Data-centric Advanced Knowledge Interface for Legal Archives

An NLP-Enhanced Tax Statute Text Search System in the Philippines

Matthew Roque
Department of Electronics and Computer Engineering
De La Salle University
Manila, Philippines
matthew_roque@dlsu.edu.ph

Shirley Chu
College of Computer Studies
De La Salle University
Manila, Philippines
shirley.chu@dlsu.edu.ph

Melvin Cabatuan
Department of Electronics and Computer Engineering
De La Salle University
Manila, Philippines
melvin.cabatuan@dlsu.edu.ph

Nicole Abejuela
Department of Electronics and Computer Engineering
De La Salle University
Manila, Philippines
nicole_abejuela@dlsu.edu.ph

Michael Ng
Department of Electronics and Computer Engineering
De La Salle University
Manila, Philippines
michael_stevens_ng@dlsu.edu.ph

Hans Nolasco
Department of Electronics and Computer Engineering
De La Salle University
Manila, Philippines
hans_nolasco@dlsu.edu.ph

Jenine Valencia
Department of Electronics and Computer Engineering
De La Salle University
Manila, Philippines
jenine_valencia@dlsu.edu.ph

ABSTRACT

This study presents an innovative and novel approach to section retrieval from the National Internal Revenue Code (NIRC) of the Philippines, leveraging advanced Natural Language Processing (NLP) techniques. The study shares the first documented dataset on manually annotated questions from the NIRC bar exam review materials, as well as a methodology for generating synthetic data using instruction-tuned Large Language Models such as Mistral 7b. Utilizing text embedding models, the research also explores the efficacy of preprocessing techniques, the impact of learning rates on model performance, and the computational considerations of using language models like Dense Passage Retrieval (DPR) and Jina Embeddings v2. The findings reveal that Jina Embeddings v2, trained on the combination of the original and synthetic datasets, delivers the highest accuracy, successfully retrieving a single relevant section out of 311, 70.52% of the time. A web application (DAKILA) was developed to house the system and serve as an interface between the retrieval pipeline and the user. Feedback was collected from students and practicing professionals from the fields of law and accountancy using the System Usability Scale (SUS), demonstrating strong user satisfaction and DAKILA's potential as a legal research tool.

KEYWORDS

Natural language processing, large language models, information retrieval, text embedding models, Philippine Tax Code, National Internal Revenue Code, statute law retrieval

1 INTRODUCTION

Natural Language Processing (NLP), a subset of Artificial Intelligence (AI) related to Machine Learning (ML) and Deep Learning (DL), has made significant strides in multiple industries within the last decade. Its impact on the legal sector is particularly noteworthy, offering innovative tools that serve both professionals and the public. NLP-driven predictive models have been instrumental in elevating the capabilities of legal practitioners, including lawyers and researchers. These models enhance case analyses and support the formulation of compelling legal arguments, marking a significant leap forward in legal practice [1, 26]. Moreover, advancements in automated summarization and question-answering technologies have dramatically improved the comprehension of complex legal documents. These tools efficiently distill extensive, intricate texts into manageable summaries, alleviating the burden of manual analysis. Consequently, they significantly reduce the likelihood of misinterpretation and variability in understanding, making legal information more accessible and navigable [13, 20, 21].

In the Philippine legal domain, various NLP methodologies have been applied, focusing on Supreme Court cases. One study predicted outcomes of criminal cases using text analysis with machine learning techniques like random forest classifiers and support vector machines [25]. Another explored semantic analysis for retrieving relevant case laws using Doc2Vec and cosine similarity measures [18]. Additionally, Juris2Vec, leveraging a mix of Word2Vec, GloVe, and fastText, created domain-specific embeddings tailored for Philippine legal texts, enhancing the precision and relevance of legal research in the Philippines [14].

