Explainable Artificial Intelligence, Lawyer's Perspective

Łukasz Górski
lgorski@icm.edu.pl
Interdisciplinary Centre for Mathematical and Computational Modelling
Warsaw, Poland

Shashishekar Ramakrishna
shashi792@gmail.com
Freie Universität Berlin
Berlin, Germany

ABSTRACT

Explainable artificial intelligence (XAI) is a research direction that was already put under scrutiny, in particular in the AI&Law community. Whilst there were notable developments in the area of (general, not necessarily legal) XAI, user experience studies regarding such methods, as well as more general studies pertaining to the concept of explainability among the users are still lagging behind. This paper firstly, assesses the performance of different explainability methods (Grad-CAM, LIME, SHAP), in explaining the predictions for a legal text classification problem; those explanations were then judged by legal professionals according to their accuracy. Secondly, the same respondents were asked to give their opinion on the desired qualities of (explainable) artificial intelligence (AI) legal decision system and to present their general understanding of the term XAI. This part was treated as a pilot study for a more pronounced one regarding the lawyer's position on AI, and XAI in particular.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

explainable artificial intelligence, SHAP, LIME, Grad-CAM, survey, XAI

ACM Reference Format:
Łukasz Górski and Shashishekar Ramakrishna. 2021. Explainable Artificial Intelligence, Lawyer's Perspective. In *Eighteenth International Conference for Artificial Intelligence and Law (ICAIL'21), June 21–25, 2021, São Paulo, Brazil.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462757.3466145

1 INTRODUCTION

Explainable artificial intelligence (XAI) is a *research field that aims to make AI systems results more understandable to humans* [1]. Other domains notwithstanding, law is a domain that needs focus regarding explainability. Due process of law and fair trial requirements make an explanation of AI-based decision models a must. However, the legal and technological discourses pertaining to the creation of explainable AI seem to be separated. In particular, the methods used to create explainable systems seem to favour the needs of AI-engineers. This work was thought as the first step towards the identification of requirements of explainable AI-based systems that would involve legal perspective to a greater extent.

For the purpose of this paper, a two-way investigation was performed. Firstly, to assess the performance of different explainability methods (Grad-CAM, LIME, SHAP), we have used a convolutional neural network (CNN) for text classification and used those methods to explain the predictions; those explanations were then judged by legal professionals according to their accuracy. Secondly, the same respondents were asked to give their opinion on the desired qualities of (explainable) artificial intelligence (AI) legal decision system and to present their general understanding of the term XAI. This part was treated as a pilot study for a more pronounced one regarding the lawyer's position on AI, and XAI in particular. Our results can be treated as a stepping stone towards a more pronounced survey-based research.

This work contributes by:

- Presenting a comparison of different explainability methods when applied to legal classification neural network.
- Giving an assessment of explainable artificial intelligence as understood by lawyers.

2 RELATED WORK

XAI is a research direction that was already put under scrutiny, in particular in the AI&Law community. The general consensus is that the opening of black-box models is a *conditio sine qua non* for assuring their trustworthiness and compliance with normative background [12]. Moreover, in [33] it was noted that the development of explainable technical solutions is necessarily connected with the regulatory frameworks. Importantly, *AI and legal specialists need to engage in a dialogue when tackling explainability questions in legal AI* [33]. However, currently, we are unaware of any study that aimed to investigate the understanding of XAI and elicit system's requirements from the lawyers. Yet, obviously, legal perspective was brought into the discourse with the advent of legislation such as European GDPR or envisioned ePrivacy or Digital Services Act, other areas of law, like tort, notwithstanding [9]. As far as the legal perspective is concerned, in [7] it was noted that the judiciary will play an important role in the production of different forms of XAI, especially because judicial reasoning, *using case-by-case consideration of the facts to produce nuanced decisions, is a pragmatic way to develop rules for XAI* [7]. Whilst this way of thinking sounds very attractive and reminds how seminal works on case-based-reasoning in AI originated in the AI&Law community [24], on the other hand, the cited passage clearly alludes to common law judge mode of thinking and civil law judge may be more concerned with general rules.

ICAIL'21, June 21–25, 2021, São Paulo, Brazil

ACM ISBN 978-1-4503-8526-8/21/06…$15.00
https://doi.org/10.1145/3462757.3466145

This train of thought should also remind of historical developments regarding the explainability in AI&Law. Historically, case-based reasoners and rule-based reasoners (symbolic AI) are inherently interpretable [3]. Yet, producing viable explanations for deep learning algorithms (sub-symbolic AI) remains a challenge. Atkinson *et al.* cite Robbins, who remarked that *[a]ny decision requiring an explanation should not be made by machine learning (ML) algorithms. Automation is still an option; however, this should be restricted to the old-fashioned kind of automation whereby the considerations are hard-coded into the algorithm* [3] [25]. Consequently, historical machine learning-based systems (like SMILE) were capable of achieving good performance, though introspection of their domain models repeatedly shows that those were often incomplete or faulty [3]. Development of twin systems, with traditional rule-based or case-based reasoner explaining the results of neural network was also a subject of scrutiny. Research in this respect continues to test different newly developed methods to achieve good explanations. Usage of Attention Network, accompanied with attention-weight-based text highlighting was explored in [4], with the negative conclusion regarding this approach's feasibility. In general, Competition on Legal Information Extraction and Entailment (COLIEE) is a venue for applying state-of-the-art research regarding the domain-specific information selection and justification generation methods [21][20].

