A Simple but Effective Context Retrieval for Sequential Sentence Classification in Long Legal Documents

Anas Belfathi, Nicolas Hernandez, Laura Monceaux, Richard Dufour
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000, France
firstname.lastname@univ-nantes.fr

Abstract

Sequential sentence classification extends traditional classification, especially useful when dealing with long documents. However, state-of-the-art approaches face two major challenges: pre-trained language models struggle with input-length constraints, while proposed hierarchical models often introduce irrelevant content. To address these limitations, we propose a simple and effective document-level retrieval approach that extracts only the most relevant context. Specifically, we introduce two heuristic strategies: **Sequential**, which captures local information, and **Selective**, which retrieves the semantically similar sentences. Experiments on legal domain datasets show that both heuristics lead to consistent improvements over the baseline, with an average increase of $\sim$5.5 weighted-F1 points. Sequential heuristics outperform hierarchical models on two out of three datasets, with gains of up to $\sim$1.5, demonstrating the benefits of targeted context.

1 Introduction

Sequential sentence classification (SSC) is the task of classifying each sentence based on its semantic role within a document. Since a sentence's meaning is often shaped by its surrounding context, SSC is particularly useful in structured texts such as legal cases. Identifying key rhetorical components (e.g., preamble, issue, or analysis; see Figure 1) benefits downstream tasks such as information retrieval (Neves et al., 2019; Safder and Hassan, 2019) and document summarization (Kalamkar et al., 2022; Muhammed et al., 2024).

Recent SSC approaches rely on hierarchical models that process full-document sequences to capture broader context (Jin and Szolovits, 2018; Brack et al., 2021; Kalamkar et al., 2022). However, processing all sentences is not always beneficial, as it may introduce noise from irrelevant content (Shi et al., 2023). This issue is compounded by the fact that pre-trained language models (PLMs) remain constrained by input-length limitations (Warner et al., 2024), even with advances in large language models (LLMs) (BehnamGhader et al., 2024). Overcoming these limitations by retrieving only the most relevant context offers a way to both reduce noise and improve the efficiency of SSC models, particularly when using PLMs.

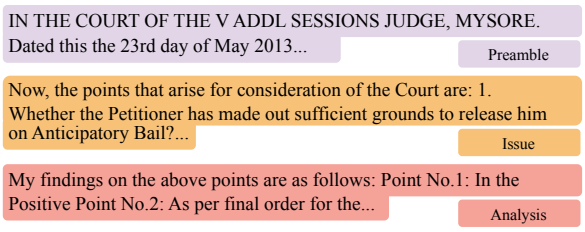


Figure 1: A segment of a legal document with sentences labeled by their function.

Several studies have begun exploring strategies for retrieving relevant informations (Amalvy et al., 2023; Lan et al., 2024). However, to our knowledge, no prior work has specifically addressed how to retrieve the most relevant sentence-level context to optimize PLMs performance for the SSC task. We focus only on encoder-only models, which currently combine effectiveness with low computational cost compared to LLMs for classification tasks (Roccabruna et al., 2024).

In this paper, our contributions are twofold: (1) we analyze the role of context in SSC by introducing two heuristic retrieval strategies—*Sequential*, which assumes that the most informative context lies in positional proximity, and *Selective*, which retrieves semantically similar sentences regardless of their position in the document; and (2) we demonstrate that these strategies enhance PLM performance by providing more relevant context and can outperform state-of-the-art hierarchical models.

We evaluate on document-level datasets in the legal domain, the primary benchmark for SSC task. To foster transparency and reproducibility, we release our code under an open-source license[1].

[1] https://github.com/AnasBelfathi/ACL-2025

Proceedings of the 12th Argument Mining Workshop, pages 160–167

2 Related Work

2.1 Input Sequence Constraints in PLMs

Encoder-only models such as BERT (Devlin et al., 2019) offer a strong tradeoff between size and performance, making them a compelling alternative to larger generative models for classification tasks. However, the quadratic complexity of self-attention in vanilla Transformer models limits their effective input length, posing challenges for processing long documents. To mitigate this, sparse attention mechanisms have been introduced to reduce computational costs (Zaheer et al., 2020; Wang et al., 2020; Beltagy et al., 2020). While these methods extend the range of accessible context, they still struggle to effectively aggregate the task-relevant information needed for fine-grained sentence classification in long-document settings (Warner et al., 2024; Nussbaum et al., 2025).

