Transductive Legal Judgment Prediction Combining BERT Embeddings with Delaunay-Based GNNs

Hugo Attali and **Nadi Tomeh**
LIPN, CNRS UMR 7030, Université Paris Sorbonne Nord, France

Abstract

This paper presents a novel approach to legal judgment prediction by combining BERT embeddings with a Delaunay-based Graph Neural Network (GNN). Unlike inductive methods that classify legal documents independently, our transductive approach models the entire document set as a graph, capturing both contextual and relational information. This method significantly improves classification accuracy by enabling effective label propagation across connected documents. Evaluated on the Swiss-Judgment-Prediction (SJP) dataset, our model outperforms established baselines, including larger models with cross-lingual training and data augmentation techniques, while maintaining efficiency with minimal computational overhead.

1 Introduction

Modeling legal texts have attracted lots of interest recently in two directions (Cui et al., 2023). The first is to gather large collections of legal text such as the MultiLegalPile corpus (Niklaus et al., 2024) and train legal large language models (LLMs) such as (Colombo et al., 2024). The second focuses on smaller, manually annotated and specialized datasets and benchmarks such as the Swiss Judgment Prediction[1] (SJP) dataset (Niklaus et al., 2021), LexGLUE (Chalkidis et al., 2022) and LEXTREME (Niklaus et al., 2023), and train smaller supervised models, mainly by finetuning BERT-like models, sometimes applying cross-lingual transfer and data augmentation (Niklaus et al., 2022).

General-purpose LLMs like ChatGPT often perform poorly on legal tasks in zero and few-shot settings (Chalkidis, 2023; Niklaus et al., 2023), though they can be useful as components in larger frameworks (Wu et al., 2023). Specialized models, fine-tuned with supervised learning (Niklaus et al., 2021, 2022, 2023), require significant resources to improve performance, such as applying cross-lingual transfer, adapter-based fine-tuning, or tripling the dataset size with machine-translated documents (Niklaus et al., 2022). The suboptimal performance is likely due to the complexity of legal texts, which are long, dense, and filled with specialized terminology that generic pre-trained models struggle to understand. Additionally, these models lack sufficient exposure to the contextual and nuanced nature of legal reasoning, requiring more domain-specific data to adapt effectively.

In this paper, we hypothesize that transductive learning techniques (Gammerman et al., 1998; Joachims, 1999) are well adapted to Legal Judgment Prediction (LJP) as it has been shown to work well in few-shot scenarios (Liu et al., 2019; Colombo et al., 2023) and on small training datasets (Li et al., 2021; Lin et al., 2021). Along these lines, we construct a single graph with all training (labeled) and test (unlabeled) documents as nodes, allowing a Graph Neural Network (GNN) to learn from the entire dataset simultaneously. This approach leverages the relationships between documents for effective label propagation and context-aware classification, improving generalization by using both labeled and unlabeled data. It also captures domain-specific knowledge through connections like citations and shared terminology, adapts dynamically to the test set, and reduces overfitting by integrating test data into the learning process.

Our model (§3) is a simple and efficient graph-based approach that achieves state-of-the-art results on the Swiss Judgment Prediction (SJP) task (Niklaus et al., 2021) without additional resources. It is also simpler than existing transductive graph-based models for document classificaiton (Lin et al., 2021). Experiments (§4) show it outperforms strong baselines from the literature and a new zero-shot SaulLM-7B baseline (Colombo et al., 2024).

[1] We use the term prediction in the machine learning sense and not in the juridical sense (Medvedeva and Mcbride, 2023).

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 187–193
November 16, 2024

2 Related Work

Transductive GNNs for Text Classification GNNs (Goller and Kuchler, 1996) have demonstrated effectiveness across various domains (Wu et al., 2020; Nathani et al., 2019; Schlichtkrull et al., 2018; Vashishth et al., 2020), and have been applied to various text processing tasks (Nikolentzos et al., 2020; Wang et al., 2024). Most similar to our work is their use in transducive models. For instance, BertGCN (Lin et al., 2021) which builds a heterogeneous graph over a dataset, representing documents as nodes using BERT embeddings and modeling semantic relationships between them, allowing both labeled and unlabeled data to contribute to learning. Our model differs by using Delaunay triangulation for simpler graph construction, avoiding joint BERT and GCN training to reduce memory usage, and not requiring interpolation with a separate BERT-based classifier, resulting in more efficient graph construction and faster training. KnnGCN (Benamira et al., 2019) constructs corpus-level graphs using a KNN approach, which is less suited to GNNs than our Delaunay-based method. In contrast, TextGTL (Li et al., 2021) builds three non-heterogeneous graphs (Semantic, Syntax, and Context Text Graphs) using complex techniques like canonical correlation analysis and dependency parsing, whereas our model employs simpler graph construction techniques. Furthermore, none of the previous models have been specifically applied to LJP.

