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] Jorge Poco[1]
Correspondence: lucas.domingues@fgv.edu.br

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

16 June 2025
7th Workshop on Automated Semantic Analysis of Information in Legal Text
ASAIL 2025, co-located with ICAIL 2025

Table of Contents

Motivation and contributions

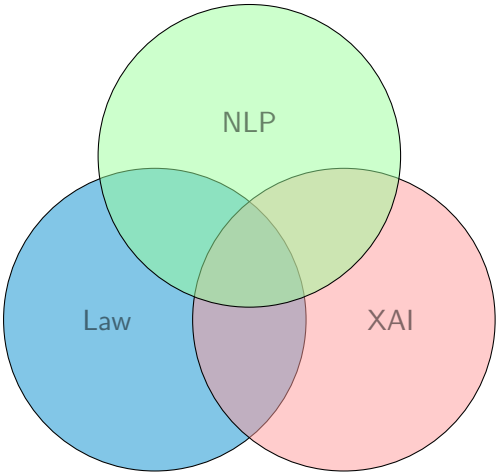
- Increasing application of NLP in the legal domain
- But limited efforts to enhance their understanding

- We present a survey of explainable AI (XAI) applied to legal NLP:
  - Intersection of NLP, law and XAI
  - Essential intersection

- Main contributions:
  1. A **survey** on explainable natural legal language processing.
  2. **Taxonomy** based on the NLP task, explanation type and technique employed.
  3. Analysis of **research trends** in types of explanations and the utilisation of XAI.
  4. Discussion of **ethics in XAI** and the legal domain, **open issues and future work**.

Table of Contents

Previous surveys



We build on previous surveys:

- Danilevsky et al. [14] and Qian et al. [30]: survey on **XAI** applied to **NLP**.
  - Categorisation of explanation types and explainability methods employed.
- Katz et al. [21]: survey on **legal NLP**.
  - Categorisation of NLLP engineering tasks.

Table of Contents

Taxonomy

We propose the following taxonomy:

Explanation type [14, 30]:

- Local vs global
- Post-hoc vs self-explaining

Explainability technique [14]:

- Feature importance
- Surrogate model
- Provenance-based
- Declarative induction

NLP task [21]:

- Classification
- Text generation
- etc.

Ethical issues:

- Ethics mentions
- Ethics mentions in the context of XAI

Table of Contents

Local post-hoc

Explanation Type	Explainability Technique	NLP Task	Representative Works
Local post-hoc	Surrogate model	Classification, Information retrieval, Resources	**Benedetto et al. [4]**, Resck et al. [33], Bhambhoria, Dahan, and Zhu [6], Domingues [16], Górski and Ramakrishna [18], Rabelo et al. [32], Rabelo et al. [31], Chhatwal et al. [13]
	Feature importance	Classification, Information retrieval, Machine summarization, Resources	*Semo et al. [37]*, *T.y.s.s et al. [40]*, Górski and Ramakrishna [18], *Malik et al. [26]*, Norkute et al. [29], Mahoney et al. [25], Waltl et al. [41], Górski, Ramakrishna, and Nowosielski [19], Landthaler, Glaser, and Matthes [22]
	Declarative induction	Classification	de Arriba-Pérez et al. [15]

- **Notable XAI techniques:**
  - LIME [34, 33, 6]
  - Input gradients [39, 36, 4, 37]
- Combination of techniques [18, 29]

- Sentence-level instead of token-level [4]
- Information retrieval [33, 22]
- Machine summarisation [29]

Local self-explaining

Explanation Type	Explainability Technique	NLP Task	Representative Works
Local self-exp.	Provenance-based	Classification	Zhao, Gao, and Guo [45], Bhambhoria et al. [7], Li et al. [23], Lyu et al. [24], **Wu et al. [43]**, Branting et al. [9], **Zhong et al. [46]**, Branting et al. [8], *Chen et al. [12]*, Jiang et al. [20], Ashley and Brüninghaus [2]
	Feature importance	Classification, Machine summarization, Resources, Text generation	Bertalan and Ruiz [5], **Nielsen et al. [28]**, Wang et al. [42], *Zhou et al. [47]*, Branting et al. [9], Chalkidis et al. [11], *Malik et al. [26]*, Norkute et al. [29], Branting et al. [8], Caled et al. [10], Ye et al. [44]
	Declarative induction	Text generation	Ye et al. [44]
	Surrogate model	Classification, Information retrieval, Resources	Resck et al. [33]

