April 2022

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.

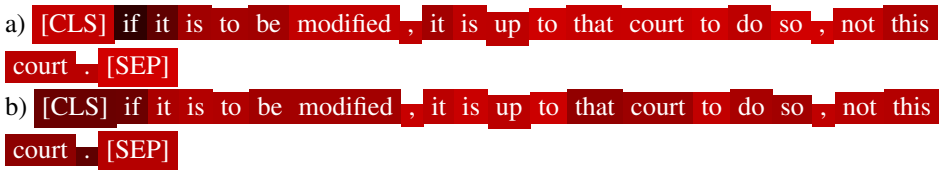

Figure 1. A cite-worthy sentence. a) LEGAL-BERT+PU classified it as citeworthy and b) LEGAL-BERT classified it as sentence non-citeworthy. Darker the color more the relative contribution.

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. 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.

Appendix

A. Citation Detection

We follow a multi-step methodology to detect all the citations present in the legal text.[6]
Regex to detect the boundary of versus type cases:

```
([A-Z][A-Za-z-']+|[A-Z]\.)(\s([A-Z]\. |of|and|&)|(?:\s[A-Z][A-Za-z-']*))*
```

The above regex was used both before and after the occurrence of 'v.' in order to identify both the parties involved in the case.
We replace all the citations detected with a placeholder *[CITATION_SPAN]*.

B. Sentence Boundary Detection

Table 9 gives toy examples after usage of different sentence boundary detectors on our dataset. Following are few examples of citations from the corpus:

1. *168 Pa. Superior Ct. 351, 77 A. 2d 706*
2. *State v. Camerlin, 117 R.I. 61, 362 A.2d 759 (1976)*
3. *Interstate Coal Co. v. Trivett, 155 Ky. 825, 160 S. W. 728*

List of acronyms/shortenings which caused incorrect sentence splitting. These were identified and replaced with their version which didn't contain a full stop.

- Inc.
- Co.
- Ltd.
- No.
- Vol.
- Corp.
- Viz.
- Mfg.
- Dist.
- Commn.
- Sec.
- Pet.
- Com.
- Eq.
- Doc.
- Ins.
- Ex.
- Cf.
- Civ.
- a.m.
- p.m.
- e.g.
- Pvt.
- Ms.
- Mr.
- Jr.
- Sr.
- Dr.
- Al.
- A.
- Q.
- Cont.
- Aff.
- Cert.
- Art.
- Bros.
- Ref.
- Mrs.
- Ed.
- Nom.
- Ch.
- Eq.
- D.C.
- i.e.

Apart from these, several instances of multiple consecutive punctuation marks were also fixed.
Once the individual sentences were split, the following regex was used to detect if a sentence was a citation in itself: `^(See)?(\s)?(eg)?(\s)?(\[CITATION\_SPAN\]\s?)+$`

[6]Code for citation detection has been submitted in the supplementary material.

C. Data Visualization

Our data was divided into 61 different jurisdictions according to data present in https://cite.case.law/. For visualizing the data, we used legal cases from the following jurisdictions:

- Louisiana (la)
- Illinois (ill)
- Arkansas (ark)
- Massachusetts (mass)
- Wisconsin (wis)

For visualizing the data, TSNE plots were made. The hyper parameter perplexity was set at 2,000 for the plot corresponding to state wise division and it was set at 200 for the plot corresponding to the century wise division.

Further details of the TSNE plots can be found at: `https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html`.

We also visualize randomly sampled court cases across different jurisdictions to analyze any underlying patterns. We randomly sampled 1,500 cases from five randomly selected jurisdictions. Using LEGAL-BERT, we extract embeddings of each case and generated the TSNE plot of these embeddings. Figure 2 shows the distribution of the data belonging to the individual jurisdictions implying that the results of the proceedings were independent of the jurisdictions.

Further, we plotted 2,500 cases randomly from every jurisdiction ranging from the 19th to 21st century. Figure 3 shows a toy example of *'illinois'* jurisdiction. These clusters indicate a gradual change in a court case's writing style.

