Explainable NLLP: Advancements in Explainable AI for Natural Legal Language Processing

Lucas Resck[1,*], Felipe Moreno-Vera[1], Tobias Veiga[1], Gerardo Paucar[1], Ezequiel Fajreldines[2], Guilherme Klafke[2], Luis Gustavo Nonato[3] and Jorge Poco[1]

[1]Getulio Vargas Foundation, Rio de Janeiro, Brazil
[2]Getulio Vargas Foundation, São Paulo, Brazil
[3]University of São Paulo, São Carlos, Brazil

Abstract

Despite the increasing application of machine learning and NLP methods in the legal domain, there has been limited effort to enhance the understanding and transparency of these algorithms. This paper addresses this gap by presenting a survey on Explainable AI (XAI) applied to **Natural Legal Language Processing (NLLP)**. To our knowledge, this survey represents the first comprehensive examination at the intersection of XAI, Law, and NLP. Building upon prior surveys focused on partial intersections of these domains, we propose a taxonomy for classifying papers based on the NLLP task, explanation type, and technique employed. Additionally, we delve into discussions surrounding Explainable NLLP, considering perspectives related to ethics, current open issues, and future work. Our analysis reveals that the categorized papers generally do not thoroughly examine the ethical implications of the explainability principle in NLP within the legal field. Furthermore, they do not discuss the role and value of explanations nor do they effectively utilize their respective XAI techniques to offer insights into the limitations of NLP systems.

Keywords

Explainable AI, Natural Language Processing, Law, Survey

1. Introduction

Natural Language Processing (NLP) falls under the umbrella of Artificial Intelligence (AI). It is dedicated to facilitating interaction between computers and human language. It aims to empower machines to comprehend, interpret, and generate human language in a manner that is not only meaningful but also contextually relevant. While there has been substantial growth in the application of Machine Learning (ML) and NLP methods within the legal domain, often referred to as "LegalAI" [1, 2, 3], relatively little has been done to enhance the comprehension and transparency of algorithms such as legal document summarization [4, 5], legal document classification [6, 7], and predictive analytics for legal outcomes [8, 9, 10, 11].

With the growing power and complexity of ML and NLP algorithms, the demand for transparency in these systems has never been more critical. Transparency in ML applications entails the capacity to comprehend, interpret, and expound upon the decisions and predictions made by these algorithms, a vital aspect within the legal domain. Within this context, Transparency in ML and Explainable Artificial Intelligence (XAI) are closely intertwined concepts, both striving to render AI and ML systems more understandable, interpretable, and accountable. Together, they tackle ethical, regulatory, and user trust concerns in AI and facilitate the widespread integration of AI technologies across various fields, particularly in Natural Legal Language Processing (NLLP). Within legal NLP, the fusion of ML transparency and XAI is indispensable for upholding fairness, compliance, and trustworthiness. This approach benefits legal professionals, stakeholders, and the public by providing insights into AI-driven legal decisions and enabling AI's responsible and ethical use within the legal domain.

Nonetheless, while a few works encompass the study, review, and synthesis of XAI & NLP [12, 13] or NLP & Law [1, 3], none of these delve into pertinent subjects or trends in XAI, Legal NLP, or the intersection of both, such as trustworthiness, fairness, and ethics. Hence, we recognize the significance of investigating Legal, NLP, and XAI. This intersection is paramount because the legal domain imposes specific constraints and requisites concerning explanations and justifications. In this vein, we advocate for thoroughly exploring the prevailing trends in techniques and explanations applied in NLLP. To address this, we present this survey, focusing on covering and addressing these topics and structuring them through developing a taxonomy rooted in XAI and NLLP. In this work, we treat explainability and interpretability interchangeably, as it is common in the literature [14, 15], despite existing debate [16].

Main Contributions. This work encompasses a survey focusing on applying Explainable AI in Natural Legal Language Processing. Additionally, we identify papers that explicitly address ethics, particularly within the context of XAI. Our primary contributions in this study are outlined as follows:

- We introduce a taxonomy for systematically categorizing papers based on the NLLP task and the specific type and explanation technique employed.
- We analyze the prevailing research trends in types of explanation and the utilization of XAI.
- We consider perspectives related to ethics in XAI and the legal domain, emphasizing the necessity of

Proceedings of the Seventh International Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2025), 16 June, 2025, Chicago, USA

*Corresponding author.
lucas.domingues@fgv.edu.br (L. Resck); felipe.moreno@fgv.br (F. Moreno-Vera); tobsv21@gmail.com (T. Veiga); carlos.malqui@fgv.br (G. Paucar); ezequiel.santos@fgv.br (E. Fajreldines); guilherme.klafke@fgv.br (G. Klafke); gnonato@icmc.usp.br (L. G. Nonato); jorge.poco@fgv.br (J. Poco)
https://lucasresck.github.io/ (L. Resck); https://www.fmorenovr.com/ (F. Moreno-Vera); https://direitosp.fgv.br/en/professor/guilherme-forma-klafke (G. Klafke); https://www.icmc.usp.br/~gnonato/ (L. G. Nonato); http://visualdslab.com/~jpocom/ (J. Poco)
0000-0001-9634-450X (L. Resck); 0000-0002-2477-9624 (F. Moreno-Vera); 0000-0001-6440-5638 (E. Fajreldines); 0000-0002-1088-794X (G. Klafke); 0000-0002-8514-8033 (L. G. Nonato); 0000-0001-9096-6287 (J. Poco)

addressing ethical concerns and pointing out current challenges.
- We comprehensively discuss the existing open issues within the realm of XAI applied to NLLP.