- **Attention weights** [45, 28, 29]
- Classifier weights [47]
- Word- vs sentence-level [45]
- Summarisation [28, 29]
- Text generation [44]
- Intermediate labels [24, 43]

Global self-explaining

Explanation Type	Explainability Technique	NLP Task	Representative Works
Global self-exp.	Feature importance	Classification	Medvedeva, Vols, and Wieling [27], Strickson and De La Iglesia [38], Aletras et al. [1]
	Declarative induction	Classification	González-González et al. [17], de Arriba-Pérez et al. [15]

- **Classical machine learning:**
  - Feature importance of support-vector machines [1, 27]
  - Declarative induction of random forests [15, 17]
  - TF-IDF [38]

Table of Contents

Ethics mentions

- We identified papers that mention ethics:
  - **Nine** papers only
- A mere **four** papers leverage ethics in XAI:
  - How the explanations **ameliorate ethical issues** like fairness and non-discrimination [46, 4].
  - The framework being used to **auxiliate judges** rather than substitute them [43].
  - Ethics of one particular experimental result [28].
- Hypothesis: most works focus on the **technical aspects** of NLP and machine learning

Table of Contents

Role and value of explanations

- Value of explanations to various stakeholders:
  - Empower judges
  - Assist lawyers and other experts
  - Support the evaluation of AI systems by model creators
  - Provide users with the ability to understand model decisions ("right to explanation")
- Legal vs model explanation:
  - Some argue that explanations should elucidate the legal outcome only [3, 35].
  - We argue that model explanations are also valuable to other stakeholders.

Limitations and directions for improvement

Directions for improvement with XAI:

- Expose legal NLP systems limitations
- Incorporation of confidence scores

Limitations:

- Limited work on global explanations
- Limited evaluation of explanations

Table of Contents

Conclusion and takeaways

- A **survey** on explainable natural legal language processing:
  - **Taxonomy** based on the NLP task, explanation type and technique employed
  - **Trends** in how XAI is utilised in legal NLP
- Papers generally do not thoroughly examine or discuss:
  - **Ethical** implications
  - The **role and value** of explanations
  - The potential of explanations for pointing out **limitations** of legal NLP systems

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] Jorge Poco[1]
Correspondence: lucas.domingues@fgv.edu.br

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

16 June 2025
7th Workshop on Automated Semantic Analysis of Information in Legal Text
ASAIL 2025, co-located with ICAIL 2025

Bibliography I

[1] Nikolaos Aletras et al. "Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective". In: *PeerJ computer science* 2 (2016), e93.

[2] Kevin D. Ashley and Stefanie Brüninghaus. "Automatically classifying case texts and predicting outcomes". en. In: *Artificial Intelligence and Law* 17.2 (June 2009), pp. 125–165. ISSN: 1572-8382. DOI: 10.1007/s10506-009-9077-9. URL: https://doi.org/10.1007/s10506-009-9077-9.

[3] Katie Atkinson, Trevor Bench-Capon, and Danushka Bollegala. "Explanation in AI and law: Past, present and future". en. In: *Artificial Intelligence* 289 (Dec. 2020), p. 103387. ISSN: 00043702. DOI: 10.1016/j.artint.2020.103387. URL: https://linkinghub.elsevier.com/retrieve/pii/S0004370220301375.

Bibliography II

[4] Irene Benedetto et al. "PoliToHFI at SemEval-2023 Task 6: Leveraging Entity-Aware and Hierarchical Transformers For Legal Entity Recognition and Court Judgment Prediction". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha et al. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1401–1411. DOI: 10.18653/v1/2023.semeval-1.194. URL: https://aclanthology.org/2023.semeval-1.194/.

[5] Vithor Gomes Ferreira Bertalan and Evandro Eduardo Seron Ruiz. "Using attention methods to predict judicial outcomes". en. In: *Artificial Intelligence and Law* 32.1 (Mar. 2024), pp. 87–115. ISSN: 1572-8382. DOI: 10.1007/s10506-022-09342-7. URL: https://doi.org/10.1007/s10506-022-09342-7.

[6] Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. "Investigating the State-of-the-Art Performance and Explainability of Legal Judgment Prediction.". In: *Canadian Conference on AI*. 2021.

Bibliography III

[7] Rohan Bhambhoria et al. "Interpretable low-resource legal decision making". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 11819–11827.

[8] Karl Branting et al. "Semi-supervised methods for explainable legal prediction". In: *Proceedings of the seventeenth international conference on artificial intelligence and law*. 2019, pp. 22–31.