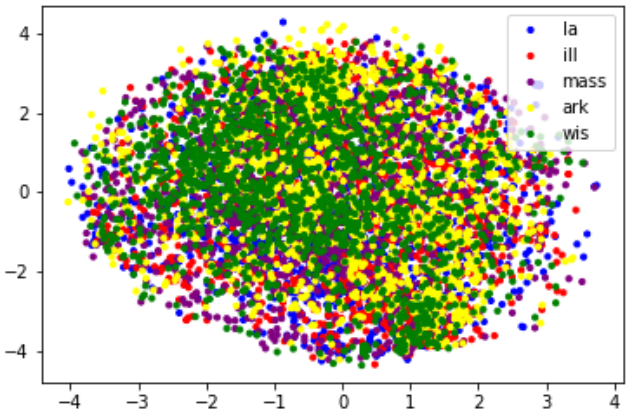

Figure 2. Visualizing LEGAL-BERT Embeddings for the dataset for five jurisdictions which shows that there are no observable differences between any two different jurisdiction. The jurisdictions chosen for this were *Louisiana ('la'), Illinois ('ill'), Massachusetts ('mass'), Arkansas ('ark') and Wisconsin ('wis').*

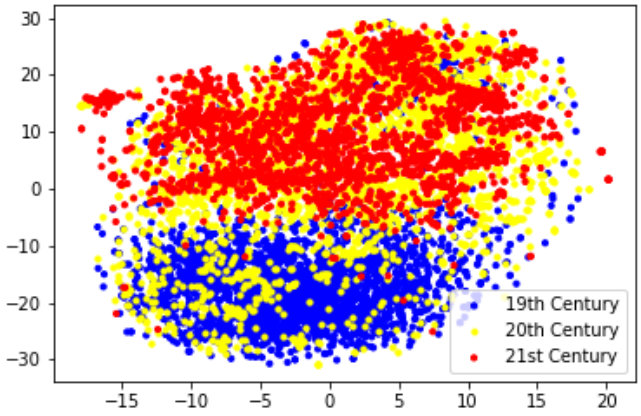

Figure 3. Visualizing LEGAL-BERT Embeddings for the 'ill' jusrisdiction over three centuries i.e. 19th, 20th and 21st, which shows the gradual shift in the writing style of the legal documents along the time period.

D. Experimental Setup

D.1. Models

• **Logistic Regression** This is a simple baseline with TF-IDF as input features.

• **CRNN** A Convolutional Recurrent Neural Network in similar architecture as [3] with little modifications.

• **Transformer** A transformer [12] is trained from scratch.

• **Longformer** A transformer-based model designed to handle longer sequences and uses a sparse attention mechanism introduced by [20].

• **BERT** A popular model with a strong reputation due to its performance on various tasks, we selected BERT [13] for our classification task.

• **LEGAL-BERT** A member of the family of BERT model pre-trained on large legal corpora spanning across different countries. Developed by [1] for legal domain.

• **LEGAL-BERT+PU** The intention is to reduce subjectivity involved while adding a citation to a sentence. Figure **??** shows the working diagram of positive unlabeled learning. Previously, PU learning has shown promising results in rumor detection on Twitter and citation needed detection in Wikipedia [21]. The basis of PU learning is to suppose that positive, i.e., cite-worthy data is labeled, and non-citation-worthy data is unlabelled. A classifier is trained on the positive and unlabeled data to estimate that a given sample is labeled. Using the classifier, we estimate whether a sample is positive, given its unlabelled. We then combine positive samples with one copy of unlabelled samples marked as positive and the other as negative. The unlabelled samples are then weighed by the first classifier's estimate of the probability of the sample being positive. Finally, a classification model is trained on the task of citation-worthiness.

D.2. Infrastructure

A system with 48 cores with muliple GPUs and $\sim$ 500GB (not even 40% utilised) RAM was used in the experimentation. All training were done using GeForce RTX 3090 with a memory of 24,268 MB (24GB).