2.2 SSC for Long Documents

Early work on SSC focused on hierarchical models to incorporate broader context into sentence representations. Hierarchical Sequential Labeling Network (HSLN) was among the first to process full-document sequences for contextualized representations (Jin and Szolovits, 2018; Brack et al., 2021; Kalamkar et al., 2022). More recent studies have explored refined learning strategies: T.y.s.s. et al. (2024) applied contrastive and prototypical learning to enhance sentence representations by leveraging semantic similarities, while Santosh et al. (2024) introduced a hierarchical curriculum learning framework to progressively improve the model's ability to distinguish rhetorical labels at different levels of granularity.

While these studies have primarily focused on improving HSLN, our work addresses a different challenge: overcoming PLM input-length constraints by retrieving only the relevant context, thus reducing noise and improving efficiency in SSC.

3 Context Retrieval

We propose a simple yet effective set of heuristics to enhance SSC in long documents. The motivation for introducing the two types—**Sequential** and **Selective**—is to explore complementary definitions of contextual relevance. Sequential heuristics are based on the assumption that the most useful context comes from nearby sentences, leveraging **positional proximity**. In contrast, Selective heuristics test whether **semantically similar** sentences, regardless of their position, provide better context, particularly in long structured documents.

Sequential Heuristics extract context from sentences adjacent to the target sentence within the same document. We consider three widely adopted strategies:

- **Before**: Selects the k sentences immediately preceding the target sentence.
- **After**: Selects the k sentences immediately following the target sentence.
- **Surrounding**: Selects $\frac{k}{2}$ sentences before and after the target sentence.

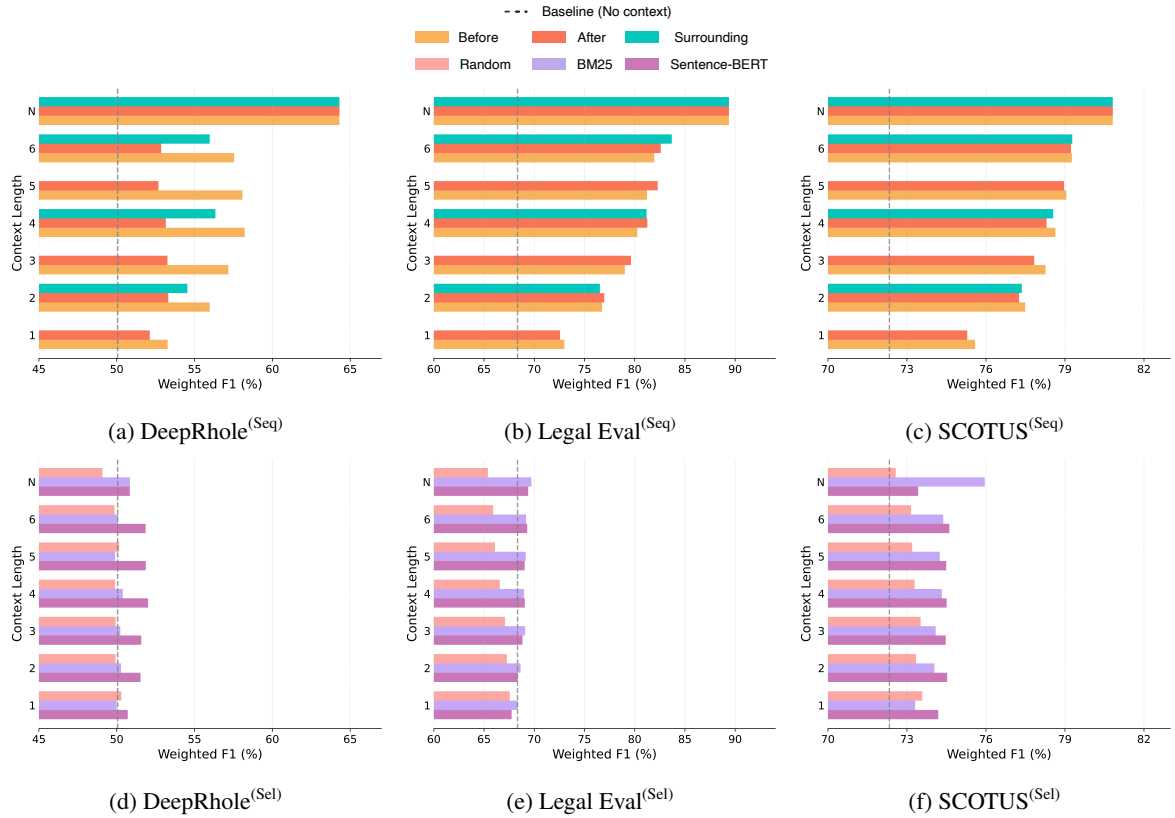
Selective Heuristics , unlike sequential strategies, retrieve sentences from anywhere in the document, independent of their position relative to the target sentence. We explore three selection techniques:

- **Random**: Randomly selects k sentences from the entire document.
- **BM25**: Retrieves the k most relevant sentences using a ranking function based on term frequency-inverse document frequency (TF-IDF) weighting (Trotman et al., 2014), widely used in information retrieval for lexical relevance scoring.
- **Sentence-BERT**: Selects the k semantically closest sentences to the target sentence using embeddings that capture sentence-level similarity via a siamese BERT network (Reimers and Gurevych, 2019).

Given computational constraints, we limit our analysis to $k = 6$. Table 2 in the Appendix provides illustrative examples.

Sentence Ordering We further investigate whether the order of retrieved sentences impacts SSC performance. Inspired by NAREOR (Gangal et al., 2022), which explores sentence reordering to analyze narrative coherence in storytelling, we examine whether maintaining full document sentences $(k = N)$ while altering their order affects performance.

To evaluate this, we use our heuristics. In Sequential, we retain the original human-written order to preserve logical flow. In Selective, we reorder sentences based on their relevance to the target sentence while ensuring that all remain included for a fair comparison.



(a) DeepRhole$^{(Seq)}$ (b) Legal Eval$^{(Seq)}$ (c) SCOTUS$^{(Seq)}$

(d) DeepRhole$^{(Sel)}$ (e) Legal Eval$^{(Sel)}$ (f) SCOTUS$^{(Sel)}$

Figure 2: Weighted F1 scores for different context lengths k across three datasets. The top row (a, b, c) presents results using Sequential context$^{(Seq)}$, while the bottom row (d, e, f) represents Selective context$^{(Sel)}$. $k = N$ indicates that the full document is used to address the sentence ordering question. We set k as an even number for Surrounding heuristic to ensure comparability in context length with other ones.

4 Experimental protocol

4.1 Datasets

Our experiments focus on the legal domain, as it is the only domain with datasets annotated at the document level in english. We utilize three datasets: (i) DeepRhole (Bhattacharya et al., 2023), (ii) LegalEval (Kalamkar et al., 2022), and (iii) SCOTUS (Lavissière and Bonnard, 2024), derived from Indian and U.S. legal judgments. DeepRhole contains 7 rhetorical role labels, while the others have 13 each. For evaluation, we report the weighted F1-score[2].

4.2 SSC Model for Context Analysis

To ensure that our analysis covers all sentences in a document, we build upon the hierarchical HSLN model (Brack et al., 2021), with two minor modifications: (1) Motivated by ablation studies (Jin and Szolovits, 2018; Chen et al., 2023), which identified the contextual sentence enrichment layer as HSLN's primary driver of effectiveness, we removed the conditional random field (CRF) layer, and (2) We optimize only over the target sentence, enriched with context selected by our heuristics.

Further architectural details, including our refinements, are provided in Appendix A. All results are averaged over three runs for robustness.

5 Results

5.1 Context Analysis

Figure 2 shows that Sequential Heuristics systematically improve classification as more sentences are included. In LegalEval and SCOTUS, the *Surrounding* heuristic achieves the highest F1 score (83.6% and 79.2% at $k = 6$, respectively). This indicates that rhetorical signals are distributed in both directions, and that accessing context from both sides helps to more accurately situate the current sentence within its transitional flow. However, in DeepRhole, the *Before* heuristic performs best, reaching 58.2%. This suggests that this dataset follows a progressive narrative and argumentative thread, where the meaning of each sentence is fun-

[2]All datasets were split at the document level into 80% training, 10% validation, and 10% test sets.

