Legal Knowledge and Information Systems
E. Francesconi et al. (Eds.)

doi:10.3233/FAIA220461

Recognising Legal Characteristics of the Judgments of the European Court of Justice: Difficult but Not Impossible

Alessandro CONTINI [a] Sebastiano PICCOLO [a,1], Lucia LOPEZ ZURITA [a], and Urska SADL [a]

[a] *Copenhagen University, Faculty of Law, iCourts*
ORCiD ID: Alessandro Contini https://orcid.org/0000-0002-8999-2610, Sebastiano Piccolo https://orcid.org/0000-0002-6986-3344, Lucia Lopez Zurita https://orcid.org/0000-0001-6092-0114, Urska Sadl https://orcid.org/0000-0003-4635-3816

Abstract. Machine learning has improved significantly during the past decades. Computers perform remarkably in formerly difficult tasks. This article reports the preliminary results on the prediction of two characteristics of judgments of the European Court of Justice, which require the knowledge of concepts and doctrines of European Union law and judicial decision-making: The legal importance (doctrinal outcome) and leeway to the national courts and legislators (deference). The analysis relies on 1704 manually labelled judgments and trains a set of classifiers based on word embedding, LSTM, and convolutional neural networks. While all classifiers exceed simple baselines, the overall performance is weak. This suggests first, that the models learn meaningful representations of the judgments. Second, machine learning encounters significant challenges in the legal domain. These arise doe to the small training data, significant class imbalance, and the characteristics of the variables requiring external knowledge.

The article also outlines directions for future research.

Keywords. Classification, European Court of Justice, Word embedding, LSTM, CNN

1. Introduction

Deep neural networks (DNN) and computer hardware have expanded the range of tasks in which machines outperform humans [1]. Examples include the remarkable progress in computer vision [2], natural language processing [3] and gaming [4]. Artificial intelligence has also transformed the legal profession, providing sophisticated tools for implementing computational legal reasoning, thus enabling argument extraction from legal texts [5,6]. That said, the legal analysis of rights, duties, precedent, and legal development (legal doctrine) has seemingly remained a safe space of a trained human lawyer. A deep-seated opinion is that law is a distinctively human domain, involving deep under-

[1] Corresponding Author: Sebastiano Piccolo, jvt612@ku.dk.

standing of legal sources, situation sense, the ability to read legal texts between the lines, constructing the systems of knowledge [7]. In sum, machines can not (for now) conclusively answer questions about the content of the law. This raises the question exactly how much machine learning can contribute to the analysis of judicial decisions.

The article develops classifiers based on word embedding, LSTM, and CNN (convolutional neural network), which consider the text of a judgment to *1)* Predict the legal importance of the Court's judgment (whether the Court makes a strong contribution to the legal doctrine, such as creating new concepts or principles); and *2)* Detect whether the Court gives the national judge or legislator a leeway to adopt the final decision (defers the final decision about the law to the national court or the legislator). The latter aspect is particular to European Union law, calling for the division of labor between the Court and the national courts. The article trains the classifier on the full judgments and on single paragraphs of the judgments, aggregating the single scores to obtain predictions on the judgment level.

The findings confirm the expectation that predicting legal importance is harder than detecting deference. Concretely, predicting single paragraphs and aggregating their scores is sub-optimal (around 25% lower than the performance of a classifier trained on the full judgment). Moreover, classifiers based on LSTM perform better than those based on CNN. The best score on predicting deference is F1=0.463, while the best score on predicting doctrinal outcome is F1=0.376. The findings echoes the observation from Habernal et al. [6] that legal experts rely on the context beyond the single paragraph used as input for their algorithm to label the arguments. A number of factors contribute to the weak performance of the algorithms: a relatively small training set, the high class imbalance, and the fact that the selected variables require extensive knowledge of complex legal concepts and legal doctrine. Given that all factors are intrinsic to the legal domain, future research should focus on developing more sophisticated models.

2. Data and Methods

2.1. Dataset

The dataset includes 1704 Judgements issued by the Court of Justice of the European Union between 1954 and 2020. All judgements are in English and freely available from the official portal of the European Union Eur-lex. Content-wise, the judgments concern the free movement of goods and the free movement of persons, both an ideal test bed. On the one hand, the Court has fashioned the fundamental principles of European Union law and developed its central doctrines in those areas, which makes them ideal for the prediction of legal importance (doctrinal outcome). On the other hand, with the development of fundamental principles, it became relevant whether the Court left the key practical decisions to the courts and the legislators of the Member States (deference). Legal experts (human coders) labelling each judgment specified whether the judgment was legally important (Doctrinal Outcome or DOCOUT) and whether the Court deferred the final decision to the national courts or legislators (Deference or DEF).

The data is divided in two datasets: the first contains the full judgements and their relative predicted label. The second contains single paragraphs of the judgments, each with assigned label predicted for their judgement. Compared to similar researches our

dataset is small: Wei et al. [8] sampled from a dataset composed by millions of judgments; Xiao et al. [9], in 2018, used a dataset composed by 2.6 million criminal cases published by the Supreme People's Court of China. Small training data is commonly known to result in poor performance, particularly in Deep Learning[10]. However, it is time expensive to produce hand coded training sets.