The work [4] is exemplary in including user evaluation study. However, whilst there were notable developments in the area of (general, not necessarily legal) XAI, user experience studies regarding such methods, as well as more general studies pertaining to the concept of explainability among the users are still lagging behind [31] (with, for example, <1% of papers in the area of case-based reasoning including user evaluations in the first place [12] [11]). The results presented herein can be considered as facilitating such dialogue.

On the other hand, surveys are commonly used to assess the usability of computer-aided methods for lawyers. Whilst the study of law is commonly associated rather with logical and linguistic studies of texts, interdisciplinary studies of law do include methods like survey-based study. Works that employed such methodology include [4] (aforementioned attention-weight-based text highlighting) or [5](for building the ontology of basic legal knowledge). As recognized in [5], usability studies are an important part in the human-centered design process, where they are used for, *inter alia*, requirements specification with the use of questionnaire-based methods. Such methodology was employed, for example, in [32], where A/B study was conducted to grade the understandability of student loan decision letter created by an automated decision system. However, such system was rule-based and explainability is mostly connected with deep-learning systems, which remain inherently non-interpretable.

Contemporary XAI models view human-machine interaction as a dynamic process, in which user understanding of the system is continuously impacted by the explanation they receive [12]. User's mental model of the system can further be decomposed into three parts: user model of the domain - which pertains to their understanding of the area that the system is working in, for example, legal decision making; user model of the AI system - pertaining to the knowledge of how the system is implemented, for example using CNN or rule-based classifier; user model of the explanation - knowledge how explanation strategy works with the system [12]. This tripartite division offers a useful conceptual framework through which the results of this paper can be analysed. While our respondents have deep domain knowledge, their knowledge of AI-based systems or explanation techniques is much shallower, if any.

In regard to the concrete explanation methods we've tested (Grad-CAM, LIME, SHAP), they are of different origins. Grad-CAM is a method used for the detection of the most important input data for a CNN's prediction. While it was originally used with image data, it was already proven to be of use in the case of textual input, including legal texts [8]. However, the aforementioned work does not include an end-user feasibility study and can be supplemented with the comparison of Grad-CAM with other explainability methods. They were, to our knowledge, studied in isolation. For example, there are already first studies regarding the usability of LIME for text data [18] (however, with conclusions pertaining to decision trees and linear models). Available analyses for SHAP, on the other hand, used it with CNN-based text classification models [36]. Nevertheless, it should be noted that post-hoc explanation methods were already subject of criticism, due to them representing correlations and not being faithful to the underlying model [30].

3 METHODOLOGY

3.1 Explanations generation

Herein, the effectiveness of different XAI methods when applied to the law is studied. By using a plug-in classification system based on CNN [13][8], we created a number of predictions based on the well-known PTSD dataset [34]. Those were used as a basis for generating explanations for CNN predictions. In general, our pipeline (based on previous work by other authors [29][13] with significant extensions by us) is composed of embedder, classifications CNN, visualizer, XAI module, and metric-based evaluator; we have also developed a simple pre-processing module that uses some industry *de facto* standard text processing libraries for spelling correction, sentence detection, irregular character removal, etc., enhanced with our own implementations which make them better-suited for legal texts (though it was not used in this research). Our embedding module houses a plug-in system to handle different variants of embeddings, in particular BERT and word2vec. The classification module houses simple 1D CNN which facilitates explainability methods. The XAI module integrates the SHAP, Grad-CAM and LIME models (based on library solutions, as developed, respectively by [16][29][22]). It connects the output from the CNN to arrive at explanations based on the model used. For LIME, the following parameters were customized: `kernel_width=20`, `num_features=150`, `num_samples=400`.

For the purpose of this work, as an embedding layer, we use DistilBERT [26] (a small, fast, cheap and light Transformer model trained by distilling BERT base) encoding input sentences into vectors. Huggingface's DistilBERT implementation was found to work relatively well with XAI libraries that we have used; unfortunately, some of them even rely on Huggingface's implementation details, which can be found a limiting factor when plugin architecture is considered.

