Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

doi:10.3233/FAIA230971

CiteCaseLAW: Citation Worthiness Detection in Caselaw for Legal Assistive Writing

Mann KHATRI [a,1], Reshma SHEIK [b] Pritish WADHWA [a] Gitansh SATIJA [a]
Yaman KUMAR [c] Rajiv Ratn SHAH [a] and Ponnurangam KUMARAGURU [d]

[a] *IIIT Delhi*
[b] *NIT, Trichy*
[c] *Adobe MDSR*
[d] *IIIT Hyderabad*

ORCiD ID: Mann Khatri https://orcid.org/0000-0002-5132-9223, Reshma Sheik https://orcid.org/0000-0003-3567-9757, Pritish Wadhwa https://orcid.org/0009-0009-7676-8108, Gitansh Satija https://orcid.org/0009-0003-1818-3597, Yaman Kumar https://orcid.org/0000-0001-7880-8219, Rajiv Ratn Shah https://orcid.org/0000-0003-1028-9373, Ponnurangam Kumaraguru https://orcid.org/0000-0001-5082-2078

Abstract. Complex legal language, filled with jargon, nuanced language semantics, and a high level of domain specificity, poses a significant challenge for automation in handling various legal tasks. In the realm of legal document composition, a pivotal component revolves around accurately referencing case laws and other sources to substantiate assertions and arguments. Understanding the legal domain and identifying appropriate citation context or cite-worthy sentences automatically is challenging. Our research is centered on the issue of citation-worthiness identification of a given sentence. This serves as the initial phase in contemporary citation recommendation systems, aimed at alleviating the effort involved in extracting a suitable array of citation contexts. To address this, we first introduce a labeled dataset comprising 178 million sentences, specifically tailored for detecting citation-worthy content within the legal domain. This dataset is curated from the Caselaw Access Project (CAP).[2] We proceeded to assess the performance of a range of deep learning models on this novel dataset. Among the models examined, the domain-specific pre-trained model consistently demonstrated superior performance, achieving an 88% F1-score in the task of detecting citation-worthy material. To enhance our insights, we employed inputXGradient explainable AI techniques to dissect the predictions, thereby identifying the tokens that contribute to specific citation classes.

Keywords. Legal NLP, Citation, Classification

[1] Corresponding Author: Mann Khatri, mannk@iiitd.ac.in

1. Introduction

Accurate source citation is indispensable in legal documentation, especially in case-law-based legal systems that establish crucial links between cases. Identifying sentences worthy of citation involves recognizing sentences that refer to external sources. We aim to delineate the essential components that render a sentence citation-worthy, classifying sentences into either "cite" or "not cite". This task forms the initial stage in a citation recommendation system as the effectiveness of such recommendations critically depends on precisely identifying these sentences as they steer subsequent stages of the process. This identification process facilitates intelligent writing and reduces the burden on legal professionals when composing legal documents.

Our primary aim is to curate an extensive dataset for detecting citation-worthiness at the sentence level within American legal texts. Creating this dataset involves extracting various sentence types (outlined in Section F.2 of Appendix[3]) from legal documents, annotating each sentence to denote the presence of citations, and eliminating citations and undesirable sentences. The primary challenges encountered in developing an effective citation detection system revolve around the volume and quality of data. A sizable, well-annotated legal corpus is imperative for effectively training deep learning models. Challenges related to sentence boundaries and references to external legal sources significantly impact data quality as their incorrect detection can lead to incomplete parsing and citation data, introducing noise. A substantial dataset tailored for citation detection in the legal domain needs to be improved. This dataset will serve as a cornerstone for training future applications in legal writing assistance. In our approach, we utilized machine learning algorithms such as LEGAL-BERT [1] and positive-unlabeled learning to evaluate citation detection on this dataset. Our research aims to address the following key questions: **RQ1:** How can a dataset for detecting citation worthiness in the legal domain be automatically generated with minimal noise, even without relying on domain-specific tokenizers/segmenters? **RQ2:** What techniques are the most reliable for identifying sentences worthy of citation in the legal domain? **RQ3:** To what extent do models trained on the citation worthiness dataset perform compared to established benchmarks for other legal text classification tasks? In summary, our contributions in this research are as follows:

1. We present a dataset[4] for the citation worthiness detection task extracted from the Caselaw Access Project (CAP). Our corpus comprises 178 million sentences for the citation-worthiness detection task.
2. We conducted comprehensive experiments with various state-of-the-art models, quantitatively evaluating them and establishing them as baselines for citation-worthiness detection (see Section 4). Furthermore, we conducted ablation experiments to interpret the model's performance, utilizing the explainable AI inputX-Gradient method on this binary classification task to identify token contributions in each cite class.

