Enhancing legal research through knowledge-infused information retrieval for Vietnamese labor law

Vuong Pham[1,3,4], Hoang Huy Le[1,3,5], Thinh Phu Ngo[1,3], Binh Nguyen[1,4], Diem Nguyen[2,4], Hien D. Nguyen[2,4]

[1]Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam
[2]Faculty of Computer Science, University of Information Technology, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
[4]Institute of Data Science and Artificial Intelligence, Sai Gon University, Ho Chi Minh City, Vietnam
[5]AISIA Research Lab, Ho Chi Minh City, Vietnam

Article Info

Article history:

Received Feb 19, 2024
Revised May 3, 2024
Accepted Jun 1, 2024

Keywords:

Information retrieval
Intelligent system
Legal dataset
Legal documents
Natural language processing

ABSTRACT

The role of intelligent information retrieval systems in legal research optimization has become increasingly recognized. There are many methods for exhibiting advancements in the proficient retrieval of legal documents. However, those methods fail to tackle the specific challenges encountered in real-world labor law searches. This research breaks new ground in Vietnamese labor law retrieval by leveraging a comprehensive dataset of 300,000 documents across diverse categories (20 document types and 27 legal fields) to train and evaluate retrieval models specifically designed for Vietnamese labor law. Unlike previous approaches, this work goes beyond simple information retrieval. It also constructs question & answer (Q&A) dataset specifically tailored to this legal domain. Besides, this study introduces a novel approach of incorporating a legal ontology built from the dataset itself. This knowledge infusion significantly improves retrieval performance across legal search tasks, as demonstrated through rigorous experimentation. These advancements empower intelligent systems to grasp the intricate semantic nuances of Vietnamese labor law.

Corresponding Author:

Hien D. Nguyen
Faculty of Computer Science, University of Information Technology
Ho Chi Minh City, Vietnam
Email: hiennd@uit.edu.vn

1. INTRODUCTION

The increasing of the complexity of legal documents creates a significant barrier for citizens trying to navigate their legal rights and responsibilities [1], [2]. Many individuals lack the time or legal expertise to decipher dense legalese, leaving them feeling powerless when facing legal issues [3], [4]. Traditional legal assistance can be prohibitively expensive, further limiting access to justice. To empower individuals and bridge this knowledge gap, an intelligent search system is required to provide prompt and accurate legal information, directly addressing these critical needs [5], [6]. Intelligent retrieval systems are essential tools that empower legal professionals to navigate massive amounts of legal information efficiently. Recent advances in natural language processing (NLP) and machine learning have paved the way for more sophisticated legal search engines. Nguyen *et al.* [7] proposed a recurrent neural network (RNN) to identify and label essential parts of Japanese legal documents into two kinds. It improved bidirectional-long short term memory (Bi-LSTM)-conditional random field (CRF) to recognize essential components that do not overlap and the multi-layer Bi-LSTM-CRF and

BiLSTM-multi layer perceptron (MLP)-CRF to identify necessary overlapping components. The results of this method outperformed others on the dataset in Japanese national pension law requisite-effectuation recognition.

The machine learning methods are also concerned with predicting court decisions based on legal document retrieval [8]. Besides, pretrained language models like bidirectional encoder representations from transformers (BERT) [9] have achieved state-of-the-art results on legal text retrieval tasks. However, the performance of these intelligent systems depends heavily on the quality and size of the underlying training data. The existing legal dataset used in [10] has limitations that constrain model capabilities. That dataset covers a narrow domain of Vietnamese labor law, with only 23 documents and 618 question-answer pairs. Thus, expanding the diversity and volume of legal texts is necessary to provide more robust training data. Moreover, BERT also captures semantic meaning effectively, there is many room to augment these models with knowledge graphs and ontologies to inject legal domain expertise [11]-[13]. Combining neural networks with structured knowledge sources can potentially improve retrieval accuracy further.

In this paper, a larger-scale Vietnamese legal dataset spanning diverse areas of law is constructed, and the improvement of methods based on legal ontology [14] combining neural networks is proposed to implement on this dataset. This study focuses on Vietnamese labor law [15], one of the popular and vital knowledge domains for employees. Firstly, the standards of legal documents have been studied to crawl and normalise. The structure for organizing the collected documents has been studied to be suitable with the consultancy in the practice. Following the initial processing, the pipeline tackles the challenges of legal documents, often delivered in HTML. After that, the question & answering (Q&A) dataset is gathered, incorporating cite references from the legal corpus in each response. This indicates that the material was consulted on these sites to provide the relevant response. The responses are categorized using multi-labeling or indices, resulting in a list of indices corresponding to the content of each answer. Additionally, a method has been proposed for utilizing sentence transformers in legal documents. Fine-tuning and adjusting the neural network weights are necessary to better align with the Vietnamese legal context and achieve optimal results. The experimental results demonstrate that the expanded dataset and knowledge-infused models lead to significant gains in retrieval performance across different legal search tasks. The improved techniques advance the capability of intelligent systems to understand the semantics of legal texts and better aid legal research.