2. Previous Surveys

Researchers have made significant progress in summarizing and classifying the forefront of XAI, yielding an extensive body of literature that addresses this challenge from various perspectives and within diverse domains [17, 1, 12, 13, 3, 18]. In a recent paper, Schwalbe and Finzel [17] have consolidated all these prior efforts into a unified taxonomy.

This survey classifies advances in XAI within the specific Natural Legal Language Processing domain. To our knowledge, this represents the inaugural survey at the crossroads of **XAI**, **Law**, and **NLP**. We extend upon pertinent prior works that have approached the convergence of NLP and Law [1, 3] and NLP and XAI [12, 13] to construct a comprehensive taxonomy encompassing XAI, Law, and NLP.

Atkinson et al. [18] analyze explanation methodologies in AI as applied to law. However, their focus is predominantly on conventional automation-based systems, such as rule- and case-based ones. While they touch upon explainability in machine learning, it is done with a critical perspective. We delve into these critiques and their implications in Section 5.

Danilevsky et al. [13] present a survey on applying XAI in NLP. This work is complemented by an interactive browser-based system for exploring the study [12]. This body of work organizes explanations and encompasses various modalities through which explanations are extracted and visualized. Drawing inspiration from Danilevsky et al. [13] efforts, we propose our own taxonomy, particularly concerning the categorization of explanations (local vs. global, self-explanatory vs. post-hoc) and methods of explainability (e.g., feature importance, surrogate, among others).

Additionally, some works scrutinize the intersection of NLP and the legal domain, a field referred to as *LegalAI* by Zhong et al. [3]. Specifically, this research categorizes and illustrates several methods based on embeddings and symbols. It also delineates several applications of LegalAI. Finally, Katz et al. [1] provide a comprehensive overview of the current state of legal NLP. Despite their extensive analysis of hundreds of related papers, they also propose a broad taxonomy centered around engineering tasks in NLP.

3. Methodology

3.1. Search for Papers

To curate pertinent literature, we conducted a thorough search using Google Scholar and Semantic Scholar. Employing keywords such as "xai legal nlp," "legal nlp," "legal decision prediction," "nlp legal judgment prediction xai," and "nlp legal judgment prediction interpretable," without date range restrictions, we executed the search. The gathered papers underwent screening based on their title, abstracts, keywords, and full text. Ensuring a focused selection of literature, papers were screened based on relevance regarding the intersection of XAI, NLP, and Law. For instance, we included any papers that employed explainability or interpretability techniques to enhance the understanding of NLP models in legal contexts.

Additionally, we scrutinized the bibliographies of each selected paper from the initial search, incorporating those identified as pertinent into our list for meticulous examination. Following a comprehensive review of the selected papers, we arrived at a set of 40 documents, which are thoroughly discussed and organized within the proposed taxonomy (see Section 3.2). The resulting survey considers works from a diversity of venues, including the Association for Computational Linguistics Anthology (ACL, AACL, COLING, EMNLP, NAACL, etc.), AI & Law venues (Artificial Intelligence and Law and ICAIL), and preprint repositories (arXiv and SSRN).

3.2. Taxonomy

To propose a taxonomy for the convergence of NLP, XAI, & Law, we have built upon prior efforts in categorizing papers within the realms of NLP & Law and NLP & XAI (Section 2).

Explanation Type: We adhere to the approach outlined by Danilevsky et al. [13] and Qian et al. [12], organizing explanation methods into the following classifications:

- Local vs. Global: This pertains to whether the explanation is specific to a particular instance or provides an overview of the model's behavior across the entire set of instances.
- Self-explaining vs. Post-hoc: This distinguishes whether the explanation is derived directly from the model or obtained through a post-processing.

It is worth noting that only a limited number of works rely on global explanations (as shown in Table 1). Consequently, while global explanations constitute a pertinent category, our ensuing discussion will primarily center on local explanation methods.

Explainability Technique: Diverse approaches exist for integrating XAI methods into a legal NLP pipeline, encapsulated by the various explainability methods employed. We also draw upon Danilevsky et al. [13]'s work for classification:

- Feature importance: This XAI method scrutinizes and assigns importance scores to the features utilized in the prediction process, such as employing attention mechanisms [19].
- Surrogate model: In this approach, another model, typically simpler and interpretable, approximates the decision-making process of the original model and serves as a stand-in for explanations, as exemplified by LIME [20].
- Example-driven: Other examples are used to justify the prediction.
- Provenance-based: This method is employed when the decision-making process involves a sequence of derivation steps, some or all presented as part of the explanation.
- Declarative induction: Human-readable representations like trees [21] serve as explanations in this category.

It is essential to note that these categories are not mutually exclusive. For instance, the LIME technique falls under both feature importance and surrogate model. When this happens, we apply the most pertinent one.

Table 1
Works categorized by the suggested taxonomy of Section 3.2: explanation type, explainability technique, and NLP task. Works with mentions of ethics are in *italic*, and works with mentions to ethics in XAI are in **bold**.

