Open Law Philippines: Legal Document Retrieval Analysis

Andres Clemente
College of Computer Studies
De La Salle University Manila
Manila, Philippines
andres_clemente@dlsu.edu.ph

Priscilla Licup
College of Computer Studies
De La Salle University Manila
Manila, Philippines
priscilla_licup@dlsu.edu.ph

Kenn Villarama
College of Computer Studies
De La Salle University Manila
Manila, Philippines
kenn_michael_villarama@dlsu.edu.ph

Joshua Permito
College of Computer Studies
De La Salle University Manila
Manila, Philippines
joshua_permito@dlsu.edu.ph

Ann Laguna
College of Computer Studies
De La Salle University Manila
Manila, Philippines
ann.laguna@dlsu.edu.ph

Donnald Miguel Robles
College of Computer Studies
De La Salle University Manila
Manila, Philippines
donnald_robles@dlsu.edu.ph

ABSTRACT

Due to the vast amount of legal document data, understanding and organizing law-related documents can be challenging, particularly for those outside the legal field. Initiatives such as Harvard's Case Law project have made strides in simplifying this process through a web application and relevant data visualization tools. This study proposes developing a novel document retrieval system tailored to the nuances of Philippine law, leveraging advanced deep learning techniques. Recent advancements in natural language processing, particularly models like Juris2vec, have shown promise in legal text analysis. Alongside these developments, the emergence of transformer models such as LEGAL-BERT, specifically pre-trained and fine-tuned for the legal domain, further enhances our ability to process legal documents accurately. By integrating BERTopic for thematic organization/filtering and Sentence-BERT (SBERT) for semantic document search, and in collaboration with legal experts, our project aims to significantly enhance the accessibility and comprehension of legal documents for a diverse range of users, including legal practitioners, scholars and the public. This endeavor not only promises to bridge the gap in legal document retrieval but also to pioneer the application of these sophisticated models in the context of Philippine law, setting a precedent for future legal informatics research.

KEYWORDS

Legal document retrieval, legal informatics, topic modeling, document similarity, Sentence-BERT (SBERT), BERTopic

1 INTRODUCTION

1.1 Overview

Legal documents, including court cases, are considered public property; however, not all are readily accessible to the public. Some of these documents are only physically available, requiring a visit to an office and incurring reproduction expenses. In this digital era, the importance of making public documents accessible online and free of charge cannot be overstated. While efforts to standardize legal documents have begun on a global scale, Philippine law repositories still lack the necessary tools for efficient document search and analysis. Locating relevant documents typically involves sifting through numerous files to identify those pertinent to a specific case or topic. To address these challenges, implementing Natural Language Processing (NLP) techniques such as topic modeling and document similarity can significantly enhance the relevance assessment of each document.

There is a current lack of efficiency when it comes to the retrieval of Philippine legal documents. In order to provide a way to harness the complexity of these documents, NLP tasks such as topic modeling and document similarity, as well as the involvement of Large Language Models like Sentence-BERT (SBERT) and topic modeling techniques like BERTopic, should be implemented to properly organize the contents of legal documents.

1.2 Objectives

The study aims to create a document retrieval system for Philippine legal documents using the transformer models based on topic and semantic similarity. More specifically, the study's objectives tackle on the following:

(1) To cluster Philippine legal documents into relevant topics using BERTopic and implement topic-based filtering for a document retrieval system.
(2) To implement a semantic retrieval system for Philippine legal documents using Sentence-BERT (SBERT).
(3) To implement a user-interface that would allow users to visualize similarities as well as retrieve relevant Philippine legal documents given a specific query.

1.3 Scope & Limitations

This study focuses on nationally recognized Philippine legal documents such as Philippine Supreme Court Decisions, Republic Acts, Senate Bills, Presidential Proclamations, and other nationally recognized legal materials. Regional documents may be considered for inclusion in later phases of the project. This initial constraint is considered sufficient for the preliminary analysis because regional cases typically do not establish jurisprudence at the national level. While regional cases are important for understanding the application of laws in specific geographic areas, they typically do not carry the same weight in terms of setting precedent and shaping the legal landscape on a national scale. Focusing on Philippine legal documents will allow for a more comprehensive and focused initial analysis, helping to understand the overarching legal framework and key legal principles that apply to the entire country. Due to the significant advancements of specific NLP techniques with regard to transformers, a focus on SBERT embeddings would be implemented in the study. This study will be especially beneficial for future researchers who are interested in the legal domain of the Philippines. However, several challenges that must be acknowledged include the computational requirements of advanced NLP models like SBERT and BERTopic as they demand computational power, especially with large datasets. The system will also handle sensitive legal documents so data privacy measures and protocols must be implemented to protect data integrity and confidentiality.