[3] Appendices, code and dataset creation steps can be found at `https://drive.google.com/drive/folders/1ZSEaaQFGGUassSiWEvZrBNKEmvQ2c5MB?usp=sharing`

[4] `https://huggingface.co/datasets/Vidhaan/LegalCitationWorthiness`

2. Related Work

Citation-Worthiness in Legal and Scientific Texts: The exploration of citation worthiness, a topic pioneered in scientific language, is crucial in the legal domain yet has received limited attention. In scientific literature, Sugiyama et al. [2] initiated this domain by creating a dataset from the ACL Anthology Reference corpus, employing heuristics to remove citation markers. Farber et al. [3] and Bonab et al. [4] utilized convolutional recurrent neural networks on diverse datasets. Context-aware citation detection was introduced by Gosangi et al. [5] with the ACL-cite dataset, integrating BiLSTMs and transformer-based embeddings. Wright et al. [6] delved into citation worthiness extensively, incorporating domain adaptation and transfer learning techniques. Zeng et al. [7] utilized BiLSTMs and highlighted the importance of adjacent sentence context. In a recent study [8], emphasis was placed on sentence-level citation worthiness, incorporating syntax-based learning and down-sampling analyses.
Works related to citations in the legal domain: In legal texts, research by Savelka et al. [9] demonstrated the challenges legal decisions pose to existing sentence boundary detection systems. Sanchez [10] explored methods to identify sentence breaks in legal language, acknowledging the complexities introduced by punctuation and syntax. Notably, Huang et al. [11] utilized the Board of Veterans' Appeal (BVA) corpus, a substantial dataset containing over a million appeal decisions, to study citation contexts in legal texts. Despite these efforts, there remains a significant gap: the need for a suitable dataset for identifying citation-worthy sentences in the legal domain.

3. Experimentation and Dicussion

We experimented with different models trained on our dataset to establish the baselines for the task of citation-worthiness detection (*RQ2*). For this assessment, we used our subset with 1M entries[5]. The split contains sentences sampled over all jurisdictions. A thorough hyperparameter search is carried out and mentioned in the Appendix Section D.3. The models are logistic regression model with tf-idf features, a CRNN [3], vanilla Transformer [12], Bert [13] and LEGAL-BERT (with and without PU learning). More details are in Section D.1 of Appendix.

Table 1 presents the classification performance of the models employed in our study. Notably, the pre-trained transformer models performed better than logistic regression and other deep-learning models. Among these, LEGAL-BERT stood out, surpassing all the mentioned models regarding classification scores. We incorporated Positive-Unlabeled (PU) learning into the LEGAL-BERT model to enhance its capabilities further. In the Appendix, in Tables 5 and 6, we provide detailed state-wise results derived from the LEGAL-BERT+PU model.

We examined the model's performance on other legal tasks using the UNFAIR-Tos [14] and LEDGAR [15] datasets (See Appendix D.4). The main objective of experimenting on these datasets is to establish a benchmark by including the task related to contracts, as contracts contain limited citations to hypothesise that our model can be used in related legal tasks. As LEGAL-BERT's training corpus included data from the European and American legal domains, the contracts given in the task are from the same.

[5] https://drive.google.com/file/d/1i8bzZNQVfTrFT_2uV3gMbJwbIztpT53S/view?usp=sharing

Model	P	R	F1
Logistic Regression	77.85	75.77	76.79
CRNN	76.54	74.72	74.93
Transformer	72.42	84.25	77.89
Longformer	87.10	86.02	86.56
BERT	**87.73**	86.56	87.14
LEGAL-BERT	87.64	**87.2**	**87.42**
LEGAL-BERT + PU	84.17	**92.86**	**88.30**

Table 1. Classification results on the dataset of different models in terms of Precision (P), Recall (R), and F1-score (F1).