[9] L. Karl Branting et al. "Scalable and explainable legal prediction". en. In: *Artificial Intelligence and Law* 29.2 (June 2021), pp. 213–238. ISSN: 1572-8382. DOI: 10.1007/s10506-020-09273-1. URL: https://doi.org/10.1007/s10506-020-09273-1.

Bibliography IV

[10] Danielle Caled et al. “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.

[11] Ilias Chalkidis et al. “Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2021. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, June 2021, pp. 226–241. DOI: 10.18653/v1/2021.naacl-main.22. URL: https://aclanthology.org/2021.naacl-main.22/.

Bibliography V

[12] Huajie Chen et al. "Charge-Based Prison Term Prediction with Deep Gating Network". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6362–6367. DOI: 10.18653/v1/D19-1667. URL: https://aclanthology.org/D19-1667.

[13] Rishi Chhatwal et al. "Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding". In: *2018 IEEE International Conference on Big Data (Big Data)*. Dec. 2018, pp. 1905–1911. DOI: 10.1109/BigData.2018.8622073. URL: https://ieeexplore.ieee.org/document/8622073.

Bibliography VI

[14] Marina Danilevsky et al. "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*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 447–459. URL: https://aclanthology.org/2020.aacl-main.46.

[15] Francisco de Arriba-Pérez et al. "Explainable machine learning multi-label classification of Spanish legal judgements". In: *Journal of King Saud University - Computer and Information Sciences* 34.10, Part B (2022), pp. 10180–10192. ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2022.10.015. URL: https://www.sciencedirect.com/science/article/pii/S1319157822003664.

[16] Lucas Emanuel Resck Domingues. "Inferring and explaining potential citations to binding precedents in Brazilian Supreme Court Decisions". In: (2021).

Bibliography VII

[17] Jaime González-González et al. "Automatic explanation of the classification of Spanish legal judgments in jurisdiction-dependent law categories with tree estimators". In: *Journal of King Saud University - Computer and Information Sciences* 35.7 (July 2023), p. 101634. ISSN: 1319-1578. DOI: 10.1016/j.jksuci.2023.101634. URL: https://www.sciencedirect.com/science/article/pii/S131915782300188X.

[18] Łukasz Górski and Shashishekar Ramakrishna. "Explainable artificial intelligence, lawyer's perspective". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 2021, pp. 60–68.

[19] Łukasz Górski, Shashishekar Ramakrishna, and Jédrzej 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.

Bibliography VIII

[20] Xin Jiang et al. "Interpretable Rationale Augmented Charge Prediction System". In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Ed. by Dongyan Zhao. Santa Fe, New Mexico: Association for Computational Linguistics, Aug. 2018, pp. 146–151. URL: https://aclanthology.org/C18-2032.

[21] Daniel Martin Katz et al. *Natural Language Processing in the Legal Domain*. arXiv:2302.12039 [cs]. Feb. 2023. DOI: 10.48550/arXiv.2302.12039. URL: http://arxiv.org/abs/2302.12039.

[22] Jörg Landthaler, Ingo Glaser, and F. Matthes. "Towards Explainable Semantic Text Matching". en. In: *Legal Knowledge and Information Systems*. Frontiers in Artificial Intelligence and Applications 313 (2018). DOI: 10.3233/978-1-61499-935-5-200. URL: https://www.semanticscholar.org/paper/Towards-Explainable-Semantic-Text-Matching-Landthaler-Glaser/e53c56ab762d51e141a3747416b7bcb489631e0f.

Bibliography IX

[23] Lin Li et al. “Charge prediction modeling with interpretation enhancement driven by double-layer criminal system”. en. In: *World Wide Web* 25.1 (Jan. 2022), pp. 381–400. ISSN: 1573-1413. DOI: 10.1007/s11280-021-00873-8. URL: https://doi.org/10.1007/s11280-021-00873-8.

[24] Yougang Lyu et al. “Improving legal judgment prediction through reinforced criminal element extraction”. In: *Information Processing & Management* 59.1 (2022), p. 102780.

[25] Christian J. Mahoney et al. “A Framework for Explainable Text Classification in Legal Document Review”. English. In: IEEE Computer Society, Dec. 2019, pp. 1858–1867. ISBN: 978-1-72810-858-2. DOI: 10.1109/BigData47090.2019.9005659. URL: https://www.computer.org/csdl/proceedings-article/big-data/2019/09005659/1hJsCablZfy.

