Improving Legal Citation Recommendation with DistilBERT: A Study on Performance and Efficiency

Louis Kapp
Boğaziçi University
Department of Computer Engineering
34342 Bebek, Istanbul, Turkey
louis.kapp@boun.edu.tr

Abstract

This paper presents an enhanced legal citation recommendation system utilizing a DistilBERT-based transformer model. The system's primary objective is to suggest suitable legal citations in real-time to legal document writers, thus speeding up the legal writing process. Previous studies indicated only slight improvements using RoBERTa over BiLSTM models, motivating the exploration of DistilBERT's potential. This research adopted an efficient offline preprocessing approach, introducing additional steps that increased available data samples and reduced training time. The results revealed a significant performance enhancement by DistilBERT over the baseline RoBERTa model across various Recall@k metrics. Particularly noteworthy is that these improvements were achieved after training for only a single epoch. These findings highlight not only the effectiveness of DistilBERT but also its practical potential, as its rapid inference capabilities make it a suitable choice for embedding within text editors. Future research will focus on potential performance enhancement through extended training epochs and hyperparameter optimization.

1 Introduction

The task of recommending legal citations involves suggesting applicable resources to legal document writers. Such resources are vital in validating the writer's arguments with legal backing. They commonly encompass statutes enacted by legislatures, regulations authored by agencies, and specific historical cases. However, researching suitable sources for citation is typically a time-intensive process, often contributing to a substantial backlog of cases and an overall deceleration of jurisdiction.

A computational recommendation system capable of suggesting pertinent, citable resources in real-time during the process of writing could significantly expedite the duties of attorneys crafting briefs or memoranda, and judges authoring opinions. The utility of such a system is heightened when it is context-aware; that is when it uses previously written text to inform citation recommendations. The system requires a (legal) understanding of the written content to achieve this.

Previous research has demonstrated that the optimal means of achieving this level of understanding leverages deep neural language models like Bi-LSTMs (Hochreiter and Schmidhuber, 1997) or BERT-based (Devlin et al., 2019) transformer models. These models generate word embeddings from the text, enabling subsequent processing by a classification model (Huang et al., 2021). Notably, RoBERTa (Liu et al., 2019) significantly outperforms BiLSTMs on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018), securing 21st place on the GLUE leaderboard with a score of 88.1, while the top BiLSTM variant occupies 85th place with a score of 70.0 [1].

The RoBERTa model's superiority can be attributed to its robust training scheme and architecture which, unlike BiLSTM, employs attention mechanisms, enabling it to understand longer dependencies in sentences. This characteristic makes RoBERTa potentially more suitable for the task of legal citation recommendation, which often requires understanding complex legal arguments spread across several sentences or even paragraphs.

[1] The GLUE Leaderboard is publicly available at https://gluebenchmark.com/leaderboard

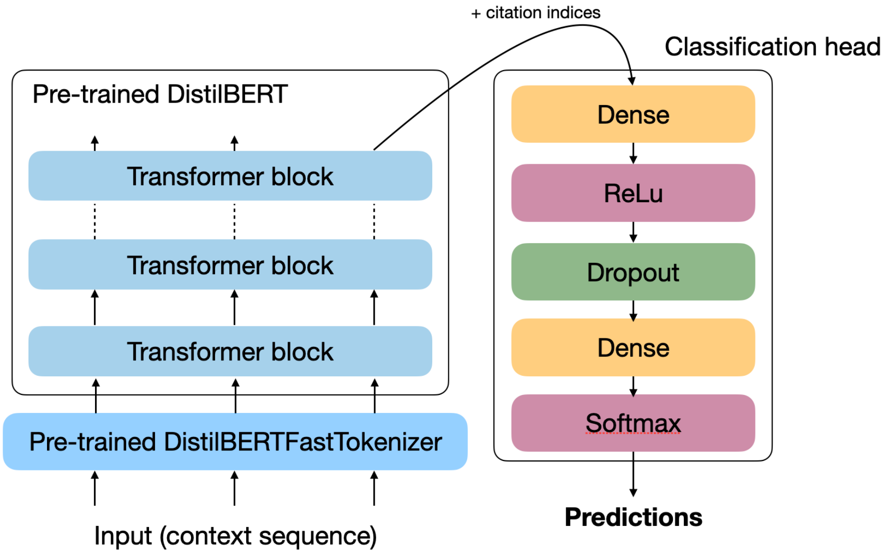

Figure 1: Visual overview of the model architecture used to make citation recommendations.