Model	**UNFAIR ToS Dataset**		**LEDGAR Dataset**	
	μ-F1	m-F1	μ-F1	m-F1
LEGAL-BERT (Reference)	96.0	83.0	88.2	82.5
LEGAL-BERT (CiteCaseLaw)	**96.2**	**84.2**	88.2	**83.0**
LEGAL-BERT w/ PU (CiteCaseLaw)	96.1	83.5	**88.4**	82.7

Table 2. Results of F1 score based on Transfer Learning on Legal datasets. Comparable performance showed that fine-tuning with cite-worthiness data did not lead to any performance degrade

4. Discussion

The outcomes presented in Table 1 emphasize the significance of incorporating domain expertise into pre-trained models, showcasing notable performance improvements [1,16]. Furthermore, integrating Positive-Unlabeled (PU) learning amplifies the model's ability to retrieve pertinent instances by augmenting token confidence. This augmentation enhances the model's resilience in tasks related to detecting citation-worthiness. To elucidate the efficacy of PU learning, we offer an illustrative example in Figure 1 in the Appendix, demonstrating how the model's predictions adapt based on token contributions to the classes.

In a study conducted by [17], the InputXGradient method [18], specifically the variant utilizing L2 normalization over neurons to derive a pre-embedding score, exhibited the highest agreement with human reasoning. This method involves post hoc multiplication of the input by the output gradient concerning the input. Building upon this foundation, we applied our domain-specific models, consistently ranking among the top performers. We computed token contributions for each class, as depicted in Figure 5 in the Appendix. Remarkably, in LEGAL-BERT, pivotal unigram tokens for classes 0 and 1 were the period (.), "report," and "requirements," respectively. However, in the context of training LEGAL-BERT with the PU setting, the distribution of contributions shifted, revealing the influence of additional tokens.

Turning attention to Table 2, we scrutinized the model's performance on diverse datasets after fine-tuning it for the citation-worthiness task. Our objective was to demonstrate that these fine-tuned models do not underperform when compared to established benchmarks. The outcomes affirm that fine-tuning the language model on our dataset significantly enhances its performance. This finding aligns seamlessly with prior research, underscoring the importance of refining language model fine-tuning with in-domain data to elevate end-task performance [19]. Consequently, we address our *RQ3* by confirming that the model not only maintains its performance post fine-tuning but also performs at least on par with established baselines.

a) [CLS] if it is to be modified , it is up to that court to do so , not this court . [SEP]
b) [CLS] if it is to be modified , it is up to that court to do so , not this court . [SEP]

Figure 1. A cite-worthy sentence. a) LEGAL-BERT+PU classified it as citeworthy and b) LEGAL-BERT classified it as sentence non-citeworthy. Darker the color more the relative contribution.

5. Conclusion and Limitations

In this study, we curated an expansive dataset tailored for the citation-worthiness task within the American legal domain. Our exploration of various models revealed the superiority of domain-specific pre-trained language models over others. This finding underscores the practical utility of these models within the legal community, particularly in identifying citation-worthy sentences during drafting judgments. Additionally, our dataset, CiteCaseLaw, serves as a valuable testing ground for transfer-learning setups, showcasing the adaptability of these models for downstream natural language understanding tasks. Beyond its immediate application, our dataset holds promise for various subsequent tasks, including citation recommendations, assessing the relevance of citations, and summarizing judgments based on citation analysis. We firmly believe that the broader research community delving into challenges within the realm of legal language processing will find both our dataset and the associated fine-tuned models to be invaluable resources.

In our research, we conducted experiments using a subset of the dataset. However, more GPU availability could have improved our ability to scale the experiments, causing each epoch to require 36 hours for processing. Another constraint we faced was validating our data, which was performed on a relatively small set of 1,000 gold standard examples due to financial limitations. Although expanding this validation capacity is feasible, it was restricted at the time of the study. In our efforts to broaden the scope of our research to encompass legal citation recommendations, one potential avenue involves incorporating metadata with citation links. Our primary objective remains the prediction of citation significance at the sentence level. However, automating the evaluation of preceding sentences for citation relevance poses significant challenges, particularly in extensive datasets [6]. This challenge is particularly pronounced within the legal domain, where input from legal professionals or experts is often indispensable for accurate assessments.