Graph-Based Methods in Legal Text Graph-based models have been explored for legal judgment prediction, similar to our approach. Zhao et al. (2022) use a graph network with heterogeneous text graphs and a GCN to predict outcomes, while LADAN (Xu et al., 2020) employs a graph neural network and attention mechanism to distinguish between confusing law articles. However, neither constructs a comprehensive graph for all documents, as we do. Other methods focus on different tasks, such as LegalGNN (Yang et al., 2021) for legal recommendations, using a heterogeneous graph with user queries, and CaseGNN (Tang et al., 2024) for legal case retrieval by modeling document-level relationships.

3 Method

In this section we describe our architecture, also depicted in Figure 1.

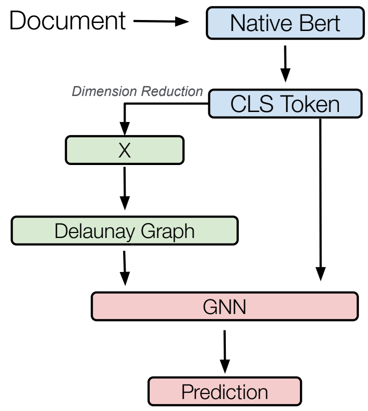

Figure 1: Our model architecture. A document is processed through a BERT model to obtain CLS tokens, which are then used alongside the Delaunay graph of documents for classification using a GNN.

Document Encoder We begin by modeling documents as a graph, using the [CLS] tokens extracted from a standard BERT model (Devlin, 2018) (up to 512 tokens) to represent each document. While this approach leverages BERT's document representation, our method is flexible and can easily incorporate other encoders that provide document representations. Documents that are longer than BERT context capacity are cut off. In contrast to our simple approach, some of the baselines we present in §4.2 handle long documents hierarchically or using larger models.

Delaunay Graph To effectively model documents as a graph, we propose using a a Delaunay graph (Attali et al., 2024). This kind of graph is particularly advantageous for information propagation by a GNN. It helps mitigating common challenges such as oversquashing (Alon and Yahav, 2021) – information loss due to bottleneck structures in the graph, and oversmoothing (Oono and Suzuki, 2020; Cai and Wang, 2020) – information mixing which can blur distinctions between nodes. In fact, Delaunay graphs do not have tight bottlenecks and large cliques (Nguyen et al., 2023). Additionally, Delaunay triangulation correlates with improved *homophily* of the graph, meaning it better captures the similarity between connected nodes.

In our approach, each document to be classified is represented as a node within this graph. To construct the graph, we employ a strategy similar to that used in Attali et al. (2024). First, we perform a Delaunay triangulation in a 2-dimensional feature space, where each [CLS] token represents the document's embedding. Since the [CLS] token is typically high-dimensional, we reduce its dimensionality using UMAP (McInnes et al., 2018) that preserves the local structure of data. Delaunay graphs basically establish relationships between documents based on their distances in feature space. This operation is computationally efficient and scalable as we show in our experiments §4.

GNN-Based Classification Finally, for classification, we use a simple GCN (Kipf and Welling, 2017). Our GCN takes as input the `[CLS]` output from BERT, which represents the document (node) embeddings, and the adjacency matrix of the Delaunay graph. We construct a single graph for training, validation and test sets.

Training To maintain simplicity and modularity, we adopt a two-stage training approach. In the first stage, we add a binary classification MLP on top of BERT's `[CLS]` token and train both BERT and the MLP to minimize the binary cross-entropy loss using the true labels from the training set. The MLP is used only during this training phase. In the second stage, we train the GNN on the Delaunay graph constructed from all document embeddings, using the same binary classification loss on the training set labels.