However, contrary to expectations, (Huang et al., 2021) have found that RoBERTa's performance was only slightly superior to that of the BiLSTM (a maximum 1% difference in Recall@k for all tested k). They proposed potential explanations for this finding, including RoBERTa's domain-agnostic pretraining and the task's inherent nature. While these explanations provided by the authors offer valuable insights, they encourage further exploration into this area. This prompted my research, which has not only aspired but also successfully realized a higher performance using a pretrained DistilBERT (Sanh et al., 2019) model.

2 System Description

The process of context-aware citation recommendation is typically approached as a classification task, whereby the model assigns a probability to each prospective resource to cite. While the recommendation could theoretically also be treated as a language generation task, the complexity of mitigating the hallucination issue in large language models - a challenge that continues to be an active area of research (Ji et al., 2023) - renders this approach less viable.

Our method for resource recommendation involves a system comprised of an encoder component and a classification head. Figure 1 gives an overview of the architecture. The text sequence necessitating a citation is introduced into the encoder, resulting in the generation of a vector representation of the sequence. Subsequently, this embedding is inputted into the classification head, which consists of a singular dense layer. Additional layers are deemed unnecessary, as the critical features have been extracted by the encoder. This methodology aligns closely with the model proposed by (Huang et al., 2021).

Rather than employing RoBERTa for the encoding, we utilize the pre-trained DistilBERT *distilbert-base-uncased* model - a more streamlined version from the BERT (Devlin et al., 2019) family. This model retains approximately 95% of BERT's performance on the GLUE benchmark while boasting a 60% increase in inference speed due to having 40% fewer parameters than *bert-base-uncased* (Sanh et al., 2019).

This characteristic allows for model fine-tuning on citation recommendation, even when computational resources for training are limited. Moreover, the expedited inference speed facilitates the deployment of the model as part of a text editor within an actual user application. A quick recommendation is vital in further enhancing the productivity of proficient jurists.

3 Experimental Setup

3.1 Dataset

The preprocessed version of the BVA Corpus (Huang et al., 2021) is utilized in this study. This corpus contains 324,309 appeal decisions made by the U.S. Board of Veterans' Appeals (BVA). Each appeal decision, encapsulated in a separate file, contains citations that have been cleaned and masked by a "@cit@" token. Another handy resource is a unique citation resource vocabulary, compiled by (Huang et al., 2021), containing 4287 unique entries. Each citable resource can be uniquely identified using its vocabulary index.

3.2 Preprocessing

In contrast to the approach taken by (Huang et al., 2021), this paper introduces additional preprocessing steps. Their method included time-consuming processes such as disk I/O operations during text retrieval, repeated tokenization of the entire text, and the need to locate "@cit@" tokens when randomly sampling an input sequence from the text. This method carried out during the data loading phase, significantly extended the training time. Moreover, this approach failed to capitalize fully on all available data samples, which, if used, could improve the model's generalization, robustness, and subsequent test performance.

In response to these challenges, a more efficient offline preprocessing procedure was devised. Initially, all texts are tokenized. For each citation within each text, the citation and its corresponding context window (a predefined span of text, extending to 256 tokens preceding the citation) are extracted. The extracted citation contexts form a part of the data samples which, in supervised classification tasks, are paired with corresponding citations to be predicted by the model. Finally, all context vectors and labels (citation indices) are stored in a single binary .pt file, which allows the direct loading of model inputs into torch tensors. This method significantly reduces data loading time, as it only involves a single large disk operation.