Explanation Type	Explainability Technique	NLP Task	Representative Works
Local self-exp.	Provenance-based	Classification	Zhao et al. [22], Bhambhoria et al. [23], Li et al. [24], Lyu et al. [25], **Wu et al. [26]**, Branting et al. [19], **Zhong et al. [27]**, Branting et al. [28], *Chen et al. [29]*, Jiang et al. [30], Ashley and Brüninghaus [31]
	Feature importance	Classification, Machine summarization, Resources, Text generation	Bertalan and Ruiz [32], **Nielsen et al. [33]**, Wang et al. [34], *Zhou et al. [2]*, Branting et al. [19], Chalkidis et al. [35], *Malik et al. [36]*, Norkute et al. [37], Branting et al. [28], Caled et al. [38], Ye et al. [39]
	Declarative induction	Classification, Resources, Text generation	Mumford et al. [40], Ye et al. [39]
	Surrogate model	Classification, Information retrieval, Resources	Resck et al. [41]
Local post-hoc	Surrogate model	Classification, Information retrieval, Resources	**Benedetto et al. [42]**, Resck et al. [41], Bhambhoria et al. [43], Domingues [44], Górski and Ramakrishna [45], Rabelo et al. [46], Rabelo et al. [47], Chhatwal et al. [48]
	Feature importance	Classification, Information retrieval, Machine summarization, Resources	**Valvoda and Cotterell [49]**, *Semo et al. [50]*, *T.y.s.s et al. [51]*, Górski and Ramakrishna [45], *Malik et al. [36]*, Norkute et al. [37], Mahoney et al. [52], Waltl et al. [53], Górski et al. [54], Landthaler et al. [55]
	Declarative induction	Classification	de Arriba-Pérez et al. [56]
	Example-driven	Classification	**Valvoda and Cotterell [49]**
Global self-exp.	Feature importance	Classification	Medvedeva et al. [57], Strickson and De La Iglesia [8], Aletras et al. [11]
	Declarative induction	Classification	González-González et al. [21], de Arriba-Pérez et al. [56]

NLP Task: Research at the intersection of NLP and Law leverages NLP techniques to address legal challenges. Hence, it is crucial to classify these studies based on their specific NLP tasks. We employ the comprehensive and succinct taxonomy proposed by Katz et al. [1, Table 1] for this purpose. However, our analysis revealed that most of the works fall within the "Classification" category, encompassing Outcome Prediction, Legal Area Classification, and Topic Modeling. It is worth noting that this prevalence is not arbitrary. While Katz et al. [1] presents a broad spectrum of legal NLP tasks, those beyond classification, machine summarization, and text generation, such as “resources,” tend to be less reliant on machine learning, if not entirely independent. Consequently, they pose challenges when applying XAI methods. Conversely, machine summarization and text generation are comparatively less common [1].

Nonetheless, certain studies (e.g., 41) are labeled with additional categories beyond classification, such as information retrieval and resources. A few others are labeled independently of classification, such as machine summarization [33] and text generation [39].

Ethical Issues: The ethical implications of applying machine learning to NLP are paramount, particularly concerning the choice of explainability methods. We identify studies that address these ethical concerns, emphasizing those that do so within the context of XAI.

4. Taxonomy Discussion

This section offers an overview of the primary XAI techniques employed in each respective XAI type. It is worth noting that the global post-hoc XAI type is omitted due to its absence in the reviewed literature. Furthermore, we present noteworthy observations applicable to all the examined studies. Table 1 thoroughly categorizes the works based on the XAI type, explainability method, and NLLP task.

4.1. Local Post-hoc

This XAI type encompasses notable XAI techniques such as LIME and input gradient methods, including Integrated Gradients [58] and Grad-CAM [59]. Recent studies employing LIME include those conducted by Resck et al. [41] and Bhambhoria et al. [43]. Similarly, Benedetto et al. [42] and Semo et al. [50] have undertaken investigations utilizing input gradient techniques. In several studies within this XAI type, researchers have employed a combination

of or explored at least two different approaches (types or techniques) to provide explanations, e.g., Górski and Ramakrishna [45] and Norkute et al. [37]. Noteworthy is the work by Benedetto et al. [42], which distinguishes itself by offering explanations at the sentence level and by conducting comparisons against ground truth. Conversely, other works primarily generate explanations at the word level.Information retrieval frameworks, e.g., text similarity, are employed by Resck et al. [41] and Landthaler et al. [55] — the retrieval is explained with additional text similarity and LIME, respectively. In machine summarization, Norkute et al. [37] also explore whether adding textual similarity highlights as an explanation can help users evaluate the summarization of legal documents.

In one particularly interesting approach, given a primary model tasked with a main classification, a secondary model autonomously computes additional pertinent classes or text segments capable of elucidating the prediction. Representative works have emanated from the competitions COLIEE 2019 [47] and 2020 [46]. Due to the independent nature of the secondary model from the main model's predictions, the former can generate predictions in advance of the latter. Consequently, the nomenclature *post-hoc* may not entirely encapsulate the essence of this technique.

4.2. Local Self-explaining

This XAI type primarily employs attention weights of deep learning architectures, e.g., Transformer- [22] and LSTM-based [38] models, as its main approach — an exception is the work by Zhou et al. [2], which employs classifier weights, commonly used by global explanations but aiming individual samples. Typically, local self-explaining methods, except for provenance-based, emphasize the word level [32, 41, 33, 38]. However, in the study by Zhao et al. [22], attention scores are sometimes extended to encompass entire sentences, thereby providing an alternative explanation at the textual level. For instance, a whole factual statement may be deemed significant at the sentence level. In contrast, mentioning a concept may hold importance at the word level.

In the context of legal summarization, Nielsen et al. [33] and Norkute et al. [37] have explored the use of attention highlights as explanations. When evaluating legal document summarization, attention highlights improved completion time, trust, and preference [37]; the use of attention highlights did not affect the temporal allocation of user attention, but spatiotemporal allocation has evidence of being affected [33]. Similarly, Ye et al. [39] explored attention scores to interpret the text generation of court views, which are analogous to natural language explanations for charge predictions.

In a different approach, secondary models predict other relevant and interpretable labels, which subsequently serve as features for the primary model responsible for the main prediction task. This approach encompasses a substantial body of work [25, 26, 23, 27, 31] and is a subset of provenance-based methods.