While advancements in NLP have notably improved legal analysis and document accessibility in the Philippines, a significant

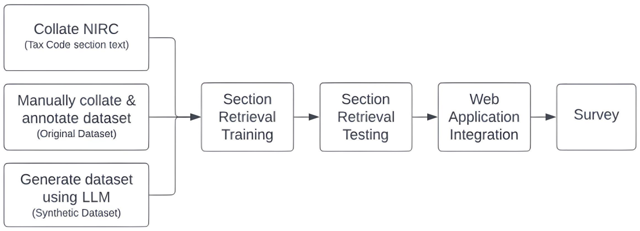
Collate NIRC
(Tax Code section text)
Manually collate & annotate dataset
(Original Dataset)
Generate dataset using LLM
(Synthetic Dataset)
Section Retrieval Training
Section Retrieval Testing
Web Application Integration
Survey

Figure 1: Main Process Flowchart

area for development remains in making statute laws more comprehensible to laypersons and more easily navigable for professionals. Current efforts focus on analyzing Supreme Court decisions, but there is a critical need for initiatives that simplify the statutory text for non-specialists. Enhancing NLP applications to decode complex legal terms into understandable language would not only empower the public with better legal understanding but also aid legal experts by facilitating faster access to essential statutory details. This highlights the importance of dedicated research towards creating tools specifically designed for the intricacies of Philippine statute law, aiming to eliminate the barrier posed by legal jargon.

This study aimed to pioneer a novel and credible section retrieval NLP dataset based on the National Internal Revenue Code (NIRC) as of 2023, also known as the Philippine Tax Code. The creation of NLP datasets tailored to the Philippine legal context is vital for capturing the unique linguistic nuances and sociocultural elements that influence legal terminology in the country [1, 3, 13, 21]. By aligning NLP resources with local contexts, this initiative aims to mitigate ambiguities and bridge interpretative divides that often arise when legal documents are analyzed from a purely foreign standpoint [2]. This localized approach not only enhances the precision of legal NLP applications but also broadens their usability and relevance across different cultural and legal frameworks. Moreover, this project sought to develop the accompanying NLP pipeline for statute law section retrieval. Various models and techniques were employed and evaluated within each task of the ensemble, i.e., dataset generation and relevant section retrieval. Integrating it into a web application, serving as the third objective of the study, facilitated the collection of essential feedback from its targeted primary users—Philippine professionals and students engaged with the law—via a System Usability Scale (SUS) survey.

2 METHODOLOGY

The methodology for this project adopts a comprehensive strategy to develop a robust section retrieval system for the NIRC. It starts with the assembly of a detailed dataset comprised of questions derived from the NIRC and corresponding relevant sections, done through both manual curation and the generation of synthetic data using advanced Large Language Models (LLMs) fine-tuned to follow specific instructions. This dual approach ensures a rich and varied corpus. The core of the system leverages text embedding models, meticulously fine-tuned to grasp the subtleties of Philippine legal texts, for the alignment of queries with pertinent sections. Subsequently, integration of the finalized system into a web application and its beta testing provides insights as to how utilizing NLP to tackle the distinct challenges presented by legal documents can create a retrieval system that excels in accuracy, computational efficiency, and usability.

2.1 Datasets

The dataset comprised three key components: (1) questions based on the NIRC updated as of 2022, (2) relevant section number, and (3) the text of that relevant section, ensuring each question is matched to only one relevant section.

2.1.1 Manual Annotation for Original Dataset. The original dataset drew its content from reputable Philippine bar exam review materials from 1994 to 2022. Each item was thoroughly checked against the NIRC for validity. To address limitations in quantity and enhance the dataset's quality, strategies such as negation and paraphrasing were employed, enriching the dataset while mitigating any potential biases [2]. Furthermore, to maintain a focus on the universal applicability, specific names in situational questions were substituted with generic identifiers.

2.1.2 Large Language Models for Generation of Synthetic Dataset. In the interest of further expanding the dataset to encourage diversity and better representation of the NIRC, the utilization of Large Language Models (LLMs) to generate a synthetic dataset was explored. The comparison of model capabilities and constraints shown in Table 1 informed the selection process. The Mistral-7B-Instruct-v0.2 model [7], referred to as Mistral, was chosen for its high performance on the MT-bench [26] and MMLU benchmark [6], indicating strong capabilities in multilingual translation and understanding across a range of subjects—key qualities for creating a diverse legal dataset.

Additional models within the computational scope were Llama-7-2b-chat [23] and Zephyr-7b-beta [24], both compatible with the hardware limits of Google Colab's A100 GPU with 40 GB of VRAM. Zephyr is trained on the Mistral 7B model, ensuring a high level of performance similar to Mistral. Despite their potential, Mistral's top-tier performance on benchmarks central to the project's goals ultimately solidified its selection. In contrast, the larger SOLAR-10.7B-Instruct-v1.0 model [9] was not considered further due to exceeding the hardware's computational capabilities. Thus, Mistral was chosen, balancing advanced language generation with the available resources to effectively enhance the dataset with synthetic legal text.