4 Experiments

4.1 Dataset

To assess the effectiveness of our method, we utilize the task of Legal Judgment Prediction, aiming to forecast the verdict of a case based on the provided facts (Aletras et al., 2016; Zhong et al., 2018; Chalkidis et al., 2019a; Niklaus et al., 2021; Cui et al., 2023). For this evaluation, we use the Swiss-Judgment-Prediction dataset (Niklaus et al., 2021), a comprehensive multilingual resource comprising 85,000 cases from the Swiss Federal Supreme Court (FSCS). Each case in this dataset is annotated with a binarized judgment outcome, indicating either approval or dismissal. See Table 1 for dataset statistics.

Dataset	#Train	#Val	#Test	#Time
Italian	3,072	408	812	$\approx 11s$
German	35,452	4,705	9,725	$\approx 50s$
French	21,179	3,095	6,820	$\approx 30s$

Table 1: Dataset statistics. Time indicates the total time required to construct the graph, including the time spent on dimensionality reduction.

4.2 Baselines

Finetuned LMs We compare our architecture with three types of **monolingual** baselines as presented by Niklaus et al. (2021). The simplest ones use standard BERT (Devlin, 2018) for German (Branden Chan and Yeung, 2019), French (Martin et al., 2019), and Italian (Parisi et al., 2020), handling up to 512 tokens. Long BERT is an extended version of Standard BERT that includes additional positional encodings, allowing it to process longer texts of up to 2048 tokens. Hierarchical BERT, on the other hand, first processes text segments of up to 512 tokens each with a standard BERT, and then combines these segment encodings using a BiLSTM (Chalkidis et al., 2019b). We also compare to **multilingual** baselines that use pre-trained XLM-R (Conneau, 2019) along with **data augmentation** techniques based on machine translation and cross-lingual transfer as presented by Niklaus et al. (2022).

Zero-shot LLM (SaulLM-7B) In this baseline, we use a role-based prompt instructing the model to evaluate legal cases as a Swiss judge, analyzing the facts step-by-step and determining whether to dismiss or approve the request in a chain-of-thought style (Wei et al., 2024). SaulLM-7B (Colombo et al., 2024) is employed through a text generation pipeline, generating responses with a limit of 600 tokens. The outputs are parsed using regular expressions and conflict resolution rules to identify patterns indicating each class.

4.3 Experimental Setup

For the experiments, we follow the same training procedure as described in (Niklaus et al., 2021). For our method, we use the standard BERT `[CLS]` token embedding (up to 512 tokens). For the final classification we use a GCN (Kipf and Welling, 2017). We fix the number of layers to 2 and the dropout rate to 0.5, in line with (Pei et al., 2020; Attali et al., 2024). We fine-tune the learning rate, testing values of {0.005, 0.0005, 0.0001}, and the weight decay among {5e-05, 5e-6, 5e-07} on the validation set. The main results are presented in Table 2, where we report the average macro-averages F1-score for each method across 5 runs. We use the macro-averaged F1-score instead of the micro-average to give equal weight to all classes, ensuring that the performance on less frequent classes is fairly represented.

Model	De	Fr	It
Majority	44.5	44.9	44.8
Stratified	50.0	50.0	48.8
Linear (BoW)	52.6	56.6	53.9
BERT	63.7	58.6	55.2
Long BERT	67.9	68.0	59.8
Hierarchical BERT	68.5	70.2	57.1
Hierarchical BERT+MT	70.0	71.0	71.9
XLM-R+Adapters+CL	69.9	71.8	70.7
XLM-R+Adapt.+CL+MT	70.3	72.1	72.3
SaulLM-7B	51.0	52.0	52.0
BERT+Delaunay+GCN	**79.2**	**77.5**	**74.4**

Table 2: Main results. The baselines including BERT and XLM-R are taken from (Niklaus et al., 2021, 2022). Best scores are in bold. Our method achieves standard deviations ranging between 0.5 and 0.7 across different languages, making it the most stable method compared to the baselines.