4.3. Global Self-explaining

While less commonly observed in the reviewed literature, this XAI type offers valuable insights. Aletras et al. [11] and Medvedeva et al. [57] employ the feature importance of an SVM model, achieved through an analysis of the SVM kernel weights. Similarly, de Arriba-Pérez et al. [56] and González-González et al. [21] leverage declarative induction within a Random Forest model. This entails identifying, for any given class, all the tree paths from root to leaf that contribute to the score of the respective class. Both methodologies apply the models to text that has undergone preprocessing using TF-IDF. Additionally, Strickson and De La Iglesia [8] directly analyze the most important TF-IDF features. The inherent simplicity of these techniques plays a crucial role in generating thoroughly explainable models. Remarkably, they consistently demonstrate commendable performance despite the anticipated trade-off between interpretability and performance [60, 61, 62].

5. Ethics Discussion

In this work, we identify papers that explicitly address ethical concerns, particularly within the context of XAI (Section 3.2). This section discusses these works, the necessity of addressing ethics in XAI, and the ethical implications of applying XAI in the legal domain.

5.1. Ethics Mentions

The ongoing discourse on ethics guidelines for developing and operating AI systems emphasizes the importance of explainability, transparency, and accountability as ethical principles within the AI domain. Systematic and scope reviews substantiate that these principles rank among the most frequently referenced in this field [63, 64, 65]. Floridi and Cowls [66] assert that "explicability," construed as "intelligibility" and "accountability," stands as the singular novel structural principle that has been appended to the established quartet of bioethics principles — "beneficence, non-maleficence, autonomy, and justice."

Given the prominence of these ethical tenets, it is surprising that only ten papers broach the ethical implications of AI applications, with a mere five specifically addressing the ethical facets of XAI [42, 33, 26, 27, 49]. Among these, two delve into how their proposed XAI solutions ameliorate ethical concerns, particularly about fairness and non-discrimination in legal cases, outperforming similar techniques [42, 27]. Wu et al. [26] proactively include a disclaimer elucidating that their framework ought to be perceived as an auxiliary tool for judges rather than an automatic decision-making system, a distinction made on ethical grounds. Valvoda and Cotterell [49] suggest careful use of their work to automate legal decisions, given that their results indicate unaligned precedents between models and judges. Nielsen et al. [33] call for legal ethicists' attention to a specific experimental result.

It is non-trivial to point out why most categorized papers do not fully address the ethical concerns of their works in a domain as sensitive as law. Perhaps the lack of ethical discussions is due to the focus on technical aspects, which is the main goal in NLP and machine learning and could erroneously indicate that no ethical issues exist [67]. Additionally, authors may be discouraged by the lack of a dedicated space in targeted venues. For instance, there was no extra space for ethical considerations, limitations, and impact statements in *ACL publications until 2021 [67].

5.2. Ethical Implications

Notably, the categorized papers generally do not furnish exhaustive accounts of the ethical implications of the explainability principle in the context of NLP within the legal domain. For instance, there exists potential for discerning between a "legal explanation" and a "model explanation" (refer to Section 6.1), given the longstanding academic discourse on what constitutes sound legal reasoning or a morally and legally sound decision. This discussion is introduced by Atkinson et al. [18][1], in alignment with Robbins [68], who provides a more nuanced perspective on the "explicability principle" itself and critiques the prevailing notion that it should encompass an explication of the algorithm's decision-making process. The author contends in favor of elucidating results rather than processes. Subsequently, Robbins [68] expounds on two overarching approaches to XAI and addresses certain misconceptions about this principle. Even the critiques articulated by Robbins [68] and Atkinson et al. [18] do not fully establish the value of an explanation in the realm of legal decision-making (Section 6.1).

Additional challenges and considerations in explaining AI within the legal domain include the need for judges to maintain control over automated decision-making systems and fully understand their processes [26]. This requirement is critical for AI models to function as supportive tools rather than replace human judgment, thereby reducing the risk of discrimination and bias inherent in models and datasets, which can be particularly damaging in sensitive areas such as family law. Ensuring transparency and fairness in legal case decisions is essential to avoid unjust outcomes [42, 27].

6. Open Issues and Future Work

Despite advances in XAI within NLLP, several unresolved challenges persist. This section outlines some of these issues and suggests potential research directions. We explore the role and value of explanations in the legal domain, propose ways to enhance NLP systems using explanations, and highlight limitations in the current literature.

6.1. Role and Value of Explanations

Researchers exert significant efforts to keep XAI in step with the dynamic landscape of NLP. However, a more concerted endeavor must contemplate XAI's role and implications in the specific legal domain. Explanations are central in most automated decision regulations [63, 64, 65], being deemed critical for ensuring quality control, accountability, and justice [66]. The ethical consequences of algorithms impacting decisions on critical legal matters make the need for clear and interpretable explanations even more pressing, as they help alleviate moral concerns. Explanations are pivotal for several legal stakeholders: they empower judges in their decision-making process [26], assist lawyers and other experts in the analysis of court understandings [41], support the evaluation of AI systems by model creators [37], and provide users with the ability to understand AI-driven decisions, supporting the "right to explanation" [69] or model trust. However, this area remains underexplored in the literature. While complex, a stronger focus on XAI's legal and ethical implications could enhance its perceived importance and the resources devoted to advancing it.

[1]Not categorized as it is a survey.