The synthetic dataset generation process harnessed the instruction-tuning capability of the Mistral model. Each section of the NIRC was inputted into the model sequentially, with a directive to craft five distinct questions based on the content of the section, along with corresponding answers. Additional instructions were given to prevent leakage of section numbers into the questions and to format the output as "Q: [Question] A: [Answer]". This structured prompt was crucial for the model to generate the desired output. Post-generation, regular expressions (RegEx) were employed to meticulously extract the data and store it in a CSV file. This streamlines the process of dataset creation and ensures ease of integration into the retrieval system. Fifty randomly sampled entries were sent to a lawyer who deemed it to be an adequate reflection of the NIRC.

Table 1: Open-source Large Language Model Comparison for Dataset Generation

Model	No. of Parameters	MT Bench	MMLU
Llama-2-7b-chat	6.74B	6.27	45.80
Mistral-7B-Instruct-v0.2	7.24B	7.60	60.80
Zephyr-7b-beta	7.24B	7.20	52.70
SOLAR-10.7B-Instruct-v1.0	10.70B	7.58	66.20

2.1.3 Dataset Splitting. The project created six distinct sets—three for training and three for testing. The original dataset underwent a stratified train-test split. A parallel process was applied to the synthetic dataset. The last pair of datasets involved combining the original and synthetic datasets prior to the stratified train-test split. Impacts of the LLM-fabricated data inclusion, and the placement of relevant section redistribution within the dataset preparation pipeline, were assessed.

2.2 Section Retrieval Models

The methodology for section retrieval from the NIRC considered several leading embedding models as shown in Table 2, evaluating them on key aspects such as sequence length handling, performance on benchmarks, and accessibility. BERT [4] set a foundational standard for deep bidirectional representation, while SBERT [19] refined this for sentence-level embeddings, and DPR [8] focused on dense passage retrieval. Despite their advancements, these models are constrained by a token limit of 512, which is insufficient for the detailed sections of the NIRC. Cohere AI's cohere-embed-english-v3.0 model and OpenAI's text-embedding offerings were also reviewed. They exhibit an ability to handle longer contexts, a necessity for the intricate legislative text. However, their proprietary and paywalled API-only nature limits their practicality for extended use, especially considering the potential scale of section retrieval tasks for the NIRC.

The jina-embeddings-v2-base-en model (herein referred to as Jina V2) emerges as a strong contender, primarily due to its impressive token limit of 8192 [5]. This is particularly relevant given the dense nature of legal documents; for instance, the longest section of the NIRC comprises 6390 words, resulting in 8060 tokens upon processing, as shown in Table 3.

In addition to Jina V2, the Dense Passage Retrieval (DPR) model was also considered for the premise of its specialization in retrieving long passages given short queries. For sections that exceeded its token limit of 512, a chunking approach was employed, dividing the text into overlapping chunks to maintain context continuity. These chunks were individually processed through DPR, and the resultant embeddings were aggregated to compute a representative mean for the entire section. For DPR, embeddings are derived from a small number of questions and passages using a dual-encoder framework, utilizing separate BERT-based models for processing both queries and context. This technique balances the model's token constraints with the need to preserve the depth and meaning of the legal text, making DPR a valuable addition to the testing framework.

The decision to explore BERT and its derivatives for section retrieval within the NIRC was informed by their notable success in legal AI competitions such as the Competition on Legal Information Extraction/Entailment (COLIEE) [10, 15–17] and Automated Legal Question Answering Competition (ALQAC) [22], as well as research findings outside of these competitions [3]. These sources collectively underscore the utility of BERT-based models in parsing and understanding complex legal texts, a capability that aligns with the project's objectives. Moreover, recognizing that the legal AI field may not fully leverage the latest NLP and deep learning advancements, there was a deliberate effort to integrate more recent NLP technologies. This approach aims to harness underutilized deep learning advancements, potentially enhancing the performance and sophistication of legal document analysis beyond current legal AI applications.

Moreover, Jina V2's architecture, which includes mechanisms such as Attention with Linear Biases and Gated Linear Units, provides a robust framework for the nuances of legislative language processing. The pre-training on the "Colossal Cleaned Common Crawl (C4)" dataset and further fine-tuning ensures that the model is well-adjusted to English, suitable for the language of the NIRC. The model's capabilities are further validated by its performance on benchmarks like MTEB and LoCo, indicating significantly high results in embedding-related tasks, including long-context document handling. The high scores suggest that Jina V2 can create embeddings that capture deep semantic meanings, essential for accurately matching queries to the relevant sections of the NIRC.