The next section presents related work in the processing of legal documents. Section 3 introduces the method for building datasets in legal documents and the method for utilizing sentence transformers in this domain. The testing and experimental results are shown in section 4. The last section concludes this paper and presents some future works.

2. RELATED WORK

Nowadays, there is an imperative to cultivate a heightened awareness of legal documents within the entirety of the national populace. Presently, many methodologies are exhibiting advancements in the proficient retrieval of legal documents [16]-[18]. In Vietnam, employees are affected by regulations within the constraints of labor law 2019 [15] and law on employment 2013 [19]. As a result, it is critical to properly process legal documents properly, facilitating easy access for employees to find information pertaining to their rights and benefits.

Le *et al.* [10] proposed two-stage systems that allow users to search for legal information more efficiently and accurately. This research extracts a processed dataset containing questions and official answers from Vietnamese labor law [15]. The dataset focuses on insurance for employees, such as social insurance, health insurance, and unemployment insurance. The proposed method in this study is also evaluated for its accuracy when compared to other baseline methods. However, the crawled dataset is too small to implement in practice.

Unsupervised document clustering aims to automatically organize documents into groups based on their inherent similarities. Within each cluster, documents exhibit a higher degree of thematic coherence than documents from different clusters. Venkatesh [20] exemplifies this application by clustering legal judgments through a two-step approach. First, hierarchical latent dirichlet allocation (hLDA) is employed to extract prominent topics from the corpus. Subsequently, a similarity measure between these topics and individual documents guides the clustering process. This methodology facilitates not only document organization but also enables the generation of document summaries based on the identified topics. However, the accuracy of this method is not high, it only is approximate 55%; thus, it very hard to apply in the practice.

Approaches based on deep learning have attracted a lot of interest in solving the challenge of answering questions. According to Van *et al.* [21], the problem is viewed as the extraction of answers using a pre-trained RoBERTa, and the model output consists of the beginning and ending places within an input sequence. Information retrieval (IR) models with varying variants in terms of sentence embeddings and document databases were used by HUKB [22]. Nonetheless, these studies did not work on a standard dataset for Vietnamese labor law. Nguyen *et al.* [23] used a similar method for medical texts. The authors built a Vietnamese healthcare question answering dataset (ViHealthQA), including 10,015 question-answer passages, in which questions from health-interested users were asked on prestigious health websites and answers from highly qualified experts. They also proposed a two-stage QA system based on sentence-bidirectional encoder representations from transformers (SBERT) [24] using multiple negatives ranking (MNR) loss combined with BM25 [25].

Large language models exhibit proficiency in text generation, language translation, creative content composition, and providing informative responses [26], [27]. These models demonstrate adaptability in executing diverse tasks related to the interrogation of legal documents, encompassing activities like discerning pertinent documents, condensing document content, and extracting pertinent information [28]. The acquired knowledge can subsequently be employed to address inquiries pertaining to legal documents within a knowledge framework dedicated to legal documentation. This framework encapsulates insights such as the interpretation of specific clauses or the ramifications of a judicial decision [29]. However, those models did not give accurate answers to inputted questions.

These are motivations for doing these researches in this manuscript. Firstly, two datasets about the Vietnamese labour law dataset and the Q&A dataset are built. After that, the labelling approach to improve retrieval efficiency is proposed to rebuilt, normalise, and transform both datasets into a new form hierarchically.

3. THE METHOD FOR BUILDING DATASETS IN LEGAL DOCUMENTS AND UTILIZING SENTENCE TRANSFORMERS

The method in this study is proposed based on the results in [10]. The proposed method makes numerous changes to improve its performance, that is, retrieving the proper legal document to answer the input question accurately. The scope of the legal datasets is broadened. In order to serve the experiment effectively, the raw data were crawled, then processed, and datasets generated in a new structure. This section also describes in further detail how the datasets were constructed and developed the proposed model.

3.1. A brief about legal document dataset

According to data crawled from the Vietnam legal library [30], presently encompasses approximately 300,000 legal normative documents spanning 20 different categories and encompassing 27 diverse fields. The link of dataset is at https://link.uit.edu.vn/Aguyx. The essential attributes characterizing a legal normative document include its nomenclature, document number, type, issuing agency, signatory, date of issuance, effective date, promulgation date, and the promulgation act. A pivotal constituent of these documents is the table of contents; however, it is noteworthy that not all documents incorporate this element, and no standardized format exists for its inclusion. Commonly employed indices within the table of contents encompass parts, chapters, sections, articles, clauses, and items. Data structure of a legal document are described as follows in two main tables: Table 1 is a index table, and Table 2 is a document table.

A sample of the legal document is shown in Figure 1. In this figure, the red box represents a chapter, the yellow box represents an article, the blue box represents a clause, and the green box represents a point. Figure 1 shows the text content of a legal document with the visualization of its indices. Each color box in the figure represents a different index type, such as section, chapter, item, article, clause, or point. The indices are crucial for outlining the structure of the documents. The hierarchical nature of the indices is evident, with each index type nested within another, reflecting the organizational structure of the document. The indices are essential for efficient navigation and retrieval of specific sections in response to user queries.