D.3. Hyperparameter Tuning

Here, except for logistic regression, we used Ax search. It can be found here: `https://docs.ray.io/en/latest/tune/api_docs/suggestion.html#ax-tune-suggest-ax-axsearch`
For logistic we used sklearn's built in *RandomizedSearchCV*. See Table 3

Model	LR (x10^{-6})	Decay	Warmup Steps	Epochs
BERT	9.8276	0.01	1,000	8
LEGAL-BERT	9.6984	0.1	300	11
LEGAL-BERT + PU	6.6823	0.0	400	3
Longformer	7.2936	0.1	2,000	9

Table 3. Selected hyperparameters for different models

D.3.1. Logistic Regression

For Logistic regression, a search space of *spacy.stats.uniform(loc=0, scale=4)* was taken with L1 and L2 regularization. Selected parameters were C: 0.1151395399 and regularization: L2.

Further documentation for *uniform* function can be found at: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.uniform.html`.

D.3.2. CRNN

For CRNNs the search space for learning rate, epoch and batch size is [1e-3, 1e-2], [3, 15], {4,8,32,128} respectively. The selected parameters as learning rate: 0.00523737; epochs: 3 and batch size: 32.

D.3.3. Tranformer

For transformers a search space for learning rate, weight decay, warmup steps, epochs, feed forward layers, number heads, epochs and dropouts are: [1e-7, 1e-3], {0.0, 0.0001, 0.001, 0.01, 0.1}, {0, 100, 200, 300, 400, 500, 1,000, 1,500, 2,000, 2,500, 5,000}, [3, 35], {128, 256, 512, 1,024, 2,048}, {1,2,3,4,5,6,10,12}, {0.0,0.1,0.2,0.3,0.4,0.5} respectively. The selected values in the mentioned order are: 0.000174364, 0.1, 14, 128, 5, 4, 0.5.

D.3.4. Xformer

The learning rate was tuned in the range [1e-7, 1e-4], while with BERT, the rate is in the range [1e-8, 1e-5]. We used a triangular learning rate. Search space for Weight decay, Warmup steps and epochs are {0.0, 0.0001, 0.001, 0.01, 0.1}, {0, 100, 200, 300, 400, 500, 1,000, 1,500, 2,000, 2,500, 5,000 }, [3, 15] respectively. The batch size was taken as 4. The selected parameters for the models are listed in Table 3.

D.4. Transfer learning

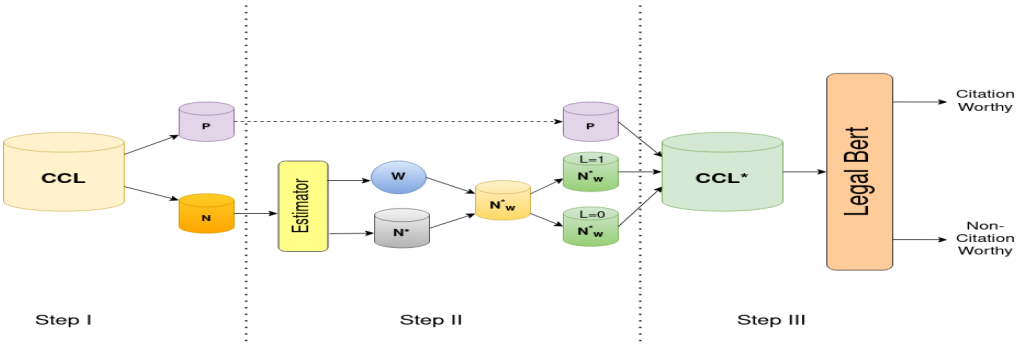

Figure 4. In Step I, the CiteLegal dataset (CCL) is categorized into positive (P) and negative (N) samples. The positive samples remain unchanged, while the negative samples are processed through an estimator to determine their weights. Each sample is then multiplied by its respective weight $(N_{w[L=1]})$. To create samples labeled as citation unworthy, we duplicate these weighted samples $(N_{w[L=0]})$ in Step II. These two sets are combined to form the augmented CiteLegal dataset (CCL^*), which is subsequently used for label prediction in Step III.