The MTEB (Massive Text Embedding Benchmark) [12] is a multifaceted benchmarking suite that specifically gauges the performance of text embedding models on a variety of retrieval-related tasks. It encompasses a broad spectrum of challenges, including document retrieval, question answering, and semantic search, among others. The aim is to measure how effectively a model can understand and match query intent with relevant text content from a large dataset, a crucial capability for information retrieval applications.

The LoCo (Long Context) benchmark complements this by testing a model's proficiency in managing and interpreting lengthy text inputs. It is especially pertinent for evaluating a model's performance on extended passages, which is a common characteristic of legal documents like the NIRC. LoCo assesses whether a model can maintain its performance when dealing with long sequences, which is essential for the retrieval of comprehensive sections that contain the nuanced information required to accurately respond to complex queries.

In selecting the Jina V2 model for section retrieval within the NIRC, MTEB's focus on retrieval-related tasks and LoCo's focus on tasks for longer contexts provide a strong foundation for its appropriateness. These ensure that Jina V2 not only performs well in general language tasks but also excels in the specific domain of

Table 2: Embedding Model Comparison for Section Retrieval

Model	Sequence Length	MTEB	LoCo	Deployment
bert-base-uncased	512	-	-	Local
all-mpnet-base-v2	512	-	-	Local
dpr-single-nq-base	512	-	-	Local
Cohere embed-english-v3.0	512	64.47	66.60	API only
jina-embeddings-v2-base-en	8192	60.39	85.45	Local
text-embedding-3-small	8191	62.26	82.40	API only
text-embedding-ada-002	8191	60.99	52.70	API only

Table 3: Summary of NIRC Longest Section Lengths

Section Number	Number of Words	Jina V2 Tokens	BGE Reranker Tokens
34	6390	8060	8932
148	3070	4437	4851
144	2489	3286	3699
288	2199	3089	3466
22	2051	2684	2959

long context retrieval, which is central to the project's objectives. The model's high scores in these benchmarks signify its capability to create precise embeddings that facilitate the accurate matching of queries to relevant sections of the text, validating its use for this research.

Training and testing of the models were conducted with distinct approaches for each phase. Training was performed on the different datasets, with the model encoding question-relevant section pairs. The loss was calculated based on one minus the cosine similarity between the embeddings, aiming to minimize the distance between semantically related question and section embeddings. For the testing phase, a retrieval task approach was adopted where section text embeddings were pre-computed, and each question was dynamically encoded. The cosine similarity between the question embedding and all section embeddings was calculated to rank the sections according to relevance; whether the correct relevant section was the top-ranked by the model or not was the basis for calculating the accuracy. This process was facilitated using Google Colab and a V100 GPU to leverage high computational power. The training process was iteratively evaluated over various learning rates with three epochs each to identify the optimal model configuration. Memory constraints necessitated the use of smaller batch sizes, coupled with gradient accumulation techniques, to achieve an effective batch size of 40, which was determined to be ideal for stability [11]. Optimization was carried out using the AdamW optimizer, ensuring efficient training dynamics.

2.3 Deployment

2.3.1 Web Development. The development of the DAKILA web application encapsulates a comprehensive user journey, starting from a straightforward search page facilitated by a navigation bar for easy access to various sections like Search, About Us, and the Tax Code. Even before receiving a query, upon setting up the server, the fine-tuned section retrieval model is pre-loaded, along with its tokenizer and pre-computed NIRC embeddings. Upon receiving a query, the DAKILA algorithm initiates a step-by-step process: first, it encodes this query using the tokenizer and fine-tuned model. Next, it calculates cosine similarity scores between the query embedding and all NIRC section embeddings, ranking them accordingly. The top relevant sections, up to a user-specified number 'n' (defaulting to 5), are then displayed to the user. This system, developed with Flask and utilizing pre-computed embeddings, ensures efficient and consistent retrieval of legal sections. This architecture ensures that user inputs are precisely processed, returning the most relevant legal sections, thereby showcasing the potential of advanced NLP techniques in making legal information more accessible and navigable not only for professionals in law and accountancy but also for the average layperson.