The CNN used in the pipeline was trained for classification. We use *The Post-Traumatic Stress Disorder* (PTSD) [19] dataset [35], for CNN training (as well as testing). It annotates a set of sentences originating from 50 decisions issued by the Board of Veterans' Appeals ("BVA") of the U.S. Department of Veterans Affairs according to their function in the decision [35] [34] [27]. The classification consists of six elements: *Finding Sentence*, *Evidence Sentence*, *(Evidence-based) Reasoning Sentence*, *Legal-Rule Sentence*, *Citation Sentence*, *Other Sentence*. This was found to be a relatively good candidate for neural network training, as the classes distinguished in this dataset are of relatively similar sizes. The output from each of the three explainer models had different value ranges, including negative values. Each value inside the value range indicated the association of the token to a specific class. For this experiment of understanding the importance of a word in the downstream task of classification, we round-off all negative explainer values to 0 and then normalize the remaining value ranges to fit a range between 0 - 1.

The software stack used for the development of this system was instrumented under Anaconda 4.8.3 (with Python 3.8.3). Tensorflow v. 2.2.0 was used for CNN instrumentation and Grad-CAMs calculations (with the code itself expanding prior implementation available at [29]). DistilBERT implementation and supporting codes were sourced from Huggingface libraries: transformers v. 3.1.0, tokenizers v. 0.8.1rc2, nlp v. 0.4.0. For XAI methods implementations, SHAP v. 0.37 and LIME v. 0.1.1.18 were facilitated. The code used for this paper is available on GitHub[1]. GPU cluster was used for calculations (with 4x Nvidia Tesla V100 32GB GPUs).

3.2 Explainability methods

The system described hereinbefore, for the purpose of this work, was used to generate three types of explanations. Those can be broadly classified into three different types [37]:

(1) Gradient based-diagnostics e.g. Grad-CAM
(2) Shapley values-based diagnostics e.g. SHAP and
(3) Surrogate model-based diagnostics e.g. LIME

In general, gradient-based methods look inside the box to obtain the gradient and feature values for attributing contributions. The other two methods explain a given model from outside of the box by looking at what is fed into the model and what is produced. The more detailed description of concrete XAI methods chosen for this paper is as follows:

3.2.1 Grad-CAM. Grad-CAM is an explainability method originating from computer vision [28]. It is a well established post-hoc explainability technique when CNNs are concerned. Moreover, the Grad-CAM method passed independent sanity checks [2]. Whilst it is mainly connected with the explanations of deep learning networks used with image data, it has already been adapted for other areas of application. In particular, CNN architecture for text classification was described in [13], and there exists at least one implementation which extends this work with Grad-CAM support for explainability [29]. Grad-CAMs were already used in the NLP domain, for (non-legal) document retrieval [6]. With Grad-CAM technique it is possible to produce a class activation map (heatmap) for a given input sentence and predicted class. Each element of the class activation map corresponds to one token and indicates its importance in terms of the score of the particular (usually the predicted) class. The class activation map gives information on how strongly the particular tokens present in the input sentence influence the prediction of the CNN.

3.2.2 SHAP. SHAP [SHapley Additive exPlanations] [17] is an explainability method based on the shapely concept. SHAP values are used for identifying the contribution of each concept/words/features in a given sentence. Shapley value is the average of the marginal contributions across all permutations, i.e. average of all the permutations for each concept/feature to get each entity's contribution. In terms of local interpretability, each observation gets its own set of SHAP values. For our problem here we consider the global interpretability, where Collective SHAP values can show how much each predictor contributes, either positively or negatively, towards influencing predictions of the CNN. SHAP has shown to provide good consistency in attributing importance scores to each feature [15].

3.2.3 LIME. LIME [Local Interpretable Model (Agnostic) Explanations] [10][23] is another explainability method that provides explanations to the predictions of any classifier by learning interpretable or simpler models locally around the prediction. While the Shapley values consider all possible permutations, a united approach to provide global and local consistency, the LIME approach builds sparse linear models around individual predictions in its local vicinity.

3.3 Comparison metrics

The XAI methods described hereinbefore are used to visualize (highlight) the parts of the input texts that are of most importance for the network's predictions. Such generated heatmaps (saliency maps) are in turn presented to a legal professional for (qualitative) grading. On the other hand, heatmaps are not easily susceptible to quantitative analysis and can be supported by more formal metrics in this respect [14][8]. Quantitative analysis and comparison of heatmaps generated by different XAI methods are thus facilitated by recalling the following metrics [8]:

(1) Fraction of elements above relative threshold t ($\mathcal{F}(v,t)$)
(2) Intersection over union with relative thresholds t_1 and t_2 ($\mathcal{I}(v_1,v_2,t_1,t_2)$)

The first metric, $\mathcal{F}(t)$, aims to quantify what portion of input data is taken into account by CNN when making a prediction. It can be defined as $\frac{|\{x \in v | x > t \times max(v)\}|}{len(v)}$.

The second metric, $\mathcal{I}(v_1,v_2,t_1,t_2)$, is used to compare two saliency maps, v_1 and v_2 and can be used to compare whether the same parts of input sentence affect the CNN's prediction. It takes as arguments two heatmaps (v_1 and v_2), binarizes them using relative thresholds (t_1 and t_2) and finally calculates $\frac{|v_1 \cap v_2|}{|v_1 \cup v_2|}$. It quantifies the relative overlap of words considered important for the prediction by each of two models.