1.4 Significance

The study is essential because it can potentially spur numerous implications concerning the retrieval and interpretation of Philippine legal documents through Large Language Models. Moreover, it could also dictate the evolution and current condition of the Philippine legal system based on the data gathered by the researchers. Overall, the study holds significance for the NLP community by advancing techniques tailored for the legal domain in the Philippines, offering a system for enhanced legal research and document analysis within its unique context.

2 RELATED WORK

We present the related studies that discuss document similarity and topic modeling.

2.1 Document Similarity

The evolution of document similarity analysis in legal documents has seen significant progress through the adoption of advanced NLP technologies, transitioning from traditional techniques such as TF-IDF to more sophisticated methods including Word2Vec and transformer models like BERT and JurisBERT. The advent of sentence transformer models, exemplified by Sentence-BERT (SBERT), marks a substantial leap in overcoming previous computational hurdles, offering notable improvements in processing speed and accuracy for semantic textual similarity tasks. SBERT, in particular, demonstrates a considerable enhancement in computational efficiency over traditional methods, facilitating more practical applications in demanding NLP tasks [1]. JurisBERT represents a notable advance in domain-specific modeling, achieving significant gains in precision and training efficiency over multilingual BERT and BERTimbau for legal text analysis [2]. This model's development emphasizes the growing focus on tailoring NLP technologies to specific fields, enhancing their applicability and effectiveness in real-world scenarios. Further investigations into the comparative performance of BERT-like models and traditional methods in semantic similarity highlight the potential of tailored embedding strategies. A study delving into Russian news similarity detection emphasizes the advantages of in-domain pre-training and fine-tuning on specialized datasets [3]. Despite the computational demands, such as the slow GPU performance noted with SBERT-WK's QR matrix decomposition, these advancements contribute critical insights into optimizing NLP models for specific content analysis, pointing towards the ongoing refinement of document similarity approaches within the legal domain and beyond. Additionally, a noteworthy development in retrieval tasks is the introduction of the Relevance Score Proportional to Relevance (RPRS) [4]. This efficiently leverages SBERT bi-encoders for comprehensive text coverage without memory constraints, showing significant potential in legal domain retrieval tasks as they were also tested using the COLIEE 2021 dataset which is drawn from an existing collection of predominantly Federal Court of Canada case law. Consequently, empirical studies demonstrate that SBERT significantly outperforms traditional methods like TF-IDF, GloVe embeddings, InferSent, and Universal Sentence Encoder, and baseline BERT models in terms of accuracy, efficiency, and scalability [1] [5]. Not only does SBERT achieve higher MAP scores and better Spearman correlation values in semantic similarity tasks, but it also drastically reduces computational overhead, enabling the processing of large datasets in seconds rather than hours [1] [6].

2.2 Topic Modeling

Topic modeling has evolved significantly with the shift from classical models like LDA and NMF to advanced text embedding techniques using BERT variants, improving thematic discovery within documents by capturing contextual nuances. The introduction of BERTopic marked a significant advancement, utilizing BERT for embeddings, HDBSCAN for clustering, and class-based TF-IDF for topic prediction, enhancing interpretability and relevance of identified topics [7]. Despite its assumption of a single topic per document, BERTopic's approach to topic modeling has shown promise, especially within the legal domain, as seen with LEGAL-BERT [8]. This model, tailored for legal texts, demonstrated high accuracy in capturing the essence of complex legal documents, suggesting potential for automating legal document summarization. Additionally, BERTopic's effectiveness across domains was highlighted in a study, showing its adaptability and robustness, including in multilingual contexts, compared to other models like Top2Vec and classical approaches, which struggle with context and relationships between topics [9]. This evolution in topic modeling techniques represents a leap forward in extracting meaningful insights from vast text corpora, offering a more nuanced understanding of document themes.

3 OPEN LAW PHILIPPINES

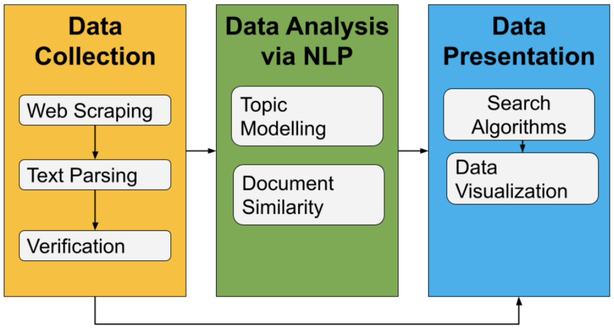

Figure 1: Conceptual Framework of the Study.

Open Law Philippines is the web application which would incorporate the document retrieval system for legal documents. The development of the study follows a 3-stage pipeline namely data collection, data analysis via NLP, and data presentation (Figure 1).