Model	Seq	DeepRhole	Legal Eval	SCOTUS
BERT (baseline)	512	52.23	69.74	75.58
+ Before		**67.18**†	78.41†	79.74†
+ After		56.72†	**79.74**†	**81.34**†
+ Surrounding		62.87†	77.27†	75.47
+ Random		46.86	67.05	74.70
+ BM25		51.59	69.43	75.96
+ Sentence-BERT		52.23	68.98	76.24
Nomic-BERT (baseline)	2048	50.32	68.90	75.50
+ Before		**67.89**†	80.54†	81.12†
+ After		57.75†	**81.11**†	**81.32**†
+ Surrounding		65.51†	78.20†	80.81†
+ Random		51.61	68.43	75.73
+ BM25		53.90	70.82‡	77.06†
+ Sentence-BERT		54.02‡	70.76‡	77.17‡
BERT-HSLN (SOTA)	$512 \times N$	54.45	93.06	79.66

Table 1: Performance of PLMs using the best configuration observed in context analysis for $k \leq 6$ for each heuristic. Bold values represent the best improvement over the baseline (w/o context), while underlined values indicate the second-best. BERT-HSLN is the SOTA for the SSC task. Markers † and ‡ denote statistical significance over the baseline at $p = 0.05$ and $p = 0.01$, respectively.

damentally built upon what has been previously developed.

In contrast, Selective Heuristics yield marginal gains, with *BM25* being the most effective, reaching $\approx 74\%$ F1 in SCOTUS when $k \leq 6$.
The limited effectiveness of those heuristics could be attributed to two factors: (1) When documents lack semantically similar sentences, heuristics retrieve unrelated ones, adding noise (as observed in DeepRhole), and (2) The rhetorical function of a sentence often depends on its placement within the overall argumentative structure, rather than on its intrinsic semantics alone.

At $k = N$, the Sentence Ordering experiment confirms that SSC is sensitive to how context is structured—with the highest scores observed when the document's logical flow is preserved. Conversely, reordering sentences using Selective heuristics suggests that taking the full document may not be necessary; instead, prioritizing only the most relevant ones yields competitive performance.

5.2 Context Enrichment for PLMs

To examine how PLMs benefit from contextual enrichment[3], we conduct experiments with BERT (Devlin et al., 2019) and the recently introduced model Nomic-BERT (Nussbaum et al., 2025), as shown in Table 1.

[3]Context sentences were integrated with the target sentence into the PLM input while maintaining the natural human order for sequential heuristics.

Our results indicate that Sequential heuristics typically yield the largest improvements, significantly outperforming both the no-context baseline and state-of-the-art BERT-HSLN[4]. We attribute the substantial improvement observed, particularly in DeepRhole, to a statistical property of the dataset: on average, a rhetorical label persists across approximately 8.56 consecutive sentences before shifting to another[5][6]. Consequently, fully hierarchical models like BERT-HSLN, which process entire document sequences, may dilute the relevant signal by incorporating structurally irrelevant or conflicting content. In contrast, a simpler PLM guided by a well-targeted *Before* context can focus more effectively on the most informative local cues, resulting in more accurate and efficient predictions.

However, LegalEval remains challenging, as these PLMs have not yet matched SOTA performance. A plausible explanation is its higher label complexity, making it difficult for small models like BERT to achieve strong discrimination, as noted in SCOTUS annotation guidelines (Lavissière and Bonnard, 2024).

Finally, our retrieval-based models offer substantial efficiency gains compared to BERT-HSLN. With $k = 6$, our models typically process around 500 tokens per example using BERT as the backbone, whereas BERT-HSLN requires additional components for enriching representations and processes entire document sequences. This results in a $\sim 3\times$ to $5\times$ reduction in GPU memory usage and $\sim 2\times$ to $4\times$ faster training and inference time, depending on batch size and model configuration (see Appendix A for details).
Additional results with RoBERTa (Liu et al., 2019), LegalBERT (Chalkidis et al., 2020), and Longformer (Beltagy et al., 2020) are provided in Appendix B.