XAI aims to elucidate how or why the algorithm arrives at a specific conclusion as we understand it. Although this conclusion may align with or contribute to a legal evaluation, the factors influencing algorithmic decisions often diverge significantly from the reasoning employed by legal practitioners. For instance, an AI model's decision-making process may differ from a judge's, even though they can agree on the decision itself. This discrepancy raises a fundamental question regarding the role of XAI within the legal framework: Should explanations be confined to ensuring the decision is robust, meaning any other legal operator would arrive at the same conclusion, or should they also provide insights into the critical legal factors at play? In other words, should the explanation elucidate the juridical reasoning or the machine learning model's decision-making process? The former is indispensable to legal reasoning, given that societal shifts or alterations in the interpretation of legal principles may lead to different conclusions compared to well-established legal precedents. Meanwhile, the latter is crucial for understanding the model's inner workings and ensuring that it is not biased or discriminatory [42, 27]. In this sense, Atkinson et al. [18] and Robbins [68] argue that the explanation should focus on elucidating the legal outcome rather than the AI's internal processes. To affirm this, Robbins [68] assumes that the only object that requires an explanation is the juridical decision. However, as we argue in this work, an explanation has more validity than simply justifying a legal decision and is essential to different stakeholders. Understanding both the legal reasoning and the model's decision-making process is crucial for ensuring transparency, accountability, and trust in AI systems.

6.2. Directions for Improvement with XAI

Most studies reviewed do not leverage XAI techniques to expose the limitations of NLP systems. A notable exception is Bhambhoria et al. [43], where the authors observe that the Longformer model [70] exhibits higher unreliability and susceptibility to spurious correlations compared to the XGBoost model [71] despite the former's greater accuracy. Ideally, XAI insights could be used to identify specific scenarios where an NLP model excels or struggles. This would allow researchers to improve model performance while providing users with crucial information about the contexts in which the model is most dependable — a particularly important consideration in the legal domain, where the stakes of a model's failure can be significant.

Further insights into the limitations of NLP systems could be gleaned by incorporating model confidence scores into the analysis, especially given that most works fall under the "Classification" category (Table 1). Understanding a model's limitations within this context is of paramount importance. An instance of misclassification with low confidence is expected. Conversely, a misclassification by an overconfident model poses greater risks. With the aid of XAI techniques, the former can be linked to a deficiency of relevant features for the model. In contrast, the latter can be attributed to a feature that the model misuses, potentially revealing issues in the training process or model selection. Moreover, confidence scores play a crucial role in score *calibration*, a vital aspect of providing users of an NLP system with interpretable probabilities of model accuracy. Unfortunately, not all models exhibit well-calibrated curves, presenting a challenging hurdle. Thus, demonstrating the relevance

of calibration methods and results in the context of XAI is imperative. Most of the reviewed works make no mention of this crucial aspect — notable exceptions are Resck et al. [41] and Semo et al. [50]. Accurate probability estimates from machine learning classifiers help legal stakeholders assess their confidence in the model's decisions, preventing overreliance on incorrect predictions.

6.3. Limitations

Several limitations in the reviewed literature are worth highlighting. A key issue is the limited number of works focusing on global explanations, particularly the absence of global post-hoc explanations, which are crucial in XAI. Such explanations are essential because global methods can help users understand the model's behavior in general, which is important for legal stakeholders. Post-hoc methods are ideal for black-box models that are not inherently interpretable. Other surveys, such as the one from Danilevsky et al. [13], have also pointed out the scarcity of global explanations. Normally, post-hoc explanations — such as LIME [20], SHAP [72], and input gradients [58] — are employed to explain the model's decision-making process using a specific sample, which may partly explain the absence of global post-hoc explanations.

Another area for improvement is the evaluation of explanations in NLLP. While benchmarks for explanations exist [73], they are not standardized, especially within the legal domain. The lack of a consistent benchmarking framework hinders evaluating, validating, and comparing different explanation methods. This is a significant gap that needs to be addressed to advance the field. Finally, the effectiveness of certain types of explanations, such as attention scores, widely used by local self-explaining methods [32, 34, 22, 33, 37, 28, 38, 39], has been debated in previous work [74, 75].

7. Conclusion

This paper presents a comprehensive survey on the intersection of Explainable AI and Natural Legal Language Processing. We compile a wide range of studies that apply explainability techniques to NLLP tasks and categorize them based on a taxonomy derived from previous research, including explanation types, techniques, and NLP tasks. Through this categorization, we identify trends in how XAI is being applied to NLLP. We also explored works incorporating ethical considerations and discussed the implications of using XAI in the legal domain. Our findings indicate that most papers do not fully address the ethical concerns associated with their research. We outline the challenges and emphasize the need to prioritize ethical considerations when applying XAI to legal contexts. Finally, we discussed open issues and proposed directions for future research, particularly focusing on the role and value of explanations in the legal domain and potential strategies for enhancing NLP systems with more effective explanations.

References

[1] D. M. Katz, D. Hartung, L. Gerlach, A. Jana, M. J. Bommarito II, Natural Language Processing in the Legal Domain, 2023. URL: http://arxiv.org/abs/2302.12039. doi:10.48550/arXiv.2302.12039, arXiv:2302.12039 [cs].

[2] X. Zhou, Q. Liu, Y. Wu, Q. Chen, K. Kuang, Lk-ib: a hybrid framework with legal knowledge injection for compulsory measure prediction, Artificial Intelligence and Law (2023) 1–26.

[3] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, M. Sun, How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5218–5230. URL: https://aclanthology.org/2020.acl-main.466. doi:10.18653/v1/2020.acl-main.466.

[4] M.-Y. Kim, Y. Xu, R. Goebel, Summarization of legal texts with high cohesion and automatic compression rate, in: JSAI-isAI Workshops, 2012. URL: https://api.semanticscholar.org/CorpusID:38025582.

[5] E. Chieze, A. Farzindar, G. Lapalme, An automatic system for summarization and information extraction of legal information, in: Semantic Processing of Legal Texts, 2010. URL: https://api.semanticscholar.org/CorpusID:12554475.

[6] P. H. L. de Araujo, T. E. de Campos, F. A. Braz, N. C. da Silva, Victor: a dataset for brazilian legal documents classification, in: International Conference on Language Resources and Evaluation, 2020. URL: https://api.semanticscholar.org/CorpusID:219299779.