Each aspect of the framework will be discussed thoroughly in the following sections. As of this moment, the planned output of the Open Law Philippines study is a document retrieval system that incorporates SBERT embeddings for semantic search as well as assigning topics for finding and filtering relevant documents through a BERTopic model.

3.1 Data Collection

For this study, the data will be sourced from government websites in the Philippines, focusing on documents from the Official Gazette, Supreme Court Rulings, and House and Senate Bills, among others. The official gazette encompasses a variety of documents, including Executive Issuances, Presidential Speeches, Proclamations, and more. Legislative data includes senate bills, house bills, resolutions, journals, committee reports, republic acts, and treaties. Judicial data comprises Supreme Court Decisions, Resolutions, Rules of Court, and other related documents. This collection will be initially constrained to national documents, such as Legislative Acts, Republic Acts, Commonwealth Acts, Batas Pambata and Philippine Supreme Court Decisions. Regional documents will be excluded from this initial analysis as regional cases do not establish jurisprudence. Since web scraping may still have errors and inaccuracies, human verification will still be conducted after the study's methodology to ensure the veracity of data. Every data that has been verified by a human transcriber shall be noted. Due to the large amount of data that has to be collected, this system shall be implemented randomly. A legal expert would also provide their feedback on the effectiveness of the User Interface.

3.2 System Architecture

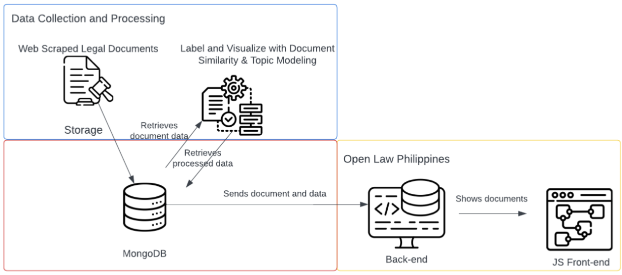

Figure 2: Diagram of Open Law Philippines Architecture.

Figure 2 shows the system architecture which encompasses data collection, processing, and storage of processed web-scraped data. Additionally, it involves document retrieval through the Open Law Philippines web application. MongoDB is chosen to store prelabeled document data and processed documents for document similarity. For the Open Law Philippines back-end, Express will serve as the API responsible for querying and saving legal expert inputs and user corrections to the database. The back-end server primarily connects the frontend application to the MongoDB database, facilitating the retrieval of pre-labeled document data and supporting data visualizations in the study. In the frontend, ReactJS, known for its component-based UI, will be used for convenient programming. Additionally, ReactJS will be utilized together with NextJS, simplifying programming with built-in functions instead of custom ones.

3.3 Target Visualization

For this study, the focus of target visualizations has been narrowed down to dimensionality reduction scatter plots, BERTopic-specific visualizations, and word clouds. These chosen visualizations will serve as essential references for conveying the NLP techniques conducted in the study. Figure 3 shows a word cloud sample from using BERTopic. Also, the deliberate selection of dimensionality reduction scatter plots (see left of Figure 4), and BERTopic-specific visualizations such as distance maps (see right of Figure 4) and bar charts (see Figure 5) aim to provide a highly relevant visualization for the included processes in the system. Additionally, the incorporation of word clouds adds a textual dimension, highlighting key terms and patterns within the collection of legal documents. Conversely, scatter plots of dimensionality reduction as an implemented visualization for document similarity allow for better interpretation due to the reduction of high-dimensional document feature vectors through spatial distribution that represents similarities between the corpus of documents. These visualizations collectively contribute to enhancing the clarity of findings in the study.

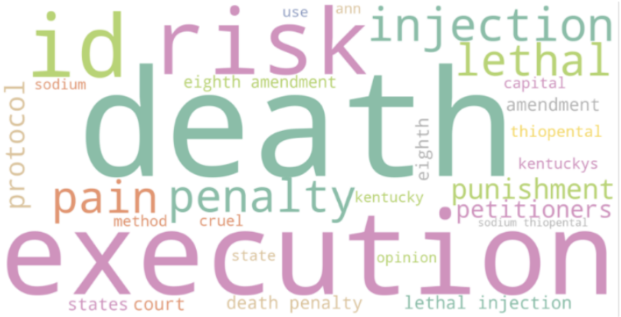

Figure 3: BERTopic's Most Relevant Words for a Specific Topic used in Legal Documents.