UNFAIR-ToS [[14]] comprises 50 Terms of Service (ToS) from online platforms like YouTube, eBay, and Facebook. The dataset annotates eight categories of unfair contractual terms or sentences that could violate user rights under EU consumer laws. Each sentence is input for the model, which outputs any unfair categories present.

LEDGAR (Labeled EDGAR) [[15]], introduced in 2020, is a dataset for classifying contract provisions (paragraphs). Contract terms are sourced from US Securities and Exchange Commission (SEC) filings available to the public through EDGAR10. The original dataset contains around 850k contract clauses categorized into 12.5k topics. It's a single-label multi-class classification task where each label represents the primary topic of the related contract clause.

E. Metrices

We used sklearn's *precision_recall_fscore_support* for the following metrices

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN}$$

Here T stands for True, F for False, P for positives and N for negatives. hence TP stands for true positives and so on.

Further documentation can be found at: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html`.

E.0.1. Macro F1

Macro F1 is the average of F1 scores of all the classes.

E.0.2. Micro F1

Micro F1 is the weighted sum of F1 scores of all the classes where weights are the class distribution in the dataset.

F. Dataset

To address **RQ1**, we employed the Caselaw Access Project (CAP), an extensive repository encompassing American legal cases from both federal and territorial courts across all US states. These legal cases are meticulously categorized into 61 jurisdictions, each containing essential information about the presiding judge, reporter, court of jurisdiction, and cited cases. Notably, older cases often exist in the form of document scans, which CAP has diligently converted into a digital format through Optical Character Recognition (OCR). Additionally, CAP provides an OCR confidence score for these digitized cases, enhancing the reliability of the data.

We extracted our dataset, denoted as version 3, from CAP. This dataset, retrieved on September 21, 2021, comprises a staggering 7 million documents spanning the period from 1600 to 2021. Comprehensive details regarding this dataset can be found in Section C of the Appendix.

F.1. Data Preprocessing and Sentence Boundary Detection

Legal documents are organized into distinct sections, with the 'Opinion' section containing the case transcript. In this section, irrelevant excerpts from other legal cases and their citations are meticulously removed to maintain the continuity of sentences. To ensure data quality, we addressed various sources of noise, such as footnotes, page numbers, and Non-ASCII tokens, by applying specific regular expression (regex) patterns. Our data preprocessing methodology was refined through an in-depth analysis of 500 documents spanning diverse jurisdictions and centuries. Subsequently, this methodology was rigorously validated using a random selection of 500 documents, ensuring that all 1,000 documents in the dataset were devoid of noise.

Accurately delineating sentence boundaries within legal documents presents a complex challenge. Standard natural language processing libraries like SpaCy, NLTK, and

SegTok [7] lack domain-specific boundary handling capabilities. To enhance sentence boundary detection, we incorporated rules from [22], addressing specific tokens that caused issues in sentence splitting. We curated a list of problematic tokens and replaced them with appropriate alternatives, ensuring accurate boundary identification. For illustrative examples, please refer to the relevant section in the Appendix (B).

The performance of sentence splitters in legal text was evaluated and recorded in Table 9. A meticulous analysis of 30 documents (comprising 1,800 sentences) revealed that only pySBD achieved an accuracy exceeding 50%. This finding was subsequently validated with a larger sample of 50 documents (consisting of 2,700 sentences), where all but four sentences were accurately split, affirming the effectiveness of our chosen approach.

F.2. Citation Detection

A citation occurred in the following format in the case documents which was detected using our regex
Part A vs Party B <volume> <Reporter> <Page Numbers> <Year/Other metadata>
After detecting the citations, we divided sentences into four following types to make our dataset.

- **Type 1:** A sentence that does not contain in-line citations and is followed by a sentence of the same type. Such sentences are labeled '**0**'.
- **Type 2:** A sentence that does not contain in-line citations but is followed by a sentence containing in-line citations. Such sentences are ignored and not included in our dataset as we cannot completely classify them as citation-worthy or not.
- **Type 3:** A sentence that contain in-line citations. Such sentences are ignored and not included in our dataset as removing citations from them may lead to the incorrect grammatical structure of the sentence.
- **Type 4:** A sentence that does not contain in-line citations and is followed by a sentence that is a citation in itself. Such sentences are labeled '**1**'.