The modifications result in 3,140,609 data samples available per epoch, almost a 10-fold increase from the initial 324,309. This is because, instead of randomly sampling a single context vector and citation index per appeal decision text, all available data samples within a text are used after one another.

3.3 Hyperparameters

The hyperparameters were chosen as a mix between those typical for DistilBERT and those that performed best in the baseline paper. The learning rate is set to $1e - 4$, the context window size is 256 tokens, and the forecasting window length is 16. The batch size is fixed at 576. To accommodate this batch size within the available memory, a 9-fold gradient accumulation strategy is adopted, where gradients are computed over smaller batches of 64, and combined for a single backpropagation. This allows for more parallel training, facilitated by large batch sizes while minimizing memory usage. All other hyperparameters are set to the default configurations of the DistilBERT Huggingface (Wolf et al., 2020) implementation for sequence classification. For tokenization, the DistilBertTokenizerFast from Huggingface is employed.

The dataset is divided into 72% for training, 18% for development, and 10% for testing. Initial training runs were conducted by shuffling the data samples in the dataset before splitting it. However, this might bias the model by artificially increasing performance, since it could access data samples from the same appeal decisions even though it did not have access to the exact validation/test data samples during training. Hence, the final experiments were conducted by splitting the ordered dataset without shuffling upfront.

In the baseline study, the best-performing RoBERTa model converged after 106 training epochs. However, given the significantly larger amount of data samples and limited computational resources, the DistilBERT model was trained for a single epoch only, and validated/tested thereafter. Future work will focus on hyperparameter optimization on the validation set.

	Number of Unique Data Samples		
Vocabulary Size	Training Set	Validation Set	Test Set
105	115,200	28,800	16,001
479	388,812	97,203	54,002
859	576,060	144,016	80,009
1431	1,065,724	266,431	148,018
4287	3,140,609	785,153	436,196

Table 1: Number of unique data samples per vocabulary size. All of these data samples are used during a single epoch.

Model	**Setting**	**Recall@1**	**Recall@5**	**Recall@20**
RoBERTa-base	vsize=4287, epochs=106	65.6%	82.8%	91.7%
DistilBERT-base	vsize=4287, split=ordered, epochs=1	**78.2%**	**91.7%**	**96.6%**
DistilBERT-base	vsize=859, split=shuffled, epochs=1.7	87.5%	96.5%	98.6%
DistilBERT-base	vsize=105, split=ordered, epochs=13.7	93.9%	99.2%	99.7%

Table 2: Model performance on the test set. In the setting column, *vsize* refers to the vocabulary size, which corresponds to the number of classes. *Split* denotes whether the dataset has been shuffled or not before splitting into train, validation, and test set. Epoch denotes the number of epochs the model was trained for.

3.4 Vocabulary Sizes

Initial training runs were conducted on smaller vocabulary sizes, as it was unclear whether satisfactory performance could be achieved with the original vocabulary size, given limited computational resources. Downsizing the vocabulary was a strategic choice to test which sizes could be feasibly trained until convergence. In general, fewer possible citation resources mean that there are fewer data samples that can be used for training because each data sample corresponds to a single citation index. Five different versions of the dataset, each corresponding to a different vocabulary size were created.

The original vocabulary of size 4287 was downsized using histogram equalization to maintain a distribution of citation frequencies akin to the original vocabulary. All citation entries were sorted according to their frequency of citation in the appeal decisions. The citation indices were then split into n equally sized groups (where n = 1429/857/476/102). The element with the lowest citation frequency from each group was selected and from these elements, a new and smaller citation vocabulary was constructed.

4 Results and Discussion

4.1 Results

Table 2 outlines the comparative performance of various models on the test set. The superior performance of the DistilBERT-base model over the baseline RoBERTa-base model for all tested Recall@k values is evident. It is especially significant given the equal vocabulary size of 4287 across both models, yet with the DistilBERT model trained for only a single epoch compared to the baseline model's extensive 106 training epochs. Potential for further performance enhancement through extended training epochs remains, however, this exploration falls beyond the scope of the present study and will be subject to future work.