3.2. Data processing

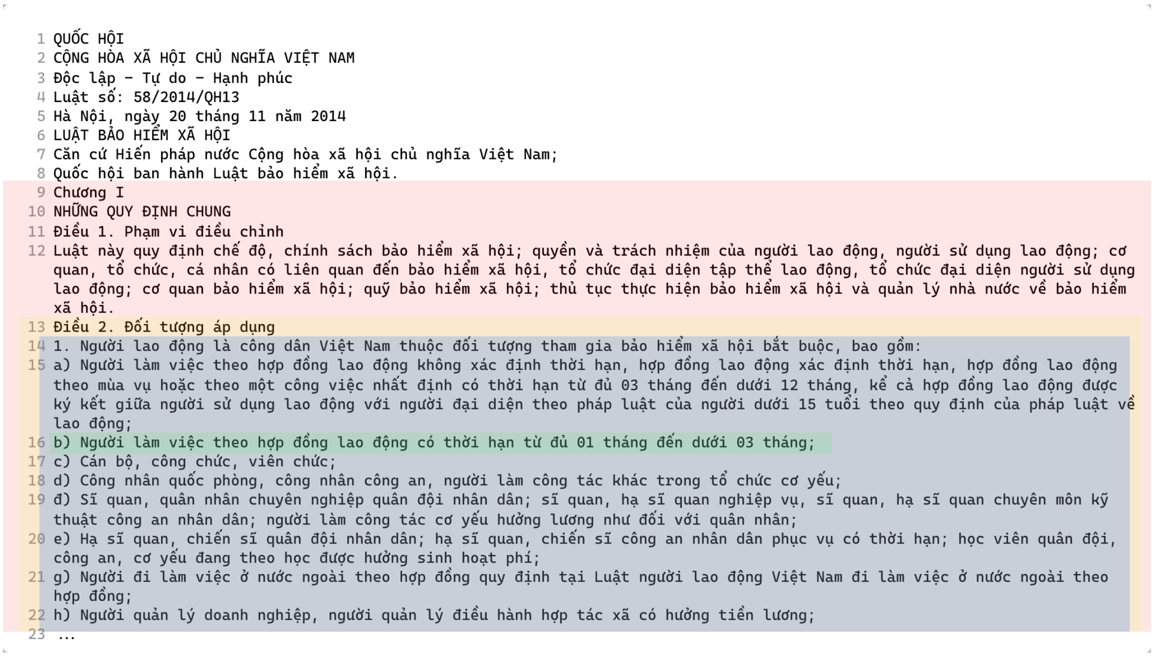
The data processing pipeline is designed to handle the complexities of legal documents, often presented in HTML format. The pipeline consists of two primary steps: HTML document handling and table of contents generation via regex patterns. This systematic approach is essential for efficient information retrieval, especially given legal texts' complex and hierarchical nature.

Table 1. Index table: contains detailed information about the structure and organization of legal normative documents through various index entries

Field name	Description
ID	Unique identifier for each index entry
Index name	The index entry's title indicates the name of the section, chapter, or specific segment within the document.
Index type	Categorizes the index entry by its structural role in the document, such as section, chapter, item, article, clause, and point, to provide clarity on the organizational hierarchy.
Start position	The position within the document where the content of this index entry begins, aiding in precise navigation to the content's start.
End position	The position within the document where the content of this index entry concludes, marking the endpoint for easy content delineation.
Parent ID	The identifier of the parent index entry, if any, under which this entry is nested. This establishes a hierarchical structure among entries, reflecting their organizational relationship in the document.
Document ID	The identifier for the associated document, linking this index entry to the specific document it belongs to, ensuring accurate referencing within a collection of documents.

