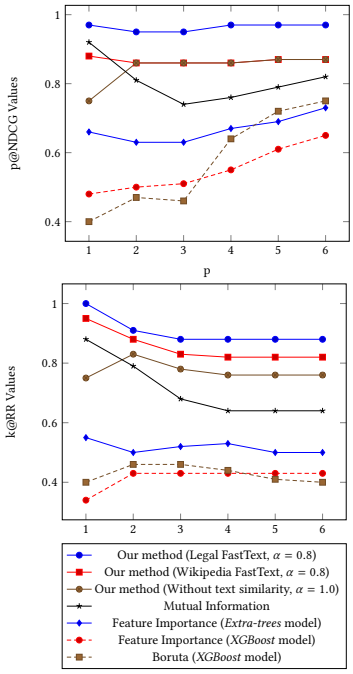

Figure 3: (a) Mean of NDCG ($p@NDCG$) values in function of p. (b) Mean of Relevance Rate($k@RR$) values in function of k (number of selected law articles)

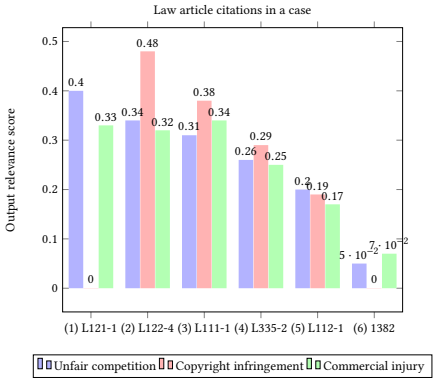

Figure 4: Output relevance score for each motion of cited law articles in French jurisprudence legal case.

6.2 Local interpretations

Fig. 4 depicts the output of our scoring method for interpreting one example of case-law. In this case-law, multiples motions were identified: *unfair competition*, *copyright infringement* and *commercial damages*. The system evaluates the importance of each legal citation for supporting a certain motion. For instance, the motions *unfair competition* and *commercial damages* are probably correlated since similar importance values were attributed for both. We can also infer that the (1) *Article L121-1 du Code de la propriété intellectuelle* is particularly pertinent for these motions. In contrast, the (1) *Article L121-1* is less relevant for the motion *copyright infringement* for this particular document.

7 CONCLUSIONS

We proposed a new method that can be used to local or global interpretations of multi-label *one-versus-rest* classifiers. Our experiments prove that it is possible to automatically detect the *most salient law* articles relative to a particular motion by using techniques of interpretability of ML models. The proposed method is also compatible to high imbalanced dataset with a large number of features, as is the case for motions argumented with few versus several law articles. From the legal field point of view, we demonstrated that word embeddings trained on legal corpus can improve the results. An interesting direction of research is to aggregate the information about facts. Going forward, our results suggest that the interpretability of ML models can address important questions in legal domain.

REFERENCES

[1] Michał Araszkiewicz. 2013. Time, Trust and Normative Change. On Certain Sources of Complexity in Judicial Decision-Making. In *International Workshop on AI Approaches to the Complexity of Legal Systems*. Springer, 100–114.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[4] Ilias Chalkidis and Dimitrios Kampas. 2018. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law* (dec 2018). https://doi.org/10.1007/s10506-018-9238-9

[5] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[6] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 93.

[7] Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. 2017. Network Analysis in the Legal Domain: A complex model for European Union legal sources. *Journal of Complex Networks* 6, 2 (2017), 243–268.

[8] Miron B Kursa, Witold R Rudnicki, et al. 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software* 36, i11 (2010).

[9] Tian-Yu Liu. 2009. Easyensemble and feature selection for imbalance data sets. In *Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS'09. International Joint Conference on*. IEEE, 517–520.

[10] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[11] 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 2017 Conference on Empirical Methods in Natural Language Processing*. 2727–2736.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[13] Yannis Panagis, Urska Sadl, and Fabien Tarissan. 2017. Giving every case its (legal) due The contribution of citation networks and text similarity techniques to legal studies of European Union law. In *30th International Conference on Legal Knowledge and Information Systems (JURIX'17)*, Vol. 302. IOS Press, 59–68.

[14] Marc Van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law* 25, 1 (2017), 65–87.

[15] Jorge R Vergara and Pablo A Estévez. 2014. A review of feature selection methods based on mutual information. *Neural computing and applications* 24, 1 (2014), 175–186.

[16] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG type ranking measures. In *Conference on Learning Theory*. 25–54.

[17] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *CoRR* abs/1807.02478 (2018).