[7] N. C. da Silva, F. A. Braz, T. E. de Campos, A. L. P. Guedes, D. B. Mendes, D. A. Bezerra, D. B. Gusmao, F. B. S. Chaves, G. G. Ziegler, L. H. Horinouchi, M. U. Ferreira, P. H. Inazawa, V. H. D. Coelho, R. V. C. Fernandes, F. Peixoto, M. S. M. Filho, B. P. Sukiennik, L. Rosa, R. P. M. Silva, T. A. Junquilho, G. H. T. Carvalho, Document type classification for brazil's supreme court using a convolutional neural network, Proceedings of The Tenth International Conference on Forensic Computer Science and Cyber Law (2018). URL: https://api.semanticscholar.org/CorpusID:69283834.

[8] B. Strickson, B. De La Iglesia, Legal Judgement Prediction for UK Courts, in: Proceedings of the 3rd International Conference on Information Science and Systems, ICISS '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 204–209. URL: https://doi.org/10.1145/3388176.3388183. doi:10.1145/3388176.3388183.

[9] V. G. F. Bertalan, E. E. S. Ruiz, Predicting judicial outcomes in the brazilian legal system using textual features, in: DHandNLP@PROPOR, 2020. URL: https://api.semanticscholar.org/CorpusID:218906340.

[10] A. Lage-Freitas, H. Allende-Cid, O. V. Santana, L. de Oliveira-Lage, Predicting brazilian court decisions, PeerJ Computer Science 8 (2019). URL: https://api.semanticscholar.org/CorpusID:165164045.

[11] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, V. Lampos, Predicting judicial decisions of the european court of human rights: A natural language processing perspective, PeerJ computer science 2 (2016) e93.

[12] K. Qian, M. Danilevsky, Y. Katsis, B. Kawas, E. Oduor, L. Popa, Y. Li, XNLP: A Living Survey for XAI Research in Natural Language Processing, in: 26th International Conference on Intelligent User Interfaces - Companion, IUI '21 Companion, Association for Computing Machinery, New York, NY, USA, 2021, pp. 78–80. URL: https://dl.acm.org/doi/10.1145/3397482.3450728.

doi:10.1145/3397482.3450728.

[13] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A Survey of the State of Explainable AI for Natural Language Processing, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 447–459. URL: https://aclanthology.org/2020.aacl-main.46.

[14] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du, Explainability for Large Language Models: A Survey, ACM Transactions on Intelligent Systems and Technology 15 (2024) 20:1–20:38. URL: https://dl.acm.org/doi/10.1145/3639372. doi:10.1145/3639372.

[15] L. Resck, I. Augenstein, A. Korhonen, Explainability and Interpretability of Multilingual Large Language Models: A Survey, 2025. URL: https://openreview.net/forum?id=KQjVhM2YhN.

[16] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, Turin, Italy, 2018, pp. 80–89. URL: https://ieeexplore.ieee.org/abstract/document/8631448. doi:10.1109/DSAA.2018.00018.

[17] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, Data Mining and Knowledge Discovery (2023). URL: https://doi.org/10.1007/s10618-022-00867-8. doi:10.1007/s10618-022-00867-8.

[18] K. Atkinson, T. Bench-Capon, D. Bollegala, Explanation in AI and law: Past, present and future, Artificial Intelligence 289 (2020) 103387. URL: https://linkinghub.elsevier.com/retrieve/pii/S0004370220301375. doi:10.1016/j.artint.2020.103387.

[19] L. K. Branting, C. Pfeifer, B. Brown, L. Ferro, J. Aberdeen, B. Weiss, M. Pfaff, B. Liao, Scalable and explainable legal prediction, Artificial Intelligence and Law 29 (2021) 213–238. URL: https://doi.org/10.1007/s10506-020-09273-1. doi:10.1007/s10506-020-09273-1.

[20] M. T. Ribeiro, S. Singh, C. Guestrin, "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, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[21] J. González-González, F. de Arriba-Pérez, S. García-Méndez, A. Busto-Castiñeira, F. J. González-Castaño, Automatic explanation of the classification of Spanish legal judgments in jurisdiction-dependent law categories with tree estimators, Journal of King Saud University - Computer and Information Sciences 35 (2023) 101634. URL: https://www.sciencedirect.com/science/article/pii/S131915782300188X. doi:10.1016/j.jksuci.2023.101634.

[22] Q. Zhao, T. Gao, N. Guo, Legal Judgment Prediction Via Legal Knowledge Fusion and Prompt Learning, 2023. URL: https://papers.ssrn.com/abstract=4341600. doi:10.2139/ssrn.4341600.

[23] R. Bhambhoria, H. Liu, S. Dahan, X. Zhu, Interpretable low-resource legal decision making, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 11819–11827.

[24] L. Li, L. Zhao, P. Nai, X. Tao, Charge prediction modeling with interpretation enhancement driven by double-layer criminal system, World Wide Web 25 (2022) 381–400. URL: https://doi.org/10.1007/s11280-021-00873-8. doi:10.1007/s11280-021-00873-8.

[25] Y. Lyu, Z. Wang, Z. Ren, P. Ren, Z. Chen, X. Liu, Y. Li, H. Li, H. Song, Improving legal judgment prediction through reinforced criminal element extraction, Information Processing & Management 59 (2022) 102780.

[26] Y. Wu, Y. Liu, W. Lu, Y. Zhang, J. Feng, C. Sun, F. Wu, K. Kuang, Towards interactivity and interpretability: A rationale-based legal judgment prediction framework, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 4787–4799.

[27] H. Zhong, Y. Wang, C. Tu, T. Zhang, Z. Liu, M. Sun, Iteratively questioning and answering for interpretable legal judgment prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 1250–1257.