F.3. Dataset Profiling

Our final dataset, available at, comprises 178 million sentences presented in three versions: small, medium, and large (original). A comprehensive overview of the dataset statistics is provided in Table 7, and Table 8 delineates the sizes of the dataset versions.

To validate the quality of the dataset, we manually examined 1,000 randomly selected sentences based on two criteria: the accuracy of the citation splitting mechanism and the precision of citation detection. Among these sentences, only eleven were inaccurately split, either forming multiple sentences or being partially split. Out of these, three instances were attributed to Optical Character Recognition (OCR) inconsistencies, resulting in an accuracy rate of 98.9%. After excluding these cases, the accuracy of the splitting mechanism rose to 99.2%. Furthermore, the labels assigned to each sentence were verified to be correct.

[7] https://github.com/fnl/segtok

(1)	**[ORIGINAL]** On appeal to this Court, we held that the railroad had acquired by condemnation proceedings a base or conditional fee, terminable on the cesser of the use for railroad purposes. Lacy v. East Broad Top Railroad and Coal Co., 168 Pa. Superior Ct. 351, 77 A. 2d 706. **[PROCESSED]** On appeal to this Court, we held that the railroad had acquired by condemnation proceedings a base or conditional fee, terminable on the cesser of the use for railroad purposes.
(2)	**[ORIGINAL]** In Tanorio v. Superior Court, 1 N.Mar.I. 4, we determined under what conditions a writ of mandamus may issue. **[PROCESSED] Ignored**. (citation present at the start of the sentence.)

Table 4. Excerpt from training samples in CiteLegal. The first example belongs to Type 4 sentences whereas the second example belongs to Type 3 sentences.

G. Dataset and statistics

We present our final dataset in jsonl format where each sentence is an object having the following parameters:

G.1. Meta-data

- **File Name:** The case file to which the sentence belongs.
- **Sentence Number:** The sentence number as present in the document.
- **Sentence:** The naturally occurring sentence in the text (after preprocessing/removing citation span.)
- **Label:** Integer value of '**0**' or '**1**'. '**0**' represents that the sentence is not citation-worthy whereas '**1**' represents that the sentence is citation-worthy.

G.2. Data Pre-processing

We handled acronyms, like 'article' instead of 'art.', 'section' instead of 'sec.', 'number' instead of 'no.' and so on. Some of these acronyms/shortenings are commonly used for e.g. 'no.', 'i.e.', 'ms.' whereas some were legal jargon like 'cf.', 'D.C.', 'Inc.'. We also identified different bodies, laws, and sections whose names contained '.' and were commonly referred to in the case laws of American legal corpus. After including the aforementioned steps, we use the pySBD module to prepare the final dataset. An example is given in Table 4.

In all the dataset versions, we have followed nearly an 80:10:10 split for train, validation, and test sets respectively. The split is document level which means that all the sentences belonging to the same document will only be present in one of the train, validation or test splits.