Table 2. Document table: contains information about legal normative documents

Field name	Description
ID	Unique identifier for the document.
Document title	The title or name of the document.
Document number	The unique document number or identifier.
Type of document	The category or type of the document.
Place of issuance	The location where the document was issued.
Signatory	The person or authority signing the document.
Date of issuance	The date when the document was issued.
Date of effect	The date when the document comes into effect.
Date of official gazette	The date when the document is published in the official gazette.
Official gazette number	The number assigned to the document in the official gazette.
Text content of the document	The textual content and details of the document.
Field/category of the document	The field or category to which the document belongs.



```
QUỐC HỘI
CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc
Luật số: 58/2014/QH13
Hà Nội, ngày 20 tháng 11 năm 2014
LUẬT BẢO HIỂM XÃ HỘI
Căn cứ Hiến pháp nước Cộng hòa xã hội chủ nghĩa Việt Nam;
Quốc hội ban hành Luật bảo hiểm xã hội.
Chương I
NHỮNG QUY ĐỊNH CHUNG
Điều 1. Phạm vi điều chỉnh
Luật này quy định chế độ, chính sách bảo hiểm xã hội; quyền và trách nhiệm của người lao động, người sử dụng lao động; cơ quan, tổ chức, cá nhân có liên quan đến bảo hiểm xã hội, tổ chức đại diện tập thể lao động, tổ chức đại diện người sử dụng lao động; cơ quan bảo hiểm xã hội; quỹ bảo hiểm xã hội; thủ tục thực hiện bảo hiểm xã hội và quản lý nhà nước về bảo hiểm xã hội.
Điều 2. Đối tượng áp dụng
1. Người lao động là công dân Việt Nam thuộc đối tượng tham gia bảo hiểm xã hội bắt buộc, bao gồm:
a) Người làm việc theo hợp đồng lao động không xác định thời hạn, hợp đồng lao động xác định thời hạn, hợp đồng lao động theo mùa vụ hoặc theo một công việc nhất định có thời hạn từ đủ 03 tháng đến dưới 12 tháng, kể cả hợp đồng lao động được ký kết giữa người sử dụng lao động với người đại diện theo pháp luật của người dưới 15 tuổi theo quy định của pháp luật về lao động;
b) Người làm việc theo hợp đồng lao động có thời hạn từ đủ 01 tháng đến dưới 03 tháng;
c) Cán bộ, công chức, viên chức;
d) Công nhân quốc phòng, công nhân công an, người làm công tác khác trong tổ chức cơ yếu;
đ) Sĩ quan, quân nhân chuyên nghiệp quân đội nhân dân; sĩ quan, hạ sĩ quan nghiệp vụ, sĩ quan, hạ sĩ quan chuyên môn kỹ thuật công an nhân dân; người làm công tác cơ yếu hưởng lương như đối với quân nhân;
e) Hạ sĩ quan, chiến sĩ quân đội nhân dân; hạ sĩ quan, chiến sĩ công an nhân dân phục vụ có thời hạn; học viên quân đội, công an, cơ yếu đang theo học được hưởng sinh hoạt phí;
g) Người đi làm việc ở nước ngoài theo hợp đồng quy định tại Luật người lao động Việt Nam đi làm việc ở nước ngoài theo hợp đồng;
h) Người quản lý doanh nghiệp, người quản lý điều hành hợp tác xã có hưởng tiền lương;
...
```

Figure 1. Sample of legal document with index visualization: the red box represents a chapter, the yellow box represents an article, the blue box represents a clause, and the green box represents a point

Step 1: HTML document handling. Firstly, the list of URL documents is crawled from the website using Scrapy [31]. Scrapy is chosen for this work because it is a powerful and flexible web scraping framework that allows us to run multiple concurrent requests and handle the complexities of web scraping. After obtaining the list of URL documents, Scrapy is used to crawl the content of each document from the list of URL documents and save it to an HTML file.

Then the plain text is extracted from the content of each document using beautiful soup [32], a Python library for pulling data out of HTML and XML files. The extracted plain text is then saved to a text file for further processing. The next objective is the normalizing of it into standard form and hierarchical order. The solution is to create and use comprehensive table content for each document to separate it into meaningful little corpus hierarchically based on section, chapter, and item.

Step 2: table of contents generation via regex patterns. This is accomplished by utilizing regular expressions (regex) to identify critical indices within the text, such as chapter titles and section headers. These indices play a crucial role in outlining the structure of the documents. However, it is important to note that this process is semi-automatic, as regex may not consistently identify all indices across different documents. Due to the inherent variability in legal texts, manual adjustments are often required to ensure accuracy and completeness. Table 3 provides a selection of regular expressions that can be used to locate indices within the text.

Table 3. Regular expressions for identifying indices in the text

Regex	Index Type	Description
`^(Phần thứ [\d\w]+.*)$`	Section	Matches indices in the format “Phần thứ <number or word> <content>".
`^(Chương [\d\w]+.*)$`	Chapter	Matches indices in the format “Chương <number or word> <content>".
`^(Mục [\dIVXLCDM]+.*)$`	Item	Matches indices in the format “Mục <number|Roman numeral> <content>".
`^(Điều \d+.*)$`	Article	Matches indices in the format “Điều <number> <content>".
`^(\d+\. .*)$`	Item	Matches indices in the format “<number>. <content>".
`^(\w\).*)$`	Point	Matches indices in the format “<letter>). <content>".

3.3. Building the legal Q&A dataset

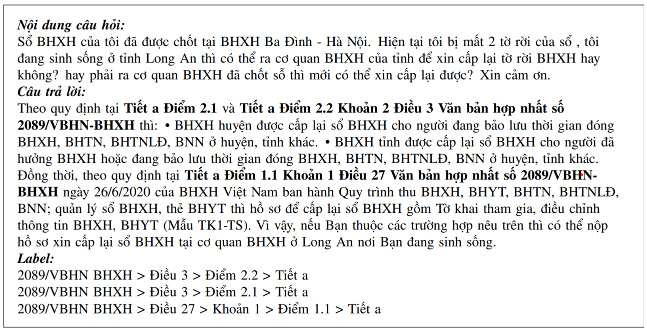
The Q&A dataset were extracted from the official portal of social insurance in Vietnam, known as Vietnam social security (VSS). These sets consist of 19,330 pairs of questions and answers categorized into various fields, as shown in Table 4. Each response includes cite references from the legal corpus (bold text in Figures 2 and 3). It shows that the looking up material in those sites to obtain the appropriate response. From here, the problem returns to multi-label classification. The answers were labeled using multi-label or indices, forming a list of indices corresponding to the content of the respective answers. This allows law consultants to easily identify the relevant legal documents for accurate responses. Additionally, the legal chunks extracted from large legal documents, as described in section 3.1, were assigned these labels, matching their reference indices. The label format is as follows:

```
[law id] > [level 0 index] > [level 1 index] > ... > [level n index].
```

The complete configuration of a Vietnamese question-answer pair and its labels is illustrated in Figure 2 (the English version is shown in Figure 3). We only included data points where the answers were referenced from legal texts. As a result, the actual dataset consists of 4,368 questions. To label the answers in the legal Q&A dataset, we used label studio.