In both cases, 0.15 was chosen as a threshold value. This was selected based on suggestions in prior work [28]. In addition to that, for the current study, a value of 0.5 was tested as well, to filter out the lower-ranking words.

[1] https://github.com/lukeg/xai_lawyers_perspective

3.4 Pilot User Study

To study prospective user's opinions on explainability in AI as well as on the particular method employed in this study, a survey was prepared. It consisted of two parts. The generic one contained three questions, pertaining to the general knowledge of XAI by the respondents. Those were as follows:

(1) Have you ever encountered the term "explainable artificial intelligence" (explainable AI, XAI)?
(2) According to you how can "explainable artificial intelligence"/ explainability in decision systems be characterized?
  (a) Explains the decision-making process and why we arrived at this result.
  (b) Conclusions are explained without a need for you to understand the inner-workings of the system.
  (c) It's more useful for a software developer rather than to a lawyer.
  (d) AI-based systems are of little use and their explainability is thus irrelevant
  (e) Explanations given by a computer system usually are not sufficient and need to be supplemented with background legal knowledge.
(3) What are your expectations regarding the justifications given by the artificial intelligence-based automated decision-making systems?

While the list of questions could have been non-exhaustive, the goal of these above questions was only to assess the depth of their knowledge in XAI. Answers to these questions help us to understand respondent's degree of confidence in the domain of XAI to evaluate/score the results generated from different XAI models.

The second part of the study was aimed to elicit the lawyer's assessment of the three XAI methods that can be used to explain CNN's predictions. In total, six correctly classified sentences from the test set were chosen as a basis for the study. Words composing those sentences were highlighted with colors of different intensities, according to their importance in the final prediction. All the words were subject to highlighting, as even stopwords can be of importance in case of legal interpretation. Respondents were asked to grade each visualization on the scale from 0 (worst) to 10 (best). We have decided to use only the correctly classified sentences, as the visualizations for incorrect ones might have looked peculiar for the respondents and could have confused them when compared with the visualizations for the correct ones. Moreover, understanding this part of the study was already not easy for the respondents, with one commenting that he skipped this part as he did not understand what the *colorful words* mean.

In the end, from the classes distinguished in the dataset, three were chosen for the visualization: *citation sentence*, *reasoning sentence (or evidence-based reasoning sentence)* and *legal rule sentence*. The classes were chosen so that each class stands out when compared with the others. Original dataset includes two additional classes, *finding sentence* and *evidence sentence*. However, even its authors admit that it may be difficult to tell apart different sentences when all the classes are compared (e.g. fact-finding ones vs. legal rule ones are easy to conflate). As our respondents do not necessarily have experience with cases under the cognition of Board of Veterans' Appeals, it was decided that using only a subset of sentences' classes was a good compromise.

The following sentences were chosen for grading:

(1) Evidence-based-reasoning sentences
  (a) Further, as discussed below, none of the medical evidence indicates that a psychiatric disorder had its onset during service, and psychiatric disorders are complex matters requiring medical evidence for diagnosis; they are not the kind of disorders that subject to lay observation.
  (b) Given the inconsistencies between the Veteran's reports and the objective evidence of record, the Veteran's credibility is diminished.
(2) Legal Rule Sentence
  (a) Service connection for PTSD requires medical evidence diagnosing the condition in accordance with 38 C.F.R. 4.125(a); a link, established by medical evidence, between current symptoms and an in-service stressor; and credible supporting evidence that the in-service stressor occurred.
  (b) There must be 1) medical evidence diagnosing PTSD; 2) a link, established by medical evidence, between current symptoms of PTSD and an in-service stressor; and 3) credible supporting evidence that the claimed in-service stressor occurred.
  (c) The Federal Circuit has held that 38 U.S.C.A. 105 and 1110 preclude compensation for primary alcohol abuse disabilities and secondary disabilities that result from primary alcohol abuse.
(3) Citation sentence
  (a) See also Mittleider v. West, 11 Vet. App. 181, 182 (1998) (in the absence of medical evidence that does so, VA is precluded from differentiating between symptomatology attributed to a nonservice-connected disability and symptomatology attributed to a service-connected disability).

4 RESULTS

4.1 Neural network training

Firstly, for classification, CNN was trained for 10 epochs, with a batch size of 1000 elements and a learning rate of 0.001. 80% of the PTSD dataset was used for training, with the remaining 20% left out as a test set. This allowed to achieve accuracy of ca. 83% on the test set. PTSD dataset was found to be a good candidate for training, as it is relatively balanced, with classes' size varying from 1941 elements (*Evidence sentence*) to 389 (*Finding sentence*) in case of the training set. Relations between the sizes of classes in the case of the test set were similar. For details on training effectiveness, Table 1 can be consulted.