H. Examples

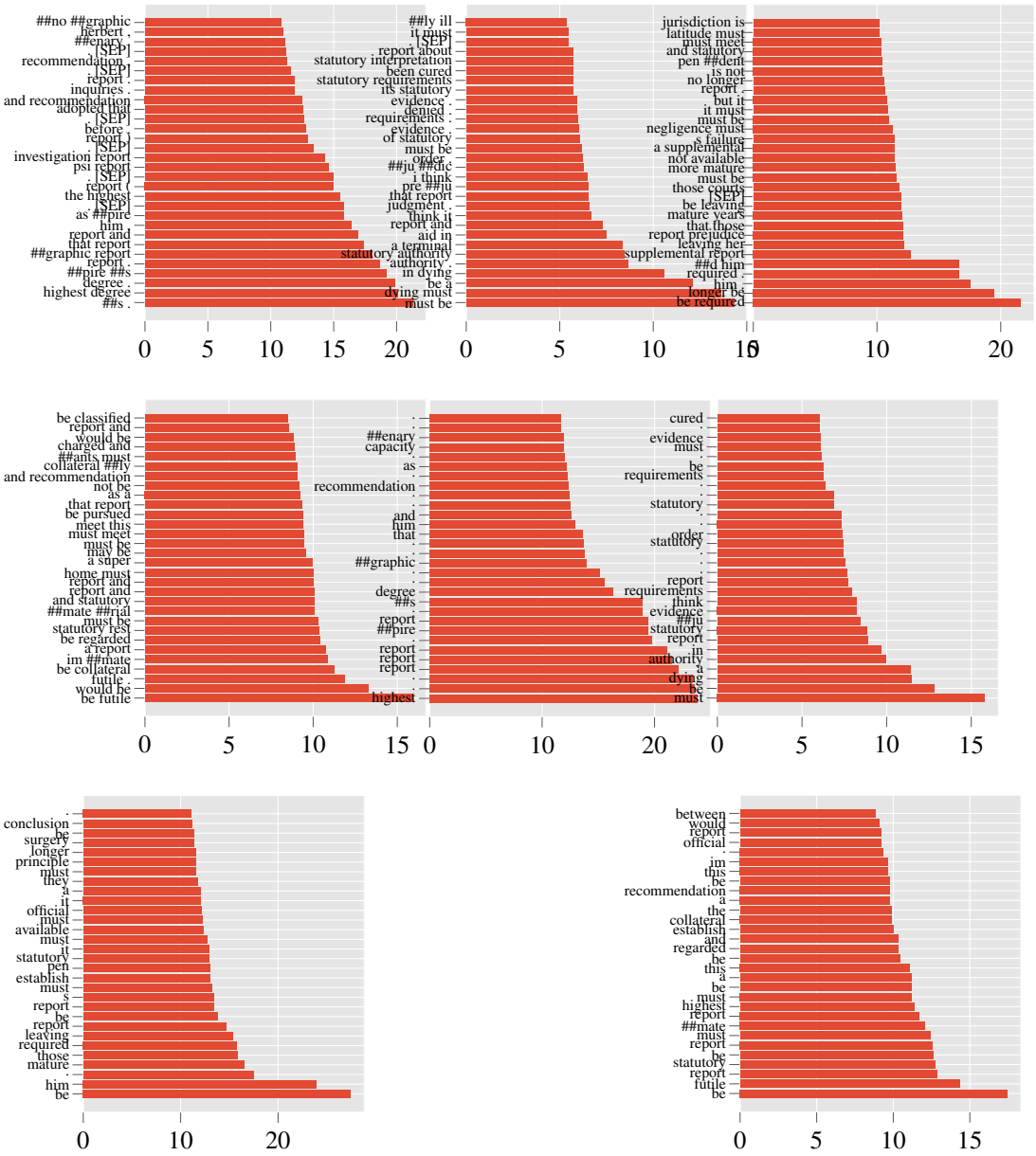

Figure 5. Bar graph showing most ”important” words for a) non-citeworthy class in LEGAL-BERT b) citeworthy class in LEGAL-BERT c) non-citeworthy class in LEGAL-BERT+pu d) citeworthy class in LEGAL-BERT+pu