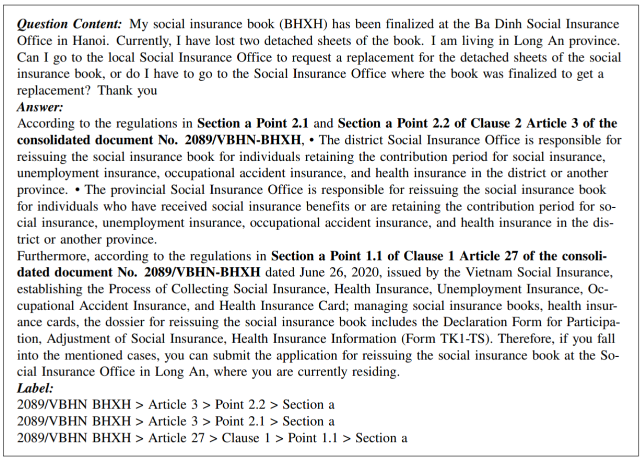
Table 4. Percentage of question types

Dataset	# Percentage
Illness, maternity	22.9
ID of social insurance and health insurance	2.66
One-time social insurance payout	3.39
Unemployment insurance	5.88
Health insurance	5.95
Payment of social insurance, health insurance, unemployment insurance	11.1
Voluntary social insurance	18.2
Retirement, life insurance	20.5
Workplace accident, occupational disease	0.22
Other questions	9.2



Nội dung câu hỏi:
Sổ BHXH của tôi đã được chốt tại BHXH Ba Đình - Hà Nội. Hiện tại tôi bị mất 2 tờ rời của sổ , tôi đang sinh sống ở tỉnh Long An thì có thể ra cơ quan BHXH của tỉnh để xin cấp lại tờ rời BHXH hay không? hay phải ra cơ quan BHXH đã chốt sỗ thì mới có thể xin cấp lại được? Xin cảm ơn.
Câu trả lời:
Theo quy định tại **Tiết a Điểm 2.1** và **Tiết a Điểm 2.2 Khoản 2 Điều 3 Văn bản hợp nhất số 2089/VBHN-BHXH** thì: • BHXH huyện được cấp lại sổ BHXH cho người đang bảo lưu thời gian đóng BHXH, BHTN, BHTNLĐ, BNN ở huyện, tỉnh khác. • BHXH tỉnh được cấp lại sổ BHXH cho người đã hưởng BHXH hoặc đang bảo lưu thời gian đóng BHXH, BHTN, BHTNLĐ, BNN ở huyện, tỉnh khác. Đồng thời, theo quy định tại **Tiết a Điểm 1.1 Khoản 1 Điều 27 Văn bản hợp nhất số 2089/VBHN-BHXH** ngày 26/6/2020 của BHXH Việt Nam ban hành Quy trình thu BHXH, BHYT, BHTN, BHTNLĐ, BNN; quản lý sổ BHXH, thẻ BHYT thì hồ sơ để cấp lại sổ BHXH gồm Tờ khai tham gia, điều chỉnh thông tin BHXH, BHYT (Mẫu TK1-TS). Vì vậy, nếu Bạn thuộc các trường hợp nêu trên thì có thể nộp hồ sơ xin cấp lại sổ BHXH tại cơ quan BHXH ở Long An nơi Bạn đang sinh sống.
Label:
2089/VBHN BHXH > Điều 3 > Điểm 2.2 > Tiết a
2089/VBHN BHXH > Điều 3 > Điểm 2.1 > Tiết a
2089/VBHN BHXH > Điều 27 > Khoản 1 > Điểm 1.1 > Tiết a

Figure 2. Format of the question and answer (Vietnamese)