2.2. Variables

Doctrinal Outcome relates to the Court's law-making activity in the narrow / legal doctrinal sense. Doctrine is defined as a set of rules and principles, which determine the scope and the content of rights and duties. The coding relies on the opinion of legal experts and lawyers. There are two possible outcomes: weak (=0) and strong (=1). The Court can entrench, strengthen or expand its doctrines, create new concepts or develop principles (DOCOUT=1). By contrast, it can moderate its strong doctrinal positions or restate and apply established doctrines, concepts and principles, without further extending their scope (DOCOUT=0).

Deference indicates whether the Court defers the final decision to the national court or the legislator; that is, whether it gives the national judge leeway as per the final decision/outcome of the case. The following language is indicative of the existence of deference (DEF=1): 'it is for the referring court to decide / establish / determine / examine', 'the national court must provide or decide'. When there are no references to the national courts, the outcome is DEF=0. Both variables present a high class imbalance.

2.3. Models

The article implements models based on two types of neural networks: Convolutional Neural Network (CNN) and Bidirectional Long short-term memory Recurrent Neural Network (LSTM), as they have been shown to be excellent methods for text classification [11]. The models' structure is organised in layers: The first is a Text Vectorisation layer followed by an Embedding layer. The Embedding layer will learn a vectorial space where similar words, or words that appear in similar contexts, are at a closer distance than words that appear in different contexts. The second is a Bidirectional LSTM layer. This layer *reads* the text sequentially and is therefore able to detect sequential dependencies as well as *remember* past information and context. Finally, there is a variable number of Dense layers, using ReLu activation function and a final output layer implementing a Sigmoid activation function. Furthermore, each hidden layer implements dropout as a means of regularization to reduce chances of over-fitting. Dropout is more effective than other standard computationally expensive regularizers [12]. The models based on CNN follow the same structure, with the difference that a CNN substitutes the bidirectional LSTM, and the CNN layer is followed by a max pooling layer. The number of hidden Dense layers and their dropout rate, as well as the size of the embedding, the number of neurons, the learning rate and the weight of every positive example are computed through the ParEGO hyperparameter tuning algorithm [13]. Table 1 reports the values.

Before fitting the model, we converted the text into lowercase and removed punctuation and numbers. We decided not to remove stopwords, as that did not help to improve the performance of our models. We restricted the size of the vocabulary for the word embedding to 30000 words. The datasets were randomly divided in a training set

Table 1. Parameters found by ParEGO for each variable to predict, dataset (whether we use the full judgments or the paragraphs), and network type. Values are rounded to the third most significant digit.

Variable	DOCOUT				DEF			
Dataset	Full Judgments		Paragraphs		Full Judgments		Paragraphs	
Net Type	LSTM	CNN	LSTM	CNN	LSTM	CNN	LSTM	CNN
Emb. dimension	36	136	204	197	30	113	38	183
Num. units	41	42	167	47	26	31	31	49
Learning rate	1.89e-4	3.9e-5	2.5e-5	4.44e-4	1.54e-4	3.9e-5	9e-3	1.47e-3
Layers Num	0	1	3	1	0	1	1	1
Pos. weight	2.775	1.13	5.02	1.21	1.09	1.129	2.662	1.165
Dropout Conv	-	0.133	-	0.046	-	0.202	-	0.208
Dropout Dense	0.079	0.203	0.2	0.226	0.14	0.219	0.0293	0.319
Conv. Kernel	-	11	-	8	-	8	-	6

and a test set with 85% and 15% of the data. The training set was furthermore randomly split into 4 cross validation folds for the model training. Finally, the performance of the models are computed on the test set. We trained our models until convergence, using an early stopping criterion that monitored the F1 score, using the Adam optimization algorithm and Binary Cross Entropy as loss functions. This is a good strategy which prevents over-fitting and saves some computational time.

3. Results and Discussion

The evaluation of the performance of the models uses the F1 score [14] and the ROC-AUC, as the accuracy is inappropriate in presence of high class imbalance [15]. In fact, a dummy classifier – that is a classifier that classifies everything as the majority class – would obtain an extremely high accuracy score because of such imbalance. The F1 and ROC are reported in Table 2 for both predicted variables, the type of model considered, and the type of prediction – i.e. classifying the full judgment or classifying single paragraphs. The comparison between the results on cross validation and the test set suggest no over-fitting of the data. However, in the case of the LSTM on full judgments, for both DOCOUT and DEF, the performances on the test set are higher than on the cross validation. This might indicate some under-fitting. The overall performance of the models is weak due to a number of difficulties: small training dataset, substantial class imbalance, and the fact that predicting doctrinal outcome is likely to require extensive knowledge of complex legal concepts and legal doctrine. The findings are nonetheless important and telling.