4.4 Results

Main Findings Our model achieves the highest scores across all languages as presented in Table 2. This demonstrates that our approach, which builds on top of a fine-tuned BERT outperforms the BERT baseline with *negligible computational overhead* and without retraining BERT. Despite being a smaller model, BERT+Delaunay+GCN outperforms Hierarchical BERT and Long BERT, and XLM-R models with cross-lingual training and data augmentation techniques like machine translation. Additionally, our transductive approach seems to mitigate the lack of resources, as seen in the results for the Italian dataset. While the Italian scores are generally lower than those for German and French, mainly due to the smaller dataset size. This underscores our model's robustness, particularly for lower-resource languages. Finally, our model outperforms the specialized legal LLM (SaulLM-7B), confirming findings from the literature that generic, powerful language models like ChatGPT underperform on this task (Niklaus et al., 2023; Chalkidis, 2023).

Running Time The Delaunay graph can be constructed efficiently including dimensionality reduction as presented in Table 1. Adding a GCN-based classification layer is highly scalable and computationally efficient. On average, a single run of classification takes 91 seconds on the German dataset, 42 seconds on the French dataset, and 5 seconds on the Italian dataset when using a T4 GPU.

	De	Fr	It
SBERT + Delaunay+GCN	44.8	47.6	51.9
BERT + KMeans	52.0	74.2	66.4
BERT + Delaunay+GCN	**79.2**	**77.5**	**74.4**

Table 3: Results of our ablation study.

Ablations To demonstrate the necessity of both (a) fine-tuning document representations for the task at hand and (b) enriching them through GNNs, we conducted a series of comparisons. First, we replaced the Delaunay+GCN part of the architecture with KMeans unsupervised clustering on `[CLS]` tokens which does not need any training. In a second experiments, we replaced the finetuned BERT with pre-trained SBERT (Reimers, 2019) without any further finetuning on the task to generate document embeddings. The results are shown in Table 3.

The results show that our method consistently outperforms both KMeans clustering and SBERT-based encoding, emphasizing the importance of first fine-tuning document representations for task-specific alignment and then further refining them with graph-based methods like Delaunay GNN. This approach effectively captures structural relationships, enhancing representation quality and leading to more accurate classification.

5 Conclusions

This paper demonstrates that a transductive legal judgment prediction method, combining BERT embeddings with Delaunay-based GNNs, significantly outperforms traditional inductive classification methods by effectively utilizing contextual and relational information between legal documents for more accurate label propagation and classification. In future work, we will study the necessity of retraining the model whenever a new batch of documents are to be classified. We will also explore semi-supervised training approaches to study the dependency of the performance on annotated data.

6 Limitations

Our study is limited by its exclusive focus on the SJP dataset, which may affect its generalizability to other legal systems. The model may also inherit biases from the training data, and we have not performed a bias analysis. While our approach improves performance, it may not fully capture all the complex factors influencing judicial decisions and may face scalability challenges with larger datasets.

7 Ethics Statement

Our work uses machine learning techniques for legal judgment prediction based on SJP dataset. We acknowledge that models trained on historical data may inherit biases, such as disparities in legal decisions or underrepresentation of certain groups. Since our model is based on cases from the Swiss Federal Supreme Court, it may not generalize to other jurisdictions or legal systems with different laws or cultural contexts. We have not tested its applicability outside the Swiss judicial system, and extending it to other settings would require careful adaptation and validation.

Our method is not intended to replace human judgment but to provide supplementary insights to legal professionals. Its outputs should be viewed as probabilistic suggestions, not definitive conclusions, and should always be used alongside human oversight to consider the broader context and ethical implications not captured in the training data.

To mitigate risks of bias and unjust outcomes, we recommend integrating our model in a way that enhances, rather than replaces, human decision-making. Any deployment should include mechanisms for regular monitoring and auditing to detect and address potential biases promptly, ensuring its alignment with fair legal practices.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Comput. Sci.*, 2:e93.

Uri Alon and Eran Yahav. 2021. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*.

Hugo Attali, Davide Buscaldi, and Nathalie Pernelle. 2024. Delaunay graph: Addressing over-squashing and over-smoothing using delaunay triangulation. In *Forty-first International Conference on Machine Learning*.

Adrien Benamira, Benjamin Devillers, Etienne Lesot, Ayush K. Ray, Manal Saadi, and Fragkiskos Malliaros. 2019. Semi-Supervised Learning and Graph Neural Networks for Fake News Detection. In *ASONAM 2019 - IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Vancouver, Canada.