Bibliography X

[26] Vijit Malik et al. "ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation". In: *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)*. ACL-IJCNLP 2021. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 4046–4062. DOI: 10.18653/v1/2021.acl-long.313. URL: https://aclanthology.org/2021.acl-long.313/.

[27] Masha Medvedeva, Michel Vols, and Martijn Wieling. "Using machine learning to predict decisions of the European Court of Human Rights". en. In: *Artificial Intelligence and Law* 28.2 (June 2020), pp. 237–266. ISSN: 1572-8382. DOI: 10.1007/s10506-019-09255-y. URL: https://doi.org/10.1007/s10506-019-09255-y.

[28] Aileen Nielsen et al. "Effects of XAI on Legal Process". In: (2023).

Bibliography XI

[29] Milda Norkute et al. “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. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–7. ISBN: 978-1-4503-8095-9. DOI: 10.1145/3411763.3443441. URL: https://doi.org/10.1145/3411763.3443441.

[30] Kun Qian et al. “XNLP: A Living Survey for XAI Research in Natural Language Processing”. In: *26th International Conference on Intelligent User Interfaces - Companion*. IUI '21 Companion. New York, NY, USA: Association for Computing Machinery, 2021, pp. 78–80. ISBN: 978-1-4503-8018-8. DOI: 10.1145/3397482.3450728. URL: https://dl.acm.org/doi/10.1145/3397482.3450728.

Bibliography XII

[31] Juliano Rabelo et al. "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.

[32] Juliano Rabelo et al. "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.

[33] Lucas Emanuel Resck et al. "Legalvis: Exploring and inferring precedent citations in legal documents". In: *IEEE Transactions on Visualization and Computer Graphics* (2023).

Bibliography XIII

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

[35] Scott Robbins. "A Misdirected Principle with a Catch: Explicability for AI". en. In: *Minds and Machines* 29.4 (Dec. 2019), pp. 495–514. ISSN: 1572-8641. DOI: 10.1007/s11023-019-09509-3. URL: https://doi.org/10.1007/s11023-019-09509-3.

[36] Ramprasaath R Selvaraju et al. "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.

Bibliography XIV

[37] Gil Semo et al. "ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US". In: *arXiv preprint arXiv:2211.00582* (2022).

[38] Benjamin Strickson and Beatriz De La Iglesia. "Legal Judgement Prediction for UK Courts". In: *Proceedings of the 3rd International Conference on Information Science and Systems*. ICISS '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 204–209. ISBN: 978-1-4503-7725-6. DOI: 10.1145/3388176.3388183. URL: https://doi.org/10.1145/3388176.3388183.

[39] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

Bibliography XV

[40] Santosh T.y.s.s et al. "Deconfounding Legal Judgment Prediction for European Court of Human Rights Cases Towards Better Alignment with Experts". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2022. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1120–1138. DOI: 10.18653/v1/2022.emnlp-main.74. URL: https://aclanthology.org/2022.emnlp-main.74/.

[41] Bernhard Waltl et al. "Semantic types of legal norms in German laws: classification and analysis using local linear explanations". en. In: *Artificial Intelligence and Law* 27.1 (Mar. 2019), pp. 43–71. ISSN: 1572-8382. DOI: 10.1007/s10506-018-9228-y. URL: https://doi.org/10.1007/s10506-018-9228-y.

Bibliography XVI

[42] Peipeng Wang et al. “Interpretable prison term prediction with reinforce learning and attention”. en. In: *Applied Intelligence* 53.2 (Jan. 2023), pp. 1306–1323. ISSN: 1573-7497. DOI: 10.1007/s10489-022-03675-1. URL: https://doi.org/10.1007/s10489-022-03675-1.

[43] Yiquan Wu et al. “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.

[44] Hai Ye et al. “Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1854–1864. DOI: 10.18653/v1/N18-1168. URL: https://aclanthology.org/N18-1168.

Bibliography XVII

[45] Qihui Zhao, Tianhan Gao, and Nan Guo. *Legal Judgment Prediction Via Legal Knowledge Fusion and Prompt Learning*. en. SSRN Scholarly Paper. Rochester, NY, Jan. 2023. DOI: 10.2139/ssrn.4341600. URL: https://papers.ssrn.com/abstract=4341600.

[46] Haoxi Zhong et al. "Iteratively questioning and answering for interpretable legal judgment prediction". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 1250–1257.

[47] Xiang Zhou et al. "LK-IB: a hybrid framework with legal knowledge injection for compulsory measure prediction". In: *Artificial Intelligence and Law* (2023), pp. 1–26.