Question Content: My social insurance book (BHXH) has been finalized at the Ba Dinh Social Insurance Office in Hanoi. Currently, I have lost two detached sheets of the book. I am living in Long An province. Can I go to the local Social Insurance Office to request a replacement for the detached sheets of the social insurance book, or do I have to go to the Social Insurance Office where the book was finalized to get a replacement? Thank you
Answer:
According to the regulations in **Section a Point 2.1** and **Section a Point 2.2 of Clause 2 Article 3 of the consolidated document No. 2089/VBHN-BHXH**, • The district Social Insurance Office is responsible for reissuing the social insurance book for individuals retaining the contribution period for social insurance, unemployment insurance, occupational accident insurance, and health insurance in the district or another province. • The provincial Social Insurance Office is responsible for reissuing the social insurance book for individuals who have received social insurance benefits or are retaining the contribution period for social insurance, unemployment insurance, occupational accident insurance, and health insurance in the district or another province.
Furthermore, according to the regulations in **Section a Point 1.1 of Clause 1 Article 27 of the consolidated document No. 2089/VBHN-BHXH** dated June 26, 2020, issued by the Vietnam Social Insurance, establishing the Process of Collecting Social Insurance, Health Insurance, Unemployment Insurance, Occupational Accident Insurance, and Health Insurance Card; managing social insurance books, health insurance cards, the dossier for reissuing the social insurance book includes the Declaration Form for Participation, Adjustment of Social Insurance, Health Insurance Information (Form TK1-TS). Therefore, if you fall into the mentioned cases, you can submit the application for reissuing the social insurance book at the Social Insurance Office in Long An, where you are currently residing.
Label:
2089/VBHN BHXH > Article 3 > Point 2.2 > Section a
2089/VBHN BHXH > Article 3 > Point 2.1 > Section a
2089/VBHN BHXH > Article 27 > Clause 1 > Point 1.1 > Section a

Figure 3. Format of the question and answer (English)

3.4. The method for utilizing sentence transformers

The model's initial training lacks a significant Vietnamese corpus, especially in specialized legal text. Therefore, fine-tuning and adjusting the neural network weights to better align with the Vietnamese legal context is necessary to achieve optimal results. This preparation involves the data for fine-tuning the model is formatted as a JSON file [33], consisting of a list of examples structured as shown in Figure 4. The form of JSON file is structured with triplets, where each triplet contains a “query" (the question), “pos" (content that answers the question), and “neg" (content that does not answer the question), and "task_name" denotes the documents' name (this JSON file can contain multiple documents). These triplets are derived from synthetic legal Q&A dataset, wherein for each Q&A pair, the question becomes the “query," the answer becomes the “pos," and the “neg" is generated through a retrieval algorithm. For creating the dataset for fine-tuning, the legal Q&A dataset is leveraged to generate examples for fine-tuning, specifically:

- For each Q&A pair in the dataset, the "query" becomes the question, and “pos" becomes the content of the labeled indices from section 3.2.
- To generate "neg", the method for text chunking, as mentioned in Approach 1, is utilized to find the top k contents. Then, we check which contents are not part of "pos"; these contents are considered "neg".

This process includes two main stages. Stage 1 is constituted of the **f**ine-**t**uning process, named FTS1. At this point, the “query", “pos," and “neg" are inputted into the pipeline, and apply the contrastive learning technique to train the embedding model Intructor-base. This would alter the weights of the model intructor-base to adapt to the Vietnamese legal framework and boost performance. After completing the fine-tuning, proceed to Stage 2, where it is continued to generate a dataset comparable to the work in Stage 1. However, instead of using basic algorithms term frequency-inverse document frequency (TF-IDF)/best match 25 (BM25) at [34], [35], the fine-tuned embedding model from Stage 1 was employed to generate "neg" samples. They were used as well in conjunction with “pos" and “query" to fine-tune the fine-tuned embedding model instructor base one more time. In this case, it is called FTS2. The fine-tuned embedding model is obtained at the end of each stage and implemented to execute experiments and evaluations. Figure 5 depicts the detailed workflow of the proposed method.

{ "***query***": ["**Represent the input question**",
"big little lies season 2 how many episodes"],
"***pos***": ["**Represent the relevant document for retrieval**",
"Big Little Lies (TV series) series garnered several accolades. It received 16 Emmy Award nominations and won eight, including Outstanding Limited Series and acting awards for Kidman, Skarsgård, and Dern. The trio also won Golden Globe Awards in addition to a Golden Globe Award for Best Miniseries or Television Film win for the series. Kidman and Skarsgård also received Screen Actors Guild Awards for their performances. Despite originally being billed as a miniseries, HBO renewed the series for a second season. Production on the second season began in March 2018 and is set to premiere in 2019. All seven episodes are being written by Kelley"],
"***neg***": ["**Represent the irrelevant document for retrieval**",
"Little People, Big World final minutes of the season two A finale, Farm Overload. A crowd had gathered around Jacob, who was lying on the ground near the trebuchet. The first two episodes of Season Two B focuses on the accident and how the local media reacted to it. The first season of Little People, Big World generated solid ratings for TLC (especially in the important 18–49 demographic), leading to the shows̈ renewal for a second season. Critical reviews of the series have been generally positive, citing the shows̈ positive portrayal of little people. Conversely, other reviews have claimed that the show has a voyeuristic bend"], }