Malte Pietsch Tanay Soni Branden Chan, Timo Möller and Chin Man Yeung. 2019. German bert. https://deepset.ai/german-bert.

Chen Cai and Yusu Wang. 2020. A note on over-smoothing for graph neural networks. *Graph Representation Learning*.

Ilias Chalkidis. 2023. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark. *Preprint*, arXiv:2304.12202.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019b. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Pierre Colombo, Michael Desa, Telmo Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, André Martins, Fabrizio Esposito, Vera Raposo, and Sofia Morgado. 2024. SaulLM-7B: A pioneering Large Language Model for Law. Working paper or preprint.

Pierre Colombo, Victor Pellegrain, Malik Boudiaf, Myriam Tami, Victor Storchan, Ismail Ayed, and Pablo Piantanida. 2023. Transductive learning for textual few-shot classification in API-based embedding models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4214–4231, Singapore. Association for Computational Linguistics.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 11:102050–102071.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik. 1998. Learning by transduction. In *UAI '98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, University of Wisconsin Business School, Madison, Wisconsin, USA, July 24-26, 1998*, pages 148–155. Morgan Kaufmann.

Christoph Goller and Andreas Kuchler. 1996. Learning task-dependent distributed representations by backpropagation through structure. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 1, pages 347–352. IEEE.

Thorsten Joachims. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 200–209, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the International Conference on Learning Representations*, ICLR.

Chen Li, Xutan Peng, Hao Peng, Jianxin Li, and Lihong Wang. 2021. Textgtl: Graph-based transductive learning for semi-supervised text classification via structure-sensitive interpolation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2680–2686. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. BertGCN: Transductive text classification by combining GNN and BERT. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1456–1462, Online. Association for Computational Linguistics.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. 2019. Learning to propagate labels: Transductive propagation network for few-shot learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Masha Medvedeva and Pauline Mcbride. 2023. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84, Singapore. Association for Computational Linguistics.

Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. 2019. Learning attention-based embeddings for relation prediction in knowledge graphs. *arXiv preprint arXiv:1906.01195*.

Khang Nguyen, Nong Minh Hieu, Vinh Duc Nguyen, Nhat Ho, Stanley Osher, and Tan Minh Nguyen. 2023. Revisiting over-smoothing and over-squashing using ollivier-ricci curvature. In *International Conference on Machine Learning*, pages 25956–25979. PMLR.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. MultiLegalPile: A 689GB multilingual legal corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094, Bangkok, Thailand. Association for Computational Linguistics.

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An empirical study on cross-X transfer for legal judgment prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only. Association for Computational Linguistics.

Giannis Nikolentzos, Antoine Tixier, and Michalis Vazirgiannis. 2020. Message passing attention networks for document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI, pages 8544–8551.

Kenta Oono and Taiji Suzuki. 2020. Graph neural networks exponentially lose expressive power for node classification. *Proceedings of the International Conference on Learning Representations*.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. *Original-date*, 55:31Z.

Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. 2020. Geom-gcn: Geometric graph convolutional networks. *arXiv preprint arXiv:2002.05287*.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling.

2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer.

Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2024. Casegnn: Graph neural networks fornbsp;legal case retrieval withnbsp;text-attributed graphs. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II*, page 80–95, Berlin, Heidelberg. Springer-Verlag.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2020. Composition-based multi-relational graph convolutional networks. ICLR.

Kunze Wang, Yihao Ding, and Soyeon Caren Han. 2024. Graph neural networks for text classification: a survey. *Artificial Intelligence Review*, 57(8).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075, Singapore. Association for Computational Linguistics.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. In *IEEE transactions on neural networks and learning systems*, volume 32, pages 4–24. IEEE.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095, Online. Association for Computational Linguistics.

Jun Yang, Weizhi Ma, Min Zhang, Xin Zhou, Yiqun Liu, and Shaoping Ma. 2021. Legalgnn: Legal information enhanced graph neural network for recommendation. *ACM Trans. Inf. Syst.*, 40(2).

Qihui Zhao, Tianhan Gao, Song Zhou, Dapeng Li, and Yingyou Wen. 2022. Legal judgment prediction via heterogeneous graphs and knowledge of law articles. *Applied Sciences*, 12(5).

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.