2.3.2 System Usability Scale Survey. Deployment of the DAKILA web application was conducted, with user testing providing practical insights about its functionality and user interface. Volunteers from law and accountancy fields interacted with the application. Their experiences were quantitatively measured through the System Usability Scale (SUS) survey accompanied by an additional Likert scale question inquiring about the perceived quality of its section retrieval output. The questions asks if the user prefers DAKILA over existing legal research tools, such as Google, CD Asia, and LawPhil, when navigating the NIRC. Through the SUS survey, we aim to evaluate the DAKILA web application's ease of use, efficiency, and overall satisfaction in comparison to existing legal research platforms. Overall, the survey seeks to understand the application's effectiveness in streamlining the search and retrieval process for tax laws, focusing on user-friendly navigation and the practical utility of the tool. Additionally, it will help highlight areas for improvement, ensuring that DAKILA meets the evolving needs of legal professionals and students by providing a more intuitive and resourceful alternative to traditional legal research methods.

3 RESULTS AND DISCUSSION

This section details experimentation results leading up to the determination of the best NLP pipeline for the purposes of this study. The Jina-V2 model trained on the training dataset combined prior to the stratified train-test split, at a learning rate of 1e-06 yielded the highest test accuracy score.

Table 4: Dataset Samples

Relevant Section Number	Dataset	Query
259	Original	Do you need a license to collect foreign payments?
172	Original	Is there a time limit on how long a package can be detained by a revenue officer without legal proceedings?
107	Original	Is an importer of flowers from abroad in 2011 liable for VAT?
60	Original	What are the various trusts subject to income tax?
24	Original	Is the 6% final tax rate imposed on all sales of real property classified as capital assets?
151	Synthetic	What is the definition of 'gross output' in the context of mineral taxation according to the Philippine Tax Code?
147	Synthetic	What are 'heated tobacco products' and how are they different from 'vapor products' as defined in this tax code?
183	Synthetic	What is the stamp tax for a life insurance policy with a coverage of PHP 1.2 million?
139	Synthetic	What happens to fermented liquor that is unfit for consumption due to damage?
283	Synthetic	How much of the excess collections from certain national taxes are distributed to local governments and how much is kept by the National Government?

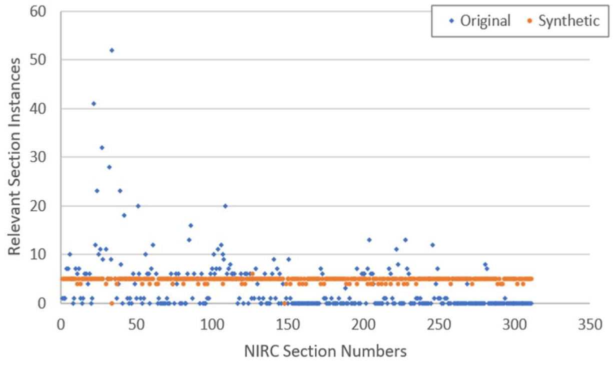

Figure 2: Relevant NIRC Section Distribution in the Original and Synthetic Datasets

3.1 Datasets

The original dataset's composition hinged on the availability of bar exam reviewers. Figure 2 presents the original dataset's skewed distribution of the NIRC sections, highlighting areas deemed crucial for legal practice. This skew may align with common search queries but does not fully reflect the breadth of the NIRC. To address potential gaps in coverage, the Mistral 7B-generated dataset aimed to create a balanced representation, with a law professional confirming the synthetic dataset's quality. Table 4 shows samples from both datasets.

The combined datasets, inclusive of both original and synthetically generated data, were foundational to the superior performance of the Jina V2 model, despite the extended training time required as seen in Tables 5 and 7. A strategic approach to combining the two datasets and partitioning with stratification for training and testing purposes contributed to a modest yet noteworthy improvement in model accuracy. This performance uptick underscores the value of a more diverse and representative dataset that ensures entries from each NIRC section.

Table 5: Summary of Datasets

Dataset	Training	Testing	Average Training Time (min)	
	90%	10%	DPR	Jina V2
Original	918	102	3.10	4.43
Synthetic	1333	149	2.13	6.42
Combined	2251	251[a]	5.18	10.87

[a] This is comprised of 97 entries from the original dataset and 154 from the synthetic dataset.

3.2 Section Retrieval Models

3.2.1 Training. In determining the best preprocessing techniques for both DPR and Jina V2 models, various methods were scrutinized individually for their impact on accuracy using the combined test set. For the DPR model, lowercasing, removal of footnotes, removal of special characters, and lemmatization emerged as beneficial, leading to improvements in accuracy. Contrastingly, the Jina V2 model performed optimally without any preprocessing. With these insights, the training and testing pipelines for each model were implemented accordingly. Experimentation with data partitioning showed that a 90:10 train-test split was optimal for the project needs.