Moreover, results derived from a DistilBERT model trained with a reduced vocabulary size of 859 demonstrate a remarkable performance improvement compared to the model trained with a vocabulary size of 4287. Notably, this improvement was attained with a shuffled dataset before splitting it into training, development, and test sets and an extended number of training steps.

Lastly, the superior performance of the smallest model, with a vocabulary size of 105, is worth noting. Despite not shuffling the dataset before splitting, this model achieved the highest performance. It was also trained for 13.7 epochs, longer than all other models.

4.2 Discussion

The superior performance of my DistilBERT-base model with a vocabulary size of 4287 suggests the effectiveness of offline preprocessing to yield more data samples. Yet, it's theoretically possible for the baseline model to access the same volume of data samples:

$$\begin{aligned} &\text{\# appeal decision texts} \times \text{\# epochs} > \text{\# total data samples} \\ \iff &324{,}309 \times 106 = 34{,}376{,}754 > 3{,}140{,}609 \end{aligned}$$

This raises the question: why does my model significantly outperform the baseline RoBERTa model? A plausible explanation may be that each training data sample in my model is exposed to the model exactly once, avoiding redundancy. This underscores the importance of judicious data sample selection and preparation, which may have significantly enhanced the learning capability of the model by reducing redundancy and the potential for overfitting.

On the other hand, the random sampling in the baseline model does not guarantee unique sample selection in each epoch, potentially leading to overfitting. Interestingly, the overfitting may not be readily observable during training. This could be due to the lack of adequate data samples per class in the early training epochs, limiting the model's capacity for generalization. As training progresses and more data samples become available, the evaluation performance improves. However, overfitting places a ceiling on achievable performance.

These explanations, while plausible, are not entirely convincing. The leap in performance could equally be attributed to alternative factors such as the different model architecture and variances in hyperparameters. Hence, pinning down the exact reasons behind the performance enhancement is not straightforward and needs further research.

The results obtained from the other tested models provide some interesting insights. A key observation is a relationship between vocabulary size and performance, where models trained with a reduced vocabulary size demonstrate superior performance. This implies a potential trade-off between performance and vocabulary size. It suggests that reducing the task's complexity, by having fewer classes to predict, can enhance the model's performance.

Furthermore, models trained for a longer number of epochs, such as the smallest model, trained for 13.7 epochs, performed better than those trained for fewer epochs. This suggests that the learning process could benefit from more exposure to the data, potentially identifying more complex patterns and relationships over time. However, there is a need to balance this against the risk of overfitting and the computational cost of longer training.

5 Conclusion

This study underscores the potential of DistilBERT-base in enhancing citation recommendation processes within legal texts. Notably, the model's performance, even when trained on a vocabulary size comparable to the baseline RoBERTa model, exhibits significant improvement over the baseline paper's results. This enhanced performance, combined with the speed of inference associated with DistilBERT, is particularly promising for practical applications.

The high Recall@k values demonstrate that jurists can expect relevant citation recommendations in a list of top suggestions, aiding in decision-making. Furthermore, DistilBERT's efficiency makes it ideal for integration within text editors, enabling real-time citation recommendations upon user interaction.

In conclusion, this study illustrates that transformer-based models, specifically DistilBERT, can effectively serve as powerful tools for citation recommendation, offering substantial improvements in terms of both recommendation quality and computational efficiency.

Acknowledgements

The well-structured dataset provided by (Huang et al., 2021) and their well-documented source code [2] were very useful throughout the process of conducting this research. My course instructor Tunga Güngör provided valuable feedback and suggestions during the research process for this paper.

References

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, June. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Zihan Huang, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel E. Ho, Mark S. Krass, and Matthias Grabmair. 2021. Context-aware legal citation recommendation using deep learning. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 79–88, New York, NY, USA. Association for Computing Machinery.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

[2] Publicly available under https://github.com/TUMLegalTech/bva-citation-prediction