4.2 Study results. Lawyers' conceptualization of XAI

For our pilot study, we have collected 21 surveys. Respondents represented a variety of legal professions, including university professors, attorneys and junior legal advisors. Results for questions 1 (prior exposure to the term "XAI") and questions 2 (closed questions regarding the respondents' understanding of XAI) are summarised in Table 2, given as a percentage of respective answers. In Table

Table 1: Classification report for CNN trained for the purpose of this paper

Class	Precision	Recall	F1-score	Support
Other Sentence	0.96	0.57	0.71	83
Reasoning Sentence	0.59	0.41	0.48	148
Finding Sentence	0.63	0.51	0.57	101
Legal Rule Sentence	0.76	0.97	0.85	207
Citation Sentence	0.99	1.00	0.99	213
Evidence Sentence	0.88	0.95	0.92	479

Table 2: Respondents' answers to questions 1 and 2

	yes	no	no opinion
Question 1			
	62%	38%	–
Question 2			
a	90%	5%	5%
b	33%	52%	14%
c	14%	76%	10%
d	0%	90%	10%
e	71%	19%	10%

Table 3: Respondents' answers to questions 2, broken down according to their prior exposure to XAI

	Prior exposure			No prior exposure		
	yes	no	no opinion	yes	no	no opinion
a	100%	0%	0%	75%	13%	13%
b	46%	46%	8%	13%	63%	25%
c	23%	69%	8%	0%	88%	13%
d	0%	85%	15%	0%	100%	0%
e	77%	15%	8%	63%	25%	13%

the answers to question 2 are further broken down by the status of prior exposure to XAI.

The majority of respondents already encountered the term "XAI". Elicitation of XAI requirements through closed question 2 shows that almost all lawyers agree that XAI is used to explain the decision-making process and the way we arrived at the result (with one person having "no opinion" and one disagreeing; both persons were unfamiliar with the term XAI, as noted in their answer to q. 1). Therefore we can conclude that - even in case of no prior exposure to this term - lawyers intuitively conceptualize the meaning of this term. Yet, certain divisions can be seen when results for question 2 are analyzed. 52% of respondents agreed that the XAI systems give explanations and the user does not need to understand its inner workings. On the other hand, 33% of respondents concluded otherwise. Therefore, the significant minority pointed out that the user of such system has to possess a deeper system's model. Despite this division, almost everyone agreed that XAI in legal decision-making systems would be of more importance to the lawyer than to computer scientists (76%); three persons disagreed, with only one of them seeing the need to understand the system's inner workings as a precondition for understanding of its explanation (as evidenced by the previous question).

Artificial intelligence & law curriculum seems to have penetrated the ranks of lawyers, as 90% of respondents see the prospects of their use (and, by extension of XAI). Remaining 10% had no opinion. 71% of respondents see the need of supplementing machine-based explanations with further background legal knowledge, 19% think otherwise.

In conclusion, lawyers seem to be affirmative regarding the usefulness of AI and XAI. Yet, more work should be devoted to prepare the explanations that do not involve prior knowledge of how a given system works. If XAI was to be deployed and aimed at non-professionals, clarity and completeness of the explanations should be the focal point, so that the decisions could be understood even without deeper background legal knowledge.

When we asked each respondent on their list of expectations regarding the justifications given by an AI-based decision-system (question 3), the majority of their expectations are in line with any general requirements necessary for such a system. Users' expectations can be divided into three broad categories. These categories are also in-line with the user's mental model of a AI system (cf. Sec. 2). General requirements, coming from users' deep knowledge of the law and its principles, like due process, contain expectations such as: *fairness, lack of bias, transparency in the decision process, consequence awareness.* Secondly, users' model of AI-based system tells them how it can increase the effectiveness of their work. Here, respondents remarked that such system should *decrease their manual effort, lessen time consumption* and *introduce automation.* Finally, the user's knowledge of XAI implementation is concerned with *supporting system's conclusions with citations* and *provision of facts relevant to system's decision.* It should be noted that those requirements apply not just to XAI systems, but also to any legal software system in general, AI-based or not. Thus, the mental model of an AI-based system in the case of prospective users is not fully developed. Fig. 1, depicts the key vocabulary set used by the respondents while providing their list of expectations.

Figure 1: User expectations regarding the justifications given by an AI-based decision-system

While a majority of the requirements described above were repeated by multiple respondents, a few respondents provided requirements that were *interesting* from an AI-driven automated legal

decision support system's perspective. They are interesting not because of their novelty, but because only a few domain experts/end users thought about it. Some such requirements are as shown below:

(1) *"It should have the option of forgetting the decisions taken in past. However, it should also have a certain standard form of code of conduct."*
(2) *"There should be scope for generating contrasting explanations based on various objectives/functions/factors."*
(3) *"Basis on which AI has neglected the other information."*
(4) *"Legally correct explanations."*
(5) *"Option for collaborative (Man-Machine) explanation of the law, context, and linguistic."*
(6) *"Understandable explanations and Better Interfaces."*