3.4 Document Similarity & Topic Modeling

For tasks related to document similarity and topic modeling, SentenceTransformers (SBERT), which is fine-tuned for semantic search, will be used as the model, while the class-based c-TF-IDF will be utilized for topic selection in topic modeling. A variation of techniques such as tokenizing, dimensional reduction and clustering will also be implemented whenever applicable in order to handle the large amount of data to be assessed for topic modeling. For document similarity, a semantic search function will be used using Approximate Nearest Neighbors (ANN) because it offers a fast and efficient way to find the most relevant documents within a large

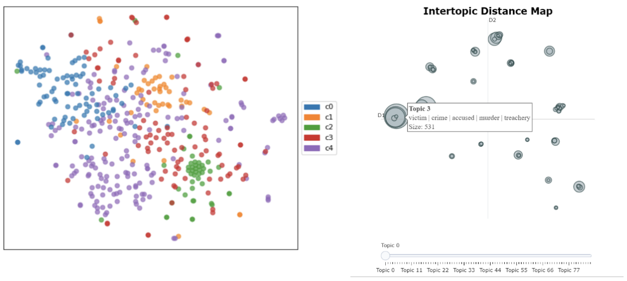

Figure 4: t-SNE Projection and Intertopic Distance Map.

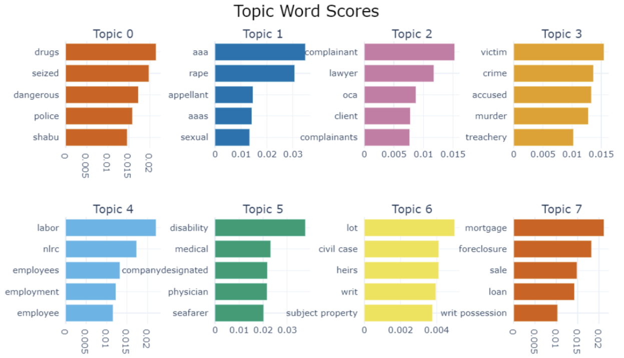

Figure 5: Topic Word Scores Bar Chart in BERTopic using a Subset of the Dataset.

dataset by approximating the nearest neighbors rather than computing the exact distances to all points in the dataset. Additionally, ANN's ability to work with high-dimensional data aligns well with the nature of document embeddings produced by SentenceTransformers, facilitating a more nuanced and accurate semantic search. Meanwhile, a list of top-assigned topics will be generated for topic modeling to provide insights into the prevalent themes within the data, allowing for effective organization, summary, and analysis of large text corpora.

4 FURTHER WORK

In an effort to address the challenges of legal document retrieval and analysis within the Philippine context, our proposal aims to enhance the retrieval and analysis of legal documents in the Philippines by integrating SBERT for semantic search and BERTopic for thematic organization. This combination is expected to improve the accessibility and usability of legal texts for professionals, scholars, and the public, using advanced natural language processing techniques to address gaps in legal informatics. Future plans include evaluating the BERTopic model with tools like OCTIS, integrating Elasticsearch with SBERT for better search capabilities, expanding the document repository to include regional texts, and supporting multiple local languages to cater to the Philippines' linguistic diversity. This ongoing work demonstrates our commitment to using cutting-edge strategies to make legal information more widely accessible.

REFERENCES

[1] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, nov 2019. Association for Computational Linguistics.

[2] Charles F. O. Viegas, Bruno C. Costa, and Renato P. Ishii. Jurisbert: A new approach that converts a classification corpus into an sts one. In Osvaldo Gervasi, Beniamino Murgante, David Taniar, Bernady O. Apduhan, Ana Cristina Braga, Chiara Garau, and Anastasia Stratigea, editors, *Computational Science and Its Applications – ICCSA 2023*, pages 349–365, Cham, 2023. Springer Nature Switzerland.

[3] A. Vatolin, E. Smirnova, and S. Shkarin. Russian news similarity detection with sbert: pre-training and fine-tuning. pages 692–697, 06 2021.

[4] Arian Askari, Suzan Verberne, Amin Abolghasemi, Wessel Kraaij, and Gabriella Pasi. Retrieval for extremely long queries and documents with rprs: A highly efficient and effective transformer-based re-ranker. *ACM Trans. Inf. Syst.*, 42(5), apr 2024.

[5] Jacob Malmberg. *Evaluating semantic similarity using sentence embeddings*. PhD thesis, 2021.

[6] Khushboo Taneja, Jyoti Vashishtha, and Saroj Ratnoo. Efficient deep pre-trained sentence embedding model for similarity search. pages 605–615, 09 2023.

[7] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

[8] Raquel Silveira, Carlos Gustavo Fernandes, Joao Araujo Monteiro Neto, Vasco Furtado, and J. Ernesto Pimentel Filho. Topic modelling of legal documents via legal-bert, Aug 2023.

[9] Roman Egger and Joanne Yu. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7, 2022.