States	μ-P	μ-R	μ-F1	m-P	m-R	m-F1
ala_xml_20210921	0.90	0.90	0.90	0.90	0.90	0.90
alaska_xml_20210921	1.00	1.00	1.00	1.00	1.00	1.00
am-samoa_xml_20210921	1.00	1.00	1.00	1.00	1.00	1.00
ariz_xml_20210921	0.87	0.87	0.87	0.90	0.86	0.87
ark_xml_20210921	0.89	0.89	0.89	0.90	0.89	0.89
cal_xml_20210921	0.75	0.75	0.75	0.80	0.80	0.75
colo_xml_20210921	0.95	0.95	0.95	0.95	0.95	0.95
conn_xml_20210921	0.80	0.80	0.80	0.80	0.79	0.79
dc_xml_20210921	0.82	0.82	0.82	0.82	0.81	0.81
del_xml_20210921	1.00	1.00	1.00	1.00	1.00	1.00
fla_xml_20210921	0.87	0.87	0.87	0.87	0.87	0.87
ga_xml_20210921	0.95	0.95	0.95	0.95	0.95	0.95
haw_xml_20210921	0.92	0.92	0.92	0.94	0.88	0.90
idaho_xml_20210921	0.91	0.91	0.91	0.91	0.93	0.91
ill_xml_20210921	0.88	0.88	0.88	0.88	0.87	0.87
ind_xml_20210921	0.88	0.88	0.88	0.87	0.89	0.88
iowa_xml_20210921	0.85	0.85	0.85	0.85	0.85	0.85
kan_xml_20210921	0.93	0.93	0.93	0.92	0.95	0.93
ky_xml_20210921	0.89	0.89	0.89	0.89	0.89	0.89
la_xml_20210921	0.87	0.87	0.87	0.87	0.87	0.87
mass_xml_20210921	0.89	0.89	0.89	0.89	0.89	0.89
me_xml_20210921	0.85	0.85	0.85	0.90	0.80	0.82
mich_xml_20210921	0.93	0.93	0.93	0.92	0.93	0.93
minn_xml_20210921	0.93	0.93	0.93	0.93	0.94	0.93
miss_xml_20210921	0.78	0.78	0.78	0.78	0.78	0.78
mo_xml_20210921	0.89	0.89	0.89	0.89	0.88	0.89
mont_xml_20210921	0.95	0.95	0.95	0.92	0.96	0.94
navajo-nation_xml_20210921	1.00	1.00	1.00	1.00	1.00	1.00
nc_xml_20210921	0.84	0.84	0.84	0.84	0.84	0.84
nd_xml_20210921	0.93	0.93	0.93	0.94	0.94	0.93
neb_xml_20210921	0.79	0.79	0.79	0.79	0.79	0.79
nev_xml_20210921	1.00	1.00	1.00	1.00	1.00	1.00
nj_xml_20210921	0.88	0.88	0.88	0.88	0.88	0.88
nm_xml_20210921	0.83	0.83	0.83	0.81	0.83	0.82
ny_xml_20210921	0.83	0.83	0.83	0.82	0.82	0.82
ohio_xml_20210921	0.89	0.89	0.89	0.89	0.89	0.89
okla_xml_20210921	1.00	1.00	1.00	1.00	1.00	1.00

Table 5. A Table showing results on different states in the US. μ is micro and m is macro. These are the state abbreviations based present under state column. You can visit it to see the expansions of these states on `https://cite.case.law`. Results are on the split they provided while bulk downloading their corpus

States (cont)	μ-P (cont)	μ-R (cont)	μ-F1 (cont)	m-P (cont)	m-R (cont)	m-F1 (cont)
or_xml_20210921	0.87	0.87	0.87	0.87	0.87	0.87
pa_xml_20210921	0.88	0.88	0.88	0.89	0.88	0.88
pr_xml_20210921	0.83	0.83	0.83	0.83	0.83	0.83
ri_xml_20210921	0.75	0.75	0.75	0.75	0.75	0.75
sc_xml_20210921	0.92	0.92	0.92	0.92	0.92	0.92
sd_xml_20210921	0.83	0.83	0.83	0.83	0.83	0.83
tenn_xml_20210921	0.80	0.80	0.80	0.80	0.78	0.78
tex_xml_20210921	0.90	0.90	0.90	0.90	0.90	0.90
tribal_xml_20210921	1.00	1.00	1.00	1.00	1.00	1.00
us_xml_20210921	0.86	0.86	0.86	0.86	0.86	0.86
utah_xml_20210921	0.83	0.83	0.83	0.85	0.86	0.83
va_xml_20210921	0.82	0.82	0.82	0.82	0.82	0.82
vi_xml_20210921	1.00	1.00	1.00	1.00	1.00	1.00
vt_xml_20210921	0.90	0.90	0.90	0.90	0.90	0.90
w-va_xml_20210921	0.87	0.87	0.87	0.89	0.88	0.87
wash_xml_20210921	0.80	0.80	0.80	0.79	0.80	0.79
wis_xml_20210921	0.88	0.88	0.88	0.87	0.89	0.87
wyo_xml_20210921	0.90	0.90	0.90	0.93	0.88	0.89

Table 6. A Table (continued) showing results on different states in the US. μ is micro and m is macro.