4.3 Study results. XAI methods' assessment

Tables 4, 5, 6 can be consulted for comparions of different XAI methods when used with our sample sentences in terms of user score as well comparison metrics (please note that the matrix for $\mathcal{I}$ metric results is symmetric, therefore repeating results were removed from the table). Below are some observation made based on these results:

- Users' expectations seem to have differed to a large extent, which can be seen from the divisions of respondent's grades of SHAP's explanation of citation sentence (Table 6 can be consulted for visualization). SHAP marked all the words as very important, while - for example, Grad-CAM marked only the reference to the source material, and omitted the summary included with the citation. For the authors of this paper, Grad-CAM has proven superior here, yet many respondents judged SHAP highly. 9 respondents gave it the lowest score out of all three (with two giving 0), 8 graded it the highest (with one person giving 10).
- While the heatmaps for the citation sentence from Grad-CAM and SHAP seem to be very different in nature (also pointed out by their F- metric), the respondents seem to provide scores that vary only by a small margin.
- Based on the $\mathcal{F}$ and $\mathcal{I}$ metric scores for all the sentence types, LIME in general seems to have higher scores compared to Grad-CAM and then SHAP.
- Both SHAP and LIME seem to be more sensitive to the change in threshold value in terms of their $\mathcal{F}$ metric when compared to Grad-CAM.
- Based on respondents' average scores, Grad-CAM and SHAP seem to perform consistently between sentence types when compared against LIME. But, based on $\mathcal{F}$ metric, both LIME and SHAP seem to have consistent values between sentence types as compared to Grad-CAM.

Herein we have presented a software system, based on CNN, capable of classifying legally relevant texts and coupled the system with an explainability module. Furthermore, this explainability module was then subject to user study. The use of ANOVA (because D'Agostino's and Pearson's test did not allow to reject the hypothesis that scores for each of the XAI methods have normal distributions and the same conclusions were arrived at when using Levene test for the equality of variances) confirms that the discrepancy between various scores is not of practical importance. The same results were achieved using the Friedman test ($\alpha = 0.05$ in all the cases). Therefore other factors should be taken into account when choosing a particular method. From a software engineering point of view, it should be noted that implementation of certain methods is dependant not only on the user's voice but also on technical feasibility. In this respect, it is of importance that Grad-CAM is a method that is used together only with CNNs and we are unaware of any works that managed to use it with other neural network architectures. LIME is dependent on a number of hyperparameters. In this work, we have chosen the largest values that allowed us to carry calculations with our hardware resources and software implementation. Yet, many works suggest that one should fine-tune those values until explanations received are close to one's expectations. However, this introduces a risk of overfitting explanatory procedures and having such fine-tuned explanations would be counterproductive with regard to our study. As far as SHAP goes, the library solution we used is still under active development, with documentation not always up-to-date and with limited support of external libraries. Regarding some of the observations presented hereinbefore, there are a few for which different evaluation techniques (technical, empirical, or mixed) need to be performed for arriving at conclusions. Such conclusions need to be backed by domain experts. Those will form a basis for future work.

5 CONCLUSION

In this paper, we have presented a comparison of different explainability methods when applied to legal classification neural network (CNN) and provided an assessment of explainable artificial intelligence as understood by lawyers. Different XAI methods were graded similarly by the users, though certain variances can be spotted. The metrics presented herein offer software engineers an option to quantify explanations given by the system. When a more general point of view is taken, one which comes from prospective users, it should be noted that the lawyers are generally looking forward to the implementation of (explainable) artificial intelligence systems and solutions that allow them to be more efficient in their use of time. Users are thus waiting for results of AI - and in particular XAI - research.

ACKNOWLEDGMENTS

This research was carried out with the support of the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw, under grant no GR81-14.

DISCLAIMER

The views reflected in this article are the views of the author (SR) and do not necessarily reflect the views of his employer organisation or its member firms.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.
[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle,

Table 4: Comparison of XAI methods on sample PTSD evidence-based-reasoning sentences

Method	Sentence	User score (avg.)	$\mathcal{F}$ metric	$\mathcal{I}$ metric vs. SHAP	$\mathcal{I}$ metric vs. LIME	$\mathcal{F}$ metric	$\mathcal{I}$ metric vs. SHAP	$\mathcal{I}$ metric vs. LIME
			$t = t_1 = t_2 = 0.15$			$t = t_1 = t_2 = 0.5$		
Grad-CAM	Further, as discussed below, none of the medical evidence indicates that a psychiatric disorder had its onset during service, and psychiatric disorders are complex matters requiring medical evidence for diagnosis; they are not the kind of disorders that subject to lay observation.	4.43	0.51	0.46	0.48	0.14	0.18	0.3
SHAP	Further, as discussed below, none of the medical evidence indicates that a psychiatric disorder had its onset during service, and psychiatric disorders are complex matters requiring medical evidence for diagnosis; they are not the kind of disorders that subject to lay observation.	3.48	0.82		0.74	0.13		0.25
LIME	Further, as discussed below, none of the medical evidence indicates that a psychiatric disorder had its onset during service, and psychiatric disorders are complex matters requiring medical evidence for diagnosis; they are not the kind of disorders that subject to lay observation.	6.66	0.85			0.45		
Grad-CAM	Given the inconsistencies between the Veteran's reports and the objective evidence of record, the Veteran's credibility is diminished.	6.15	0.92	0.89	0.81	0.78	0.22	0.19
SHAP	Given the inconsistencies between the Veteran's reports and the objective evidence of record, the Veteran's credibility is diminished.	5.15	0.96		0.85	0.26		0.31
LIME	Given the inconsistencies between the Veteran's reports and the objective evidence of record the Veteran's credibility is diminished.	5.8	0.85			0.4		