Figure 3 showcases a plot of learning rate versus accuracy across three epochs for Jina V2 on the combined dataset, revealing that a learning rate of 1e-06 was optimal, with higher epochs generally leading to reduced accuracy except at very slow learning rates. DPR and Jina V2 trained on the other variations of the dataset showed their peaks at different learning rates, which are shown in Tables 6 and 7, however, the observed decline in accuracy with

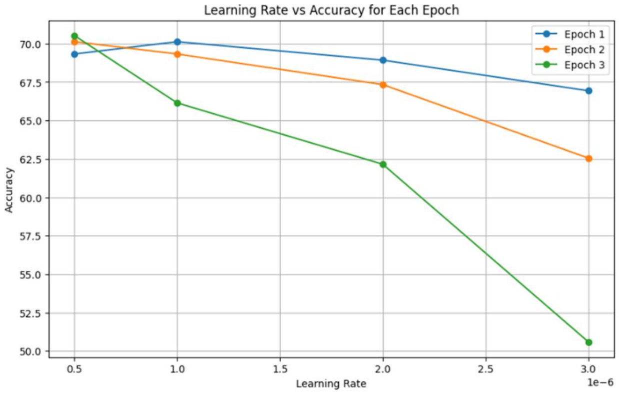

Figure 3: Learning Rate vs Accuracy for Each Epoch

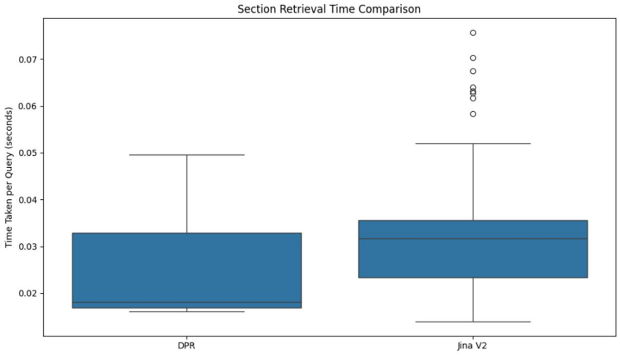

Figure 4: Testing Time per Query for Section Retrieval Models

an increase in the number of epochs ultimately holds true. The decrease in accuracy with additional epochs suggests that the models' pre-existing language comprehension could be compromised by excessive fine-tuning, leading to overfitting.

To preserve the innate capabilities of the models and optimize accuracy, the final training iterations were conducted for a single epoch, without a validation set, thereby utilizing the full scope of the training data. Typically, a validation set is utilized between epochs during model training to monitor and adjust the model's performance on data not seen during training. This intermediate evaluation helps in fine-tuning model parameters and preventing overfitting, ensuring the model generalizes well to new data. Given that training for a single epoch yielded the best results, the decision to bypass the validation set was made, eliminating the need for validation steps between epochs. This approach streamlined the training process, focusing on maximizing the efficiency and accuracy of the model.

3.2.2 Testing. For testing of the models fine-tuned on the combined training dataset, the combined testing dataset questions were first labeled as being either from the original or synthetic dataset and sorted accordingly. However, for testing the models fine-tuned on both the original and synthetic training datasets, the original and synthetic testing datasets were merged for the combined testing dataset. This was to ensure no overlap between the training and testing datasets.

Table 6 surmises that the DPR model attained its highest test accuracy of 34.02% when trained on the combined training dataset with a learning rate of 6e-07. This is significantly surpassed by Jina V2's performance shown in Table 7, regardless of the testing dataset, when fine-tuned on the combined training dataset, sporting the best test accuracy scores of 63.92% on the original testing dataset, 74.68% on the synthetic testing dataset, and 70.52% on the combined testing dataset. This suggests a heightened model performance with the integration of the synthetic dataset. This substantial difference justifies the tradeoff in choosing the latter as the final section retrieval model despite its longer testing times documented in Figure 4.