6 Conclusion and Future Work

In this study, we investigated how the role of context affects the SSC task in long legal documents. Our findings reveal that sequential heuristics, which preserve the natural flow of discourse, systematically lead to stronger performance gains than selective heuristics. An important insight is

[4]For a fair comparison, we compare against the original model, which does not include our modifications introduced in context analysis.

[5]Segment refers to consecutive annotation units (sentences) that share the same label within a document.

[6]The statistics are based on our corpus analysis.

that similarity alone is not enough—what matters more is where the sentence appears and whether the extracted relevant context forms a coherent unit. Moreover, enriching PLMs such as BERT with useful context yielded significant improvements over hierarchical models that process entire documents. Future work should give priority to (1) expanding the study to the corpus level, where multi-document signals will be explored, and (2) refining selective heuristics to extract high-quality context without increasing noise.

7 Limitations

While this study demonstrates the benefits of contextual information for SSC, a few limitations must be considered:

- We purposefully kept the heuristics basic, as our focus is not on peak performance. Nonetheless, more sophisticated approaches may yield higher scores than what we present.
- We have focused our experiments on a single document. In practice, integrating the context of several documents could potentially offer richer information for selective heuristics.
- We cannot reject the hypothesis that our findings about the utility of context may not be universally generalizable across other tasks. Our analysis centered on legal datasets, and thus further research is needed to determine whether similar gains would arise in other settings.

8 Ethical Statement

This work fully complies with the ACL Ethics Policy. To the best of our knowledge, we declare that there are no ethical issues in this paper.

9 Acknowledgments

This research was funded, in whole or in part, by l'Agence Nationale de la Recherche (ANR), project ANR-22-CE38-0004.

References

Arthur Amalvy, Vincent Labatut, and Richard Dufour. 2023. The role of global and local context in named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 714–722, Toronto, Canada. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2023. Deeprhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law*, pages 1–38.

A Brack, A Hoppe, P Buschermöhle, and R Ewerth. 2021. Sequential sentence classification in research papers using cross-domain multi-task learning. corr. *arXiv preprint arXiv:2102.06008*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Yu Chen, You Zhang, Jin Wang, and Xuejie Zhang. 2023. YNU-HPCC at SemEval-2023 task 6: LEGAL-BERT based hierarchical BiLSTM with CRF for rhetorical roles prediction. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2075–2081, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Varun Gangal, Steven Y. Feng, Malihe Alikhani, Teruko Mitamura, and Eduard Hovy. 2022. Nareor: The narrative reordering problem. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10645–10653.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3100–3109, Brussels, Belgium. Association for Computational Linguistics.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and

Ashutosh Modi. 2022. Corpus for automatic structuring of legal documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4420–4429, Marseille, France. European Language Resources Association.

Mengfei Lan, Lecheng Zheng, Shufan Ming, and Halil Kilicoglu. 2024. Multi-label sequential sentence classification via large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16086–16104, Miami, Florida, USA. Association for Computational Linguistics.

Mary C. Lavissière and Warren Bonnard. 2024. Who's really got the right moves? analyzing recommendations for writing american judicial opinions. *Languages*, 9(4).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Akheel Muhammed, Hamna Muslihuddeen, Shalaka Sankar, and M Anand Kumar. 2024. Impact of rhetorical roles in abstractive legal document summarization. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE.

Mariana Neves, Daniel Butzke, and Barbara Grune. 2019. Evaluation of scientific elements for text similarity in biomedical publications. In *Proceedings of the 6th Workshop on Argument Mining*, pages 124–135, Florence, Italy. Association for Computational Linguistics.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2025. Nomic embed: Training a reproducible long context text embedder.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. 2024. Will LLMs replace the encoder-only models in temporal relation classification? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20402–20415, Miami, Florida, USA. Association for Computational Linguistics.

Iqra Safder and Saeed-Ul Hassan. 2019. Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications. *Scientometrics*, 119:257–277.