K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., 9505–9515. https://proceedings.neurips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf

[3] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. 2020. Explanation in AI and law: Past, present and future. *Artificial Intelligence* (2020), 103387.

[4] L Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2020. Scalable and explainable legal prediction. *Artificial Intelligence and Law* (2020), 1–26.

[5] Núria Casellas. 2011. *Legal ontology engineering: Methodologies, modelling trends, and the ontology of professional judicial knowledge*. Vol. 3. Springer Science & Business Media.

[6] Jaekeol Choi, Jungin Choi, and Wonjong Rhee. 2020. Interpreting Neural Ranking Models using Grad-CAM. *arXiv preprint arXiv:2005.05768* (2020).

[7] Ashley Deeks. 2019. The judicial demand for explainable artificial intelligence. *Columbia Law Review* 119, 7 (2019), 1829–1850.

[8] Lukasz Gorski, Shashishekar Ramakrishna, and Jedrzej M Nowosielski. 2020. Towards Grad-CAM Based Explainability in a Legal Text Processing Pipeline. *arXiv preprint arXiv:2012.09603* (2020).

[9] Philipp Hacker, Ralf Krestel, Stefan Grundmann, and Felix Naumann. 2020. Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law* (2020), 1–25.

[10] Linwei Hu, Jie Chen, Vijayan N. Nair, and Agus Sudjianto. 2020. Surrogate Locally-Interpretable Models with Supervised Machine Learning Algorithms. arXiv:2007.14528 [stat.ML]

[11] Mark T Keane and Eoin M Kenny. 2019. How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems. In *International Conference on Case-Based Reasoning*. Springer, 155–171.

[12] Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence* 294 (2021), 103459. https://doi.org/10.1016/j.artint.2021.103459

[13] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[14] David Krakov and Dror G Feitelson. 2013. Comparing performance heatmaps. In *Workshop on Job Scheduling Strategies for Parallel Processing*. Springer, 42–61.

[15] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2019. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv:1802.03888 [cs.LG]

[16] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[17] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

Table 5: Comparison of XAI methods on sample PTSD legal rule sentences

Method	Sentence	User score (avg.)	$\mathcal{F}$ metric	$\mathcal{I}$ metric vs. SHAP	$\mathcal{I}$ metric vs. LIME	$\mathcal{F}$ metric	$\mathcal{I}$ metric vs. SHAP	$\mathcal{I}$ metric vs. LIME
			$t = t_1 = t_2 = 0.15$			$t = t_1 = t_2 = 0.5$		
Grad-CAM	Service connection for PTSD requires medical evidence diagnosing the condition in accordance with 38 C.F.R. 4.125 (a); a link, established by medical evidence, between current symptoms and an in-service stressor; and credible supporting evidence that the in-service stressor occurred.	5.2	0.8	0.89	0.77	0.42	0.41	0.34
SHAP	Service connection for PTSD requires medical evidence diagnosing the condition in accordance with 38 C.F.R. 4.125 (a); a link, established by medical evidence, between current symptoms and an in-service stressor; and credible supporting evidence that the in-service stressor occurred.	6.05	0.9		0.87	0.52		0.41
LIME	Service connection for PTSD requires medical evidence diagnosing the condition in accordance with 38 C.F.R. 4.125 (a); a link established by medical evidence between current symptoms and an in service stressor; and credible supporting evidence that the in service stressor occurred.	5.9	0.93			0.48		
Grad-CAM	There must be 1) medical evidence diagnosing PTSD; 2) a link, established by medical evidence, between current symptoms of PTSD and an in-service stressor; and 3) credible supporting evidence that the claimed in-service stressor occurred.	5.45	0.73	0.47	0.8	0.12	0.14	0.18
SHAP	There must be 1) medical evidence diagnosing PTSD 2) a link, established by medical evidence, between current symptoms of PTSD and an in-service stressor; and 3) credible supporting evidence that the claimed in-service stressor occurred.	5.75	0.67		0.56	0.36		0.08
LIME	There must be 1) medical evidence diagnosing PTSD; 2) a link, established by medical evidence, between current symptoms of PTSD and an in-service stressor; and 3) credible supporting evidence that the claimed in-service stressor occurred.	4.3	0.73			0.11		
Grad-CAM	The Federal Circuit has held that 38 U.S.C.A. 105 and 1110 preclude compensation for primary alcohol abuse disabilities and secondary disabilities that result from primary alcohol abuse.	5.9	0.71	0.41	0.71	0.4	0.13	0.06
SHAP	The Federal Circuit has held that 38 U.S.C.A. 105 and 1110 preclude compensation for primary alcohol abuse disabilities and secondary disabilities that result from primary alcohol abuse.	4.2	0.63		0.63	0.29		0.25
LIME	The Federal Circuit has held that 38 U.S.C.A. 105 and 1110 preclude compensation for primary alcohol abuse disabilities and secondary disabilities that result from primary alcohol abuse.	6.15	0.93			0.54		