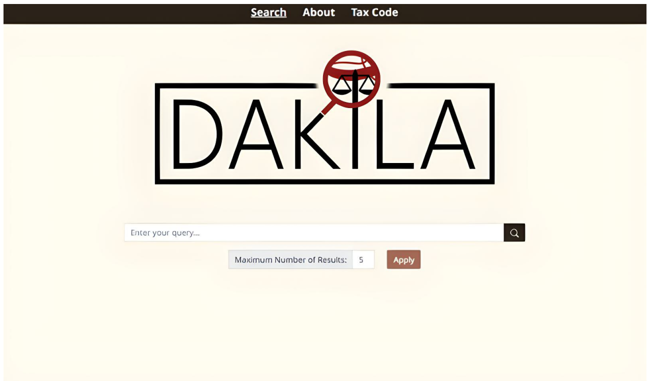

Figure 5: DAKILA Web Application

3.3 Deployment

3.3.1 Comparison to Existing Resources. The comparison between DAKILA and existing legal resources for statute law in the Philippines underscored DAKILA's advanced capabilities. Other resources often lacked the contextual understanding DAKILA offered, relying on word frequencies and direct string matches that could miss relevant information or introduce irrelevant results due to their inability to interpret the nuances of legal queries. DAKILA's semantic search, coupled with features like a higher character limit for queries and direct navigation to pertinent NIRC sections, addressed these drawbacks effectively. This not only improved the precision of search results but also enhanced user experience by streamlining access to specific legal information.

3.3.2 System Usability Scale Survey. The System Usability Scale (SUS) employed in evaluating DAKILA utilized a 5-scale rating, where participants rated their agreement with a series of statements. Positive statements were assigned to odd-numbered questions, with 5 indicating strong agreement, reflective of a favorable view. Conversely, even-numbered questions were framed negatively, where a score of 1 represented the best outcome, indicating a lack of issues or challenges. This dual-phrased approach provided a balanced

Table 6: Test Accuracy Scores for DPR Fine-tuning

		Testing Dataset		
Learning Rate	Training Dataset	Original	Synthetic	Combined
3e-06	Original	30.39%	20.81%	24.70%
1e-06	Synthetic	28.43%	23.49%	25.50%
6e-07	Combined	34.02%	20.13%	25.50%
Untrained		22.55%	21.48%	20.72%

Table 7: Test Accuracy Scores for Jina V2 Fine-tuning

		Testing Dataset		
Learning Rate	Training Dataset	Original	Synthetic	Combined
3e-06	Original	59.80%	59.06%	59.36%
7e-06	Synthetic	54.90%	59.06%	57.37%
1e-06	Combined	63.92%	74.68%	70.52%
Untrained		51.96%	57.72%	55.38%

Table 8: SUS Survey Results

No.	SUS Questions	Mean	SD
1	I think that I would like to use this system frequently.	4.57	0.56
2	I found the system unnecessarily complex.	1.90	1.22
3	I thought the system was easy to use.	4.70	0.64
4	I think that I would need the support of a technical person to be able to use this system.	1.47	0.72
5	I found the various functions in this system were well integrated.	4.67	0.54
6	I thought there was too much inconsistency in this system.	1.60	0.66
7	I would imagine that most people would learn to use this system very quickly.	4.73	0.51
8	I found the system very cumbersome to use.	1.80	1.19
9	I felt very confident using the system.	4.63	0.55
10	I needed to learn a lot of things before I could get going with this system.	1.83	1.04

measure of the system's usability from the perspectives of ease of use and potential frustrations. The survey results for DAKILA, as shown in Table 8, demonstrated strong user satisfaction across diverse demographics. Participants ranged from under 20 to over 60 years old, with the largest group being those between 21-29 years old. The professional background of respondents spanned law and accountancy, including both students, and practicing professionals. High SUS scores across all age and professional categories emphasized the system's ease of use and functionality. This broad approval showcases DAKILA's potential as a specialized tool for legal research, offering a user-centric alternative to traditional resources.

4 CONCLUSION AND RECOMMENDATIONS

The investigation into the section retrieval of the NIRC culminated a deeper understanding of the interplay between dataset quality and model training. The uptick in performance upon the introduction of the LLM-generated synthetic dataset offered a more diverse and representative dataset compared to the original, which sported a skewed section distribution. The study also confirmed a heightened effectiveness of section retrieval with slower learning rates and less training epochs to avoid diminishing the embedding model's pre-existing language comprehension due to detrimental overfitting because of excessive fine-tuning. Amidst the evaluation of multiple models Jina V2 outputted the highest test accuracy scores, showcasing the capabilities of advanced language models without the need for pre-processing steps. Overall, the relationship between careful dataset construction and targeted model training strategies for improved legal search tools is essential.