T.y.s.s Santosh, Apolline Isaia, Shiyu Hong, and Matthias Grabmair. 2024. HiCuLR: Hierarchical curriculum learning for rhetorical role labeling of legal documents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7357–7364, Miami, Florida, USA. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

Santosh T.y.s.s., Hassan Sarwat, Ahmed Mohamed Abdelaal Abdou, and Matthias Grabmair. 2024. Mind your neighbours: Leveraging analogous instances for rhetorical role labeling for legal documents. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11296–11306, Torino, Italia. ELRA and ICCL.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

A Model Overview for Context Analysis

The model consists of four key components:

- **Word Embedding:** The target sentence and its retrieved context are encoded using BERT (Devlin et al., 2019), generating word-level embeddings.
- **Sentence Encoding:** A Bi-LSTM (Hochreiter, 1997) processes these embeddings, followed by attention-based pooling to obtain sentence representations.
- **Context Enrichment:** This layer models inter-sentence relationships to refine contextualized embeddings.
- **Output Layer:** A linear transformation maps the target sentence representation to logits, with labels predicted via softmax[7].

Model	Seq	DeepRhole	Legal Eval	SCOTUS
Roberta-base (baseline)	512	52.63	72.43	76.28
+ Before		**68.29**†	78.3†	**81.75**†
+ After		60.3†	**80.12**†	81.43†
+ Surrounding		63.86†	78.40†	80.10†
+ Random		50.04	72.35	75.79
+ BM25		53.54	72.79	77.78‡
+ Sentence-BERT		53.33	73.25‡	77.84‡
Legal-BERT (baseline)	512	54.06	69.43	76.85
+ Before		**69.10**†	79.65†	81.40†
+ After		63.19†	**80.99**†	**82.81**†
+ Surrounding		67.15†	78.55†	78.72
+ Random		50.32	68.55	76.56
+ BM25		54.59	70.77‡	77.06
+ Sentence-BERT		56.30	70.55	77.47
Longformer (baseline)	4096	53.83	72.57	76.26
+ Before		**67.62**†	79.89†	**81.58**†
+ After		61.16†	**80.09**†	81.09†
+ Surrounding		64.83†	73.09†	81.35†
+ Random		52.55	72.54	75.78
+ BM25		54.82	73.22	77.44†
+ Sentence-BERT		54.3	77.95‡	77.47‡

Table 3: Performance of PLMs using the best configuration observed in context analysis for $k \leq 6$ for each heuristic. Bold values represent the best improvement over the baseline (w/o context), while underlined values indicate the second-best. Markers † and ‡ denote statistical significance over the baseline at $p = 0.05$ and $p = 0.01$, respectively.

[7]We optimize for the target sentence, eliminating the CRF layer, as supported by the ablation study in Jin and Szolovits (2018).

Dataset	Source	Sub-domain	Targets
DeepRhole	(Bhattacharya et al., 2023)	Indian law	7 classes
Legal Eval	(Kalamkar et al., 2022)	Indian law	13 classes
SCOTUS	(Lavissière and Bonnard, 2024)	U.S. law	13 classes

Table 4: Statistics of the datasets used for evaluation.

B Additional Results

We report additional results with enriching PLMs: RoBERTa (Liu et al., 2019), LegalBERT (Chalkidis et al., 2020), and Longformer (Beltagy et al., 2020) in Table 3.

Target Sentence: *"This case focuses upon the requirement of 'fair presentation.'"*

Heuristic	Extracted Sentence
Before	*"O'Sullivan v. Boerckel, 526 U.S. 838, 845 (1999)."*
After	*"Michael Reese, the respondent, appealed his state-court kidnapping and attempted sodomy convictions and sentences through Oregon's state court system."*
Surrounding	*"O'Sullivan v. Boerckel, 526 U.S. 838, 845 (1999)." "Michael Reese, the respondent, appealed his state-court kidnapping and attempted sodomy convictions and sentences through Oregon's state court system."*
Random	*"In such instances, the nature of the issue may matter more than does the legal validity of the lower court decision."*
BM25	*"For another thing, the opinion-reading requirement would impose a serious burden upon judges of state appellate courts, particularly those with discretionary review powers."*
Sentence-BERT	*"The petition provides no citation of any case that might have alerted the court to the alleged federal nature of the claim."*

Table 2: Examples of sentences extracted using different heuristics from the SCOTUS dataset.