[28] K. Branting, B. Weiss, B. Brown, C. Pfeifer, A. Chakraborty, L. Ferro, M. Pfaff, A. Yeh, Semi-supervised methods for explainable legal prediction, in: Proceedings of the seventeenth international conference on artificial intelligence and law, 2019, pp. 22–31.

[29] H. Chen, D. Cai, W. Dai, Z. Dai, Y. Ding, Charge-Based Prison Term Prediction with Deep Gating Network, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6362–6367. URL: https://aclanthology.org/D19-1667. doi:10.18653/v1/D19-1667.

[30] X. Jiang, H. Ye, Z. Luo, W. Chao, W. Ma, Interpretable Rationale Augmented Charge Prediction System, in: D. Zhao (Ed.), Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Santa Fe, New Mexico, 2018, pp. 146–151. URL: https://aclanthology.org/C18-2032.

[31] K. D. Ashley, S. Brüninghaus, Automatically classifying case texts and predicting outcomes, Artificial Intelligence and Law 17 (2009) 125–165. URL: https://doi.org/10.1007/s10506-009-9077-9. doi:10.1007/s10506-009-9077-9.

[32] V. G. F. Bertalan, E. E. S. Ruiz, Using attention methods to predict judicial outcomes, Artificial Intelligence and Law 32 (2024) 87–115. URL: https://doi.org/10.1007/s10506-022-09342-7. doi:10.1007/s10506-022-09342-7.

[33] A. Nielsen, S. Skylaki, M. Norkute, A. Stremitzer, Effects of xai on legal process (2023).

[34] P. Wang, X. Zhang, H. Yu, Z. Cao, Interpretable prison term prediction with reinforce learning and attention, Applied Intelligence 53 (2023) 1306–1323. URL: https://doi.org/10.1007/s10489-022-03675-1. doi:10.1007/s10489-022-03675-1.

[35] I. Chalkidis, M. Fergadiotis, D. Tsarapatsanis, N. Aletras, I. Androutsopoulos, P. Malakasiotis, Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, ????, pp. 226–241. URL: https://aclanthology.org/2021.naacl-main.22/. doi:10.18653/v1/2021.naacl-main.22.

[36] V. Malik, R. Sanjay, S. K. Nigam, K. Ghosh, S. K. Guha, A. Bhattacharya, A. Modi, ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, ????, pp. 4046–4062. URL: https://aclanthology.org/2021.acl-long.313/. doi:10.18653/v1/2021.acl-long.313.

[37] M. Norkute, N. Herger, L. Michalak, A. Mulder, S. Gao, Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, CHI EA '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–7. URL: https://doi.org/10.1145/3411763.3443441. doi:10.1145/3411763.3443441.

[38] D. Caled, M. Won, B. Martins, M. J. Silva, A hierarchical label network for multi-label eurovoc classification of legislative contents, in: Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23, Springer, 2019, pp. 238–252.

[39] H. Ye, X. Jiang, Z. Luo, W. Chao, Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1854–1864. URL: https://aclanthology.org/N18-1168. doi:10.18653/v1/N18-1168.

[40] J. Mumford, K. Atkinson, T. Bench-Capon, Combining a Legal Knowledge Model with Machine Learning for Reasoning with Legal Cases, in: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 167–176. URL: https://dl.acm.org/doi/10.1145/3594536.3595158. doi:10.1145/3594536.3595158.

[41] L. E. Resck, J. R. Ponciano, L. G. Nonato, J. Poco, Legalvis: Exploring and inferring precedent citations in legal documents, IEEE Transactions on Visualization and Computer Graphics (2023).

[42] I. Benedetto, A. Koudounas, L. Vaiani, E. Pastor, E. Baralis, L. Cagliero, F. Tarasconi, PoliToHFI at SemEval-2023 task 6: Leveraging entity-aware and hierarchical transformers for legal entity recognition and court judgment prediction, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1401–1411. URL: https://aclanthology.org/2023.semeval-1.194/. doi:10.18653/v1/2023.semeval-1.194.

[43] R. Bhambhoria, S. Dahan, X. Zhu, Investigating the state-of-the-art performance and explainability of legal judgment prediction., in: Canadian Conference on AI, 2021.

[44] L. E. R. Domingues, Inferring and explaining potential citations to binding precedents in brazilian supreme court decisions (2021).

[45] Ł. Górski, S. Ramakrishna, Explainable artificial intelligence, lawyer's perspective, in: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 2021, pp. 60–68.

[46] J. Rabelo, M.-Y. Kim, R. Goebel, M. Yoshioka, Y. Kano, K. Satoh, Coliee 2020: methods for legal document retrieval and entailment, in: New Frontiers in Artificial Intelligence: JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers 12, Springer, 2021, pp. 196–210.

[47] J. Rabelo, M.-Y. Kim, R. Goebel, M. Yoshioka, Y. Kano, K. Satoh, A summary of the coliee 2019 competition, in: New Frontiers in Artificial Intelligence: JSAI-isAI International Workshops, JURISIN, AI-Biz, LENLS, Kansei-AI, Yokohama, Japan, November 10–12, 2019, Revised Selected Papers 10, Springer, 2020, pp. 34–49.

[48] R. Chhatwal, P. Gronvall, N. Huber-Fliflet, R. Keeling, J. Zhang, H. Zhao, Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding, in: 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 1905–1911. URL: https://ieeexplore.ieee.org/document/8622073. doi:10.1109/BigData.2018.8622073.

[49] J. Valvoda, R. Cotterell, Towards Explainability in Legal Outcome Prediction Models, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 7269–7289. URL: https://aclanthology.org/2024.naacl-long.404/. doi:10.18653/v1/2024.naacl-long.404.