First, LSTM appears to perform better than CNN on all tasks. Future research could push the results forward by increasing the size of the training data (which implies time consuming and expensive hand coding of the data). The performance of deep learning models is known to increase with the size of the training data [10]. Alternatively, researchers could explore more complex models: from multiple LSTMs in sequence, to encoder-decoder architectures, and more recent BERT-based models[16].

Second, predicting deference is easier than predicting doctrinal outcome. This is expected, as deference has a lower class imbalance than doctrinal outcome. In fact, for the LSTM on the full judgments, the positive weight for deference selected by ParEGO

Table 2. Performance of our models on classifying full judgments or single paragraphs for doctrinal outcome (DOCOUT) and deference (DEF)

			Cross validation		Test-set	
Label	Dataset	Net Type	F1	ROC	F1	ROC
DOCOUT	Full Judgments	LSTM	0.317	0.682	0.376	0.649
		CNN	0.311	0.663	0.316	0.66
	Paragraphs	LSTM	0.570	0.870	0.569	0.884
		CNN	0.550	0.849	0.539	0.853
DEF	Full Judgments	LSTM	0.372	0.715	0.463	0.639
		CNN	0.380	0.700	0.342	0.677
	Paragraphs	LSTM	0.574	0.801	0.605	0.840
		CNN	0.589	0.837	0.605	0.863

(Table 1) is 1.09, as opposed to 2.775 for the doctrinal outcome. Additionally, from the legal perspective, it is more difficult to identify a strong or weak doctrinal outcome than a deferential outcome. The former is often implicit in the text, and often a matter of scholarly analysis rather than an information contained in the text of the judgment. [17]. The latter relies more on the text and the presence of certain expressions.

Finally, predictions on the paragraph level exhibit higher performance than prediction on the whole judgment. However, the strategy of aggregating paragraph predictions onto full judgment predictions yields performance 25% worse than those obtained through direct classification of full judgments. As such, in order to classify legal texts, we need to cope with long sequences. Besides the already mentioned encoder-decoder architectures, other ideas worth of further investigation are *1)* feeding multiple paragraphs in parallel to the prediction algorithm, thus training a network with multiple inputs, and *2)* summarising/filtering the judgments to retain only the most salient parts of the text.

4. Conclusions

The article investigated how much machine learning could contribute to the legal analysis of judicial decisions by predicting two legally interesting and demanding characteristics: legal importance (doctrinal outcome) and deference. It trained a set of classifiers based on word embedding, LSTM, and CNN on a dataset of manually labelled judgments of the European Court of Justice. The tasks proved difficult and performance were weak, with LSTM performing better than CNN.

Further analysis and experimentation would be required to understand the significance of these results. These include: developing more sophisticated models, incorporating more hand-coded judgements, and finding ways to deal with long text sequences. This work can be viewed as a starting point for studying the impact of text classification and the potential of deep learning models in very specific NLP fields, such as the legal domain. At the same time, the article suggests that the legal experts remain the final authority when it comes to legal doctrine.

References

[1] E. Brynjolfsson and T. Mitchell, "What can machine learning do? workforce implications," *Science*, vol. 358, no. 6370, pp. 1530–1534, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[3] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[4] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.

[5] K. D. Ashley, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, 2017.

[6] I. Habernal, D. Faber, N. Recchia, S. Bretthauer, I. Gurevych, C. Burchard, *et al.*, "Mining legal arguments in court decisions," *arXiv preprint arXiv:2208.06178*, 2022.

[7] F. Schauer, *Thinking like a lawyer: a new introduction to legal reasoning*. Harvard University Press, 2009.

[8] F. Wei, H. Qin, S. Ye, and H. Zhao, "Empirical study of deep learning for text classification in legal document review," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 3317–3320, 2018.

[9] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, and J. Xu, "Cail2018: A large-scale legal dataset for judgment prediction," 2018.

[10] M. Z. Alom, T. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. Nasrin, M. Hasan, B. Essen, A. Awwal, and V. Asari, "A state-of-the-art survey on deep learning theory and architectures," *Electronics*, vol. 8, p. 292, 03 2019.

[11] M. S, J. D, and D. M, "Evaluation of impact of neural networks in text classification," *Journal of University of Shanghai for Science and Technology*, vol. 23, pp. 1279–1292, 07 2021.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[13] J. Knowles, "Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 1, pp. 50–66, 2006.

[14] L. Derczynski, "Complementarity, F-score, and NLP evaluation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, (Portorož, Slovenia), pp. 261–266, European Language Resources Association (ELRA), May 2016.

[15] K. Spackman, ". signal detection theory: Valuable tools for evaluating inductive learning," *Proc. Sixth Internat. Workshop on Machine Learning. Morgan Kaufman, San Mateo, CA, pp. 160–163*, 1989.

[16] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Legal-bert: The muppets straight out of law school," *arXiv preprint arXiv:2010.02559*, 2020.

[17] U. Sadl and Y. Panagis, "What is a leading case in eu law? an empirical analysis," *European Law Review*, vol. 40, pp. 15–34, Feb. 2015.