Figure 4. JSON format of the question and answer sample

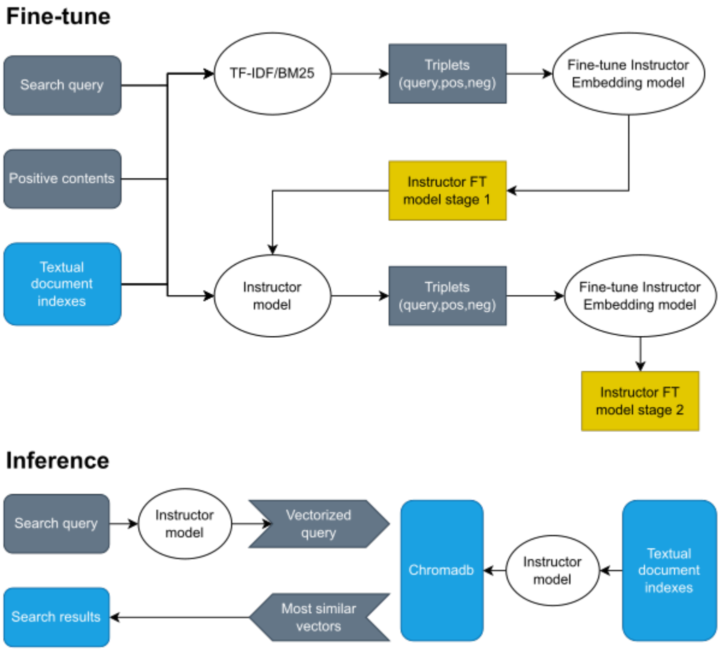


Figure 5. The workflow of the proposed method workflow

4. TESTING AND EXPERIMENTAL RESULTS

4.1. Testing

In this section, the testing process is undertaken to evaluate the effectiveness of the designed system. The primary function of this system is to provide relevant legal articles in response to user queries. Two examples illustrate how the system processes input questions and retrieves pertinent legal documentation. The terms input question and system response refer to the user's inquiry and the output that our system returns to the user after receiving it.

Example 4.1:

Input question: *Điều kiện đăng ký luật bảo hiểm xã hội là gì?* (English: What are the conditions for registering

for social insurance law?)
System response: The system will give an answer for the query as Vietnamese. The illustration is translated into English.

- As article 13, decree on 134/2015/NĐ-CP about detailing a number of articles of the law on social insurance regarding voluntary social insurance, dated on 29 December 2015:
  Article 13. Procedures for re-registration of premium payment method and monthly income on which voluntary social insurance premiums are based:
  1. A dossier for re-registration of premium payment method and monthly income on which voluntary social insurance premiums are based must comprise:
  a) The social insurance book;
  b) The declaration form for social insurance participation.
  2. Settlement of the re-registration of premium payment method and monthly income on which voluntary social insurance premiums are based:
  a) Voluntary social insurance participants shall submit a dossier prescribed in clause 1 of this article to the social insurance agency;
  b) The social insurance agency shall process the dossier within a day if it is complete as prescribed. If refusing to process the dossier, it shall give a written reply clearly stating the reason.
- As article 97, law no. 58/2014/QH13 about law on social insurance, dated on 20 November 2014:
  Article 97. Registration dossiers for participation in social insurance and grant of social insurance books:
  1. A registration dossier for first-time participation in social insurance must comprise:
  a) An employer's declaration form for participation in social insurance, enclosed with a list of employees to participate in social insurance;
  b) Employees' declaration forms for participation in social insurance.
  2. A dossier for re-grant of a lost or damaged social insurance book must comprise:
  a) An employee's application for re-grant of a social insurance book;
  b) The social insurance book, in case it is damaged.
  3. The government shall stipulate the procedures and dossier for participation in social insurance and grant of social insurance books for the subjects defined at point e, Clause 1, article 2 of this Law.

Example 4.2:
Input question: *Nghỉ việc có được hưởng lương tháng cuối không?* (English: Can I get paid at the last month when I quit my job?)
System response: The system will give an answer for the query in Vietnamese. The illustration is translated into English.

- As article 38, circular no. 59/2015/TT-BLDTBXH dated on 29 December 2015 of the Ministry of Labor, War Invalids and Social Affairs detailing and guiding the implementation of a number of articles of the law on social insurance on compulsory social insurance:
  Article 38. Benefits for employees who have decided to quit their jobs while waiting for their pension and monthly benefits to be resolved.
  Benefits for employees who have decided to quit their job while waiting for their pension and monthly benefits to be resolved are implemented according to article 25 of decree no. 115/2015/ND-CP.
- As clause 1, article 19, circular no. 59/2015/TT-BLDTBXH dated on 29 December 2015 of the Ministry of Labor, War Invalids and Social Affairs detailing and guiding the implementation of a number of articles of the law on social insurance on compulsory social insurance: Article 19. One-time social insurance
  1. One-time social insurance is implemented according to the provisions of article 60 of the law on social insurance, No. 93/2015/QH13 dated on 22 June 2015, of the national assembly on the implementation of insurance policy. one-time social benefits for workers and article 8 of decree no. 115/2015/ND-CP.

In these examples, the answer for Example 4.1 of the designed system focuses to the main meaning of the query. However, the answer of Example 4.2 is not good, it does not give a correctly content for the inputted query.