Metric	#
Total Files (train— test— validation)	**5,548,618** (4,434,179— 556,898 — 557,541)
Total Sentences (train— test— validation)	**178,459,203** (142,588,927 —17,935,336 —17,934,940)
Total citation-worthy sentence	10,487,177
Total non-citation-worthy sentences	167,972,026
Avg character length of citation-worthy Sentences	171.61
Avg character length of non-citation-worthy Sentences	172.93
Avg. number of sentences extracted per document	32.16

Table 7. Basic Statistics on the CiteLegal dataset. The table conatins details of dataset splits in terms of number of sentences, number of files, number of sentences belonging each class and average number of characters per sentence.

Dataset Version	Total Sentence Count	Citation-Worthy Sentences
Large	178,459,203	10,487,177 (∼5.87%)
Medium	10,000,000	586,999 (∼5.869%)
Small	1,000,000	58,909 (∼5.89%)

Table 8. Statistics of the different versions of the CiteLegal dataset.

Original	The copy of the hospital record, being a photostat, was admissible under Code (1427), Art. 4335, sec. 3459, and was produced by Mr. Alex, who was in charge at the time.
simple period split (.)	• The copy of the hospital record, being a photostat, was admissible under Code(1427), Art • 4335, sec • 3459, and was produced by Mr • Alex, who was in charge at the time
SegTok	• The copy of the hospital record, being a photostat, was admissible under Code (1427), Art. • 4335, sec. • 3459, and was produced by Mr. • Alex, who was in charge at the time.
Spacy blackstone	• The copy of the hospital record, being a photostat, was admissible under Code (1427), Art. • 4335, sec. • 3459, and was produced by Mr. • Alex, who was in charge at the time.
pySBD	• The copy of the hospital record, being a photostat, was admissible under Code (1427), Art. 4335, section 3459, and was produced by Mr. Alex, who was in charge at the time.

Table 9. Examples from the dataset showing performance of different segmenters for unstructured legal text. New line represents a segmented sentence. The presence of abbreviated/short terms in the legal text makes it difficult for segmenters to decide the point of segmentation. PySBD segmenter resulted in better segmentation compared to others like SegTok, spacy etc.

Type	Example
I	This statute applies to alimony obligations created by verdict. However, while an exemption should be strictly construed, the construction must still be reasonable
II	This is the ground upon which I am going to decide this case. The case does not come within the statute of 5 & 6 Will., 4, C. 65, because notice in writing was not given to two Justices under Sec. 5 of that act.
III	The case does not come within the statute of 5 & 6 Will., 4, C. 65, because notice in writing was not given to two Justices under Sec. 5 of that act.
IV	This statute applies to alimony obligations created by verdict. See Allen v. Allen,265 Ga.

Table 10. Table describes the sentence types found in the corpus before extraction of citation-worthy sentences.

Example	Sentence	Label
E1	This statute applies to alimony obligations created by verdict. **See Allen v. Allen, 265 Ga. 53 (1) (452 SE2d 767) (1995)**	cite
E2	However, while an exemption should be strictly construed, the construction must still be reasonable. **Trustees of Ind. Univ. v. Town of Rhine, 170 Wis. 2d 293, 299, 488 N.W.2d 128 (Ct. App. 1992).**	cite
E3	This leaves absolutely indefinite and uncertain what the plaintiff was to receive.	not-cite
E4	The appellant then was granted the right and did file amendments to its assignments of error.	not-cite

Table 11. Example sentences for "cite" and "not_cite" labels from CiteLegal. The bold sentence, which is removed from our dataset, confirms the claim for the previous sentence to be cite-worthy.