[50] G. Semo, D. Bernsohn, B. Hagag, G. Hayat, J. Niklaus, Classactionprediction: A challenging benchmark for legal judgment prediction of class action cases in the us, arXiv preprint arXiv:2211.00582 (2022).

[51] S. T.y.s.s, S. Xu, O. Ichim, M. Grabmair, Deconfounding legal judgment prediction for european court of human rights cases towards better alignment with experts, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, ????, pp. 1120–1138. URL: https://aclanthology.org/2022.emnlp-main.74/. doi:10.18653/v1/2022.emnlp-main.74.

[52] C. J. Mahoney, J. Zhang, N. Huber-Fliflet, P. Gronvall, H. Zhao, A Framework for Explainable Text

Classification in Legal Document Review, IEEE Computer Society, 2019, pp. 1858–1867. URL: https://www.computer.org/csdl/proceedings-article/big-data/2019/09005659/1hJsCablZfy. doi:10.1109/BigData47090.2019.9005659.

[53] B. Waltl, G. Bonczek, E. Scepankova, F. Matthes, Semantic types of legal norms in German laws: classification and analysis using local linear explanations, Artificial Intelligence and Law 27 (2019) 43–71. URL: https://doi.org/10.1007/s10506-018-9228-y. doi:10.1007/s10506-018-9228-y.

[54] Ł. Górski, S. Ramakrishna, J. M. Nowosielski, Towards grad-cam based explainability in a legal text processing pipeline. extended version, in: International Workshop on AI Approaches to the Complexity of Legal Systems, Springer, 2018, pp. 154–168.

[55] J. Landthaler, I. Glaser, F. Matthes, Towards Explainable Semantic Text Matching, Legal Knowledge and Information Systems 313 (2018). doi:10.3233/978-1-61499-935-5-200.

[56] F. de Arriba-Pérez, S. García-Méndez, F. J. González-Castaño, J. González-González, Explainable machine learning multi-label classification of spanish legal judgements, Journal of King Saud University - Computer and Information Sciences 34 (2022) 10180–10192. URL: https://www.sciencedirect.com/science/article/pii/S1319157822003664. doi:https://doi.org/10.1016/j.jksuci.2022.10.015.

[57] M. Medvedeva, M. Vols, M. Wieling, Using machine learning to predict decisions of the European Court of Human Rights, Artificial Intelligence and Law 28 (2020) 237–266. URL: https://doi.org/10.1007/s10506-019-09255-y. doi:10.1007/s10506-019-09255-y.

[58] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[59] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[60] L. Resck, M. M. Raimundo, J. Poco, Exploring the Trade-off Between Model Performance and Explanation Plausibility of Text Classifiers Using Human Rationales, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 4190–4216. URL: https://aclanthology.org/2024.findings-naacl.262. doi:10.18653/v1/2024.findings-naacl.262, also presented as a poster at the LatinX in NLP at NAACL 2024 workshop.

[61] Z. Zhang, K. Rudra, A. Anand, Explain and Predict, and then Predict Again, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, Virtual Event Israel, 2021, pp. 418–426. URL: https://doi.org/10.1145/3437963.3441758. doi:10.1145/3437963.3441758.

[62] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, L. Zettlemoyer, An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 1938–1952. URL: https://aclanthology.org/2020.emnlp-main.153. doi:10.18653/v1/2020.emnlp-main.153.

[63] E. Prem, From ethical ai frameworks to tools: a review of approaches, AI and Ethics 3 (2023) 699–716. URL: https://link.springer.com/article/10.1007/s43681-023-00258-9.

[64] T. Hagendorff, The ethics of ai ethics: An evaluation of guidelines, Minds and Machines 30 (2020) 99–120. URL: https://link.springer.com/article/10.1007/s11023-020-09517-8.

[65] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature machine intelligence 1 (2019) 389–399. URL: https://www.nature.com/articles/s42256-019-0088-2.

[66] L. Floridi, J. Cowls, A unified framework of five principles for ai in society, Harvard Data Science Review 1 (2019). URL: https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8. doi:https://doi.org/10.1162/99608f92.8cd550d1.

[67] L. Benotti, P. Blackburn, Ethics consideration sections in natural language processing papers, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4509–4516. URL: https://aclanthology.org/2022.emnlp-main.299. doi:10.18653/v1/2022.emnlp-main.299.

[68] S. Robbins, A Misdirected Principle with a Catch: Explicability for AI, Minds and Machines 29 (2019) 495–514. URL: https://doi.org/10.1007/s11023-019-09509-3. doi:10.1007/s11023-019-09509-3.

[69] A. Selbst, J. Powles, "Meaningful Information" and the Right to Explanation, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, New York, 2018, pp. 48–48. URL: https://proceedings.mlr.press/v81/selbst18a.html, iSSN: 2640-3498.

[70] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, ArXiv abs/2004.05150 (2020). URL: https://api.semanticscholar.org/CorpusID:215737171.

[71] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, et al., Xgboost: extreme gradient boosting, R package version 0.4-2 1 (2015) 1–4.

[72] S. M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems 30 (NIPS 2017), volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[73] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, OpenXAI: towards a transparent evaluation of post hoc model explanations, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2024, pp. 15784–15799. URL: https://dl.acm.org/doi/10.5555/3600270.3601418.

[74] S. Jain, B. C. Wallace, Attention is not Explanation, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceed-

ings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3543–3556. URL: https://aclanthology.org/N19-1357. doi:10.18653/v1/N19-1357.

[75] S. Wiegreffe, Y. Pinter, Attention is not not Explanation, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 11–20. URL: https://aclanthology.org/D19-1002. doi:10.18653/v1/D19-1002.