Future developments of NLP-driven Philippine statute law retrieval systems should explore appropriate annotations, methodologies, and metrics in dealing with queries with multiple relevant sections. Although positive datasets are more common, such as what was practiced in this study, incorporation of negative datasets in training may provide more valuable insights into semantic relevance. With the recent exponential rise in caliber and quantity of released LLMs arises the opportunity to expound on the effectiveness of automated dataset generation to alleviate the heavy manual labor of dataset creation and annotation while only minimally compromising quality. This addition to the ensemble should be tested across a multitude of legal documents in varying sizes and domains. Finally, reranking methods should be explored to further refine the ensemble and improve its accuracy.

ACKNOWLEDGMENTS

We would like to thank the law professionals who dedicated their service without charge to our endeavors.

REFERENCES

[1] Intisar Almuslim and Diana Inkpen. 2022. Legal judgment prediction for canadian appeal cases. In *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*. IEEE, 163–168.

[2] Dang Hoang Anh, Dinh-Truong Do, Vu Tran, and Nguyen Le Minh. 2023. The impact of large language modeling on natural language processing in legal texts: a comprehensive survey. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 1–7.

[3] Chieu-Nguyen Chau, Truong-Son Nguyen, and Le-Minh Nguyen. 2020. Vnlawbert: A vietnamese legal answer selection approach using bert language model. In *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*. IEEE, 298–301.

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

[5] Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, et al. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923* (2023).

[6] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

[7] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[8] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[9] Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166* (2023).

[10] Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2022. COLIEE 2022 summary: methods for legal document retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*. Springer, 51–67.

[11] Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. *arXiv preprint arXiv:2104.12741* (2021).

[12] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316* (2022).

[13] Varsha Naik, Rajeswari Kannan, Sanket Agarwal, Aryan Sable, and Himanshu Chaudhari. 2023. An Effective Search Algorithm for Analyzing and Extracting Indian Legal Judgments using NER and Document Summarization. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*. IEEE, 1–6.

[14] Elmer Peramo, Charibeth Cheng, and Macario Cordel. 2021. Juris2vec: Building Word Embeddings from Philippine Jurisprudence. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 121–125.

[15] Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. *The Review of Socionetwork Strategies* 16, 1 (2022), 111–133.

[16] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2020. A summary of the COLIEE 2019 competition. In *New Frontiers in Artificial Intelligence: JSAI-isAI International Workshops, JURISIN, AI-Biz, LENLS, Kansei-AI, Yokohama, Japan, November 10–12, 2019, Revised Selected Papers 10*. Springer, 34–49.

[17] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2021. COLIEE 2020: methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers 12*. Springer, 196–210.

[18] Lorenz Timothy Barco Ranera, Geoffrey A Solano, and Nathaniel Oco. 2019. Retrieval of semantically similar Philippine supreme court case decisions using Doc2Vec. In *2019 International Symposium on Multimedia and Communication Technology (ISMAC)*. IEEE, 1–6.

[19] N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084* (2019).

[20] Samarth Singhal, Siddhant Singh, Sandeep Yadav, and Anil Singh Parihar. 2023. LTSum: Legal Text Summarizer. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 1–6.

[21] Oussama Tahtah, Yassine Akhiat, Ahmed Zinedine, and Khalid Fardousse. 2023. Towards a Question/Answering System in Moroccan Legal Domain: data preparation and question classification phase using ML approaches. In *2023 7th IEEE Congress on Information Science and Technology (CiSt)*. IEEE, 140–144.

[22] Nguyen Ha Thanh, Bui Minh Quan, Chau Nguyen, Tung Le, Nguyen Minh Phuong, Dang Tran Binh, Vuong Thi Hai Yen, Teeradaj Racharak, Nguyen Le Minh, Tran Duc Vu, et al. 2021. A summary of the alqac 2021 competition. In *2021 13th international conference on knowledge and systems engineering (kse)*. IEEE, 1–5.

[23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[24] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944* (2023).

[25] Michael Benedict L Virtucio, Jeffrey A Aborot, John Kevin C Abonita, Roxanne S Avinante, Rother Jay B Copino, Michelle P Neverida, Vanesa O Osiana, Elmer C Peramo, Joanna G Syjuco, and Glenn Brian A Tan. 2018. Predicting decisions of the philippine supreme court using natural language processing and machine learning. In *2018 IEEE 42nd annual computer software and applications conference (COMPSAC)*, Vol. 2. IEEE, 130–135.

[26] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.