Acknowledgements: *We acknowledge the support of the IHUB-ANUBHUTI-IIITD FOUNDATION set up under the NM-ICPS scheme of the Department of Science and Technology, India. Thanking Rajiv Ratn Shah, who is partly supported by the Infosys Center for AI, the Center for Design and New Media, and the Center of Excellence in Healthcare at IIIT Delhi. We also want to thank Dr. Debanjan Mahata for motivating us and providing insights on the task[5].*

References

[1] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The muppets straight out of law school. arXiv preprint arXiv:201002559. 2020.

[2] Sugiyama K, Kumar T, Kan MY, Tripathi RC. Identifying citing sentences in research papers using supervised learning. In: 2010 International Conference on Information Retrieval & Knowledge Management (CAMP). IEEE; 2010. p. 67-72.
[3] Färber M, Thiemann A, Jatowt A. To cite, or not to cite? Detecting citation contexts in text. In: European conference on information retrieval. Springer; 2018. p. 598-603.
[4] Bonab H, Zamani H, Learned-Miller E, Allan J. Citation worthiness of sentences in scientific reports. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval; 2018. p. 1061-4.
[5] Gosangi R, Arora R, Gheisarieha M, Mahata D, Zhang H. On the Use of Context for Predicting Citation Worthiness of Sentences in Scholarly Articles. arXiv preprint arXiv:210408962. 2021.
[6] Wright D, Augenstein I. CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding. arXiv preprint arXiv:210510912. 2021.
[7] Zeng T, Acuna DE. Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models. Scientometrics. 2020;124(1):399-428.
[8] Roostaee M. Citation Worthiness Identification for Fine-Grained Citation Recommendation Systems. Iranian Journal of Science and Technology, Transactions of Electrical Engineering. 2022;46(2):353-65.
[9] Savelka J, Walker VR, Grabmair M, Ashley KD. Sentence boundary detection in adjudicatory decisions in the united states. Traitement automatique des langues. 2017;58:21.
[10] Sanchez G. Sentence boundary detection in legal text. In: Proceedings of the natural legal language processing workshop 2019; 2019. p. 31-8.
[11] Huang Z, Low C, Teng M, Zhang H, Ho DE, Krass MS, et al. Context-aware legal citation recommendation using deep learning. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law; 2021. p. 79-88.
[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.
[13] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.
[14] Lippi M, Pałka P, Contissa G, Lagioia F, Micklitz HW, Sartor G, et al. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law. 2019;27(2):117-39.
[15] Tuggener D, von Däniken P, Peetz T, Cieliebak M. LEDGAR: a large-scale multi-label corpus for text classification of legal provisions in contracts. In: 12th Language Resources and Evaluation Conference (LREC) 2020. European Language Resources Association; 2020. p. 1228-34.
[16] Zheng L, Guha N, Anderson BR, Henderson P, Ho DE. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. ICAIL '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 159–168. Available from: https://doi.org/10.1145/3462757.3466088.
[17] Atanasova P, Simonsen JG, Lioma C, Augenstein I. A Diagnostic Study of Explainability Techniques for Text Classification. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 3256-74. Available from: https://aclanthology.org/2020.emnlp-main.263.
[18] Kindermans PJ, Schütt K, Müller KR, Dähne S. Investigating the influence of noise and distractors on the interpretation of neural networks. arXiv preprint arXiv:161107270. 2016.
[19] Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:200410964. 2020.
[20] Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. arXiv preprint arXiv:200405150. 2020.
[21] Wright D, Augenstein I. Claim check-worthiness detection as positive unlabelled learning. arXiv preprint arXiv:200302736. 2020.
[22] Sadvilkar N, Neumann M. PySBD: Pragmatic Sentence Boundary Disambiguation. In: Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS). Online: Association for Computational Linguistics; 2020. p. 110-4. Available from: https://www.aclweb.org/anthology/2020.nlposs-1.15.