[18] Dina Mardaoui and Damien Garreau. 2020. An Analysis of LIME for Text Data. *arXiv preprint arXiv:2010.12487* (2020).

[19] Victoria Hadfield Moshiashwili. 2015. The Downfall of Auer Deference: Veterans Law at the Federal Circuit in 2014.

[20] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. [n.d.]. COLIEE 2020: Methods for Legal Document Retrieval and Entailment. ([n. d.]).

[21] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2019. A Summary of the COLIEE 2019 Competition. In *JSAI International Symposium on Artificial Intelligence*. Springer, 34–49.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 1135–1144.

Table 6: Comparison of XAI methods on sample PTSD citation sentence

Method	Sentence	User score (avg.)	$\mathcal{F}$ metric	$\mathcal{I}$ metric vs. SHAP	$\mathcal{I}$ metric vs. LIME	$\mathcal{F}$ metric	$\mathcal{I}$ metric vs. SHAP	$\mathcal{I}$ metric vs. LIME
			$t = t_1 = t_2 = 0.15$			$t = t_1 = t_2 = 0.5$		
Grad-CAM	See also Mittleider v. West, 11 Vet. App. 181, 182 (1998) (in the absence of medical evidence that does so, VA is precluded from differentiating between symptomatology attributed to a nonservice-connected disability and symptomatology attributed to a service-connected disability).	5.2	0.26	0.26	0.24	0.13	0.13	0.11
SHAP	See also Mittleider v. West, 11 Vet. App. 181, 182 (1998) (in the absence of medical evidence that does so, VA is precluded from differentiating between symptomatology attributed to a nonservice-connected disability and symptomatology attributed to a service-connected disability).	4.95	0.98		0.83	0.98		0.16
LIME	See also Mittleider v. West, 11 Vet. App. 181, 182 (1998) (in the absence of medical evidence that does so, VA is precluded from differentiating between symptomatology attributed to a nonservice-connected disability and symptomatology attributed to a service-connected disability).	4.65	0.73			0.17		

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG]

[24] Edwina L Rissland, Kevin D Ashley, and Ronald Prescott Loui. 2003. AI and Law: A fruitful synergy. *Artificial Intelligence* 150, 1-2 (2003), 1–15.

[25] Scott Robbins. 2019. A misdirected principle with a catch: explicability for AI. *Minds and Machines* 29, 4 (2019), 495–514.

[26] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL]

[27] Jaromír Savelka, Vern R. Walker, Matthias Grabmair, and Kevin D. Ashley. 2017. Sentence Boundary Detection in Adjudicatory Decisions in the United States.

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[29] Haebin Shin. [n.d.]. Grad-CAM for Text. https://github.com/HaebinShin/grad-cam-text. Accessed: 2020-08-05.

[30] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[31] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[32] Tom M van Engers and Dennis M de Vries. 2019. Governmental Transparency in the Era of Artificial Intelligence.. In *JURIX*. 33–42.

[33] Martijn van Otterlo and Martin Atzmueller. 2018. On Requirements and Design Criteria for Explainability in Legal AI. In *XAILA@ JURIX*.

[34] Vern R. Walker, Ji Hae Han, Xiang Ni, and Kaneyasu Yoseda. 2017. Semantic Types for Computational Legal Reasoning: Propositional Connectives and Sentence Roles in the Veterans' Claims Dataset *(ICAIL '17)*. Association for Computing Machinery, New York, NY, USA, 217–226. https://doi.org/10.1145/3086512.3086535

[35] Vern R. Walker, Krishnan Pillaipakkamnatt, Alexandra M. Davidson, Marysa Linares, and Domenick J. Pesce. 2019. Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning. In *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts, Montreal, QC, Canada, June 21, 2019 (CEUR Workshop Proceedings, Vol. 2385)*. http://ceur-ws.org/Vol-2385/paper1.pdf

[36] Wei Zhao, Tarun Joshi, Vijayan N Nair, and Agus Sudjianto. 2020. SHAP values for Explaining CNN-based Text Classification Models. *arXiv preprint arXiv:2008.11825* (2020).

[37] Wei Zhao, Tarun Joshi, Vijayan N. Nair, and Agus Sudjianto. 2020. SHAP values for Explaining CNN-based Text Classification Models. arXiv:2008.11825 [cs.CL]