4.2. Experiments on legal information retrieval

When a query is inputted, the system retrieves a list of articles relating to the query for answering the query. For evaluation the results, the metric evaluation - top$K$@acc: is used. Accuracy is calculated as the ratio of correct contents (contents used to answer the question) appearing in the top K returned results. L_K is a collection containing labels, or IDs of Law documents, which our system predicts are most related to the query, l_q is the query's actual collection of labels. Specifically, the formula is:

$$TopK@\text{acc} = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}1, & \text{if } 1, l_q \subseteq L_K\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

where, L_K is the set containing K labels with the most similarity to query q, and l_q is the set of correct contents of query q.

The experiments in this study are based on two approaches: Using fundamental algorithms and using transformers of sentences.

- Approach 1: Utilizing fundamental algorithms

In the first approach, basic algorithms, such as TF-IDF [34] and BM25 [35], compute the similarity between questions and content within the dataset. Subsequently, contents are sorted in descending order of similarity and presented as results. To enhance the results, additional techniques are applied for content normalization, such as removing special characters and utilizing tools from Underthesea library [36] to handle punctuation marks and word segmentation (WS).

Table 5 shows the outcomes of the initial approach utilizing two fundamental algorithms: TF-IDF and BM25. Two normalization methods were employed: one utilizing WS and the other without WS. The results obtained from this method have not been satisfactory.

Table 5. Results of the first approach

Name	Top5@acc	Top10@acc	Top20@acc	Top50@acc
TDIDF	0.1037	0.201	0.347	0.5289
BM25	0.079	0.1474	0.2556	0.4485
TDIDF_WS	0.1094	0.199	0.3344	0.5187
BM25_WS	0.0944	0.1746	0.2908	0.4709

- Approach 2: Utilizing sentence transformers

The retrieval of relevant legal documents in the Vietnamese language presents unique challenges due to the language's complexity and the specialized nature of legal terminology. To address this, the InstructorEmbedding model, an architecture rooted in sentence transformers-SBERT [23], is employed and considered state-of-the-art in this field. This model takes a string as input and returns a 768-dimensional vector, which allows for the comparison of semantic similarity between a question and relevant contents within the dataset by computing cosine similarity between the question's vector and the vectors of the dataset contents. Cosine similarity is computed using the following formula:

$$\text{similarity}(\mathbf{q}, \mathbf{d}_i) = \frac{\mathbf{q} \cdot \mathbf{d}_i}{\|\mathbf{q}\| \cdot \|\mathbf{d}_i\|} \tag{2}$$

where, $\mathbf{q}$ is the vector representation of the question, and $\mathbf{d}_i$ is the vector representation of the i-th content in the dataset.

The original instructor model consists of 3 models: base, large, and extra large XL [33]. The instructor-base model has 335 million parameters, the Instructor-large model has 500 million parameters, and the instructor-XL model has 1.5 billion parameters. Due to hardware limitations, the smallest model, which is the base model, is fine-tuned. The comparison between the original models instructor-base/large/XL and the proposed models instructor-base FTS1/FTS2 are presented in Table 6. Despite its size, the results after fine-tuning are excellent. Without finetuning on the Vietnamese dataset, the models instructor-base/large/XL produce results inferior to the typical TF-IDF/BM25 technique. However, when the model instructor-base is fine-tuned across two phases using the created dataset, the outcomes are clearly enhanced. Hence, if the resources computing are adequate to finetune the model instructor-large/XL, we would have expected much greater performance. In addition, when the dataset is expanded to a larger scale and formed using the same design, the outcome definitely improves significantly.

The expanded legal dataset covers a significantly broader domain of Vietnamese law compared to the narrow prior dataset on labor regulations. This diversity and larger volume of legal text provide more robust training data to enhance model capabilities. While pre-trained language models like BERT capture semantics effectively, injecting structured knowledge can further improve understanding of legal terminology. The knowledge-infused approach advances state-of-the-art accuracy.

The data processing pipeline also implements more sophisticated techniques like utilizing BeautifulSoup and regex for content extraction and indexing. The systematic document segmentation aligned with hierarchical indices enables more straightforward navigation and retrieval of precise sections. This structured preparation of the dataset enhances the quality and alignment critical for legal search. The experimental results demonstrate improvements from the enriched dataset and knowledge-infused models across legal search tasks. Our best approach achieves 89.12% Top-50 accuracy, significantly higher than 52.89% from basic TF-IDF (Table 5) and still far better than the best of prior research 68.12% [10]. The upgraded techniques better equip intelligent systems to comprehend legal texts and assist users in efficient access to relevant laws.

Table 6. Results of the second approach

Name	Top5@acc	Top10@acc	Top20@acc	Top50@acc
INSTRUCTOR-BASE	0.0119	0.0221	0.0416	0.0944
INSTRUCTOR-LARGE	0.0138	0.0247	0.0421	0.1023
INSTRUCTOR-XL	0.0188	0.0312	0.0537	0.1427
INSTRUCTOR-BASE FTS1	0.4832	0.5741	0.6621	0.7765
INSTRUCTOR-BASE FTS2	0.6431	0.7432	0.8123	0.8912

5. CONCLUSION AND FUTURE WORK

In this study, a larger-scale Vietnamese question-answering legal dataset, which mainly covers a wide range of labour laws in Vietnam, is built. The dataset comprises 300,000 normative legal documents, encompassing 20 different types across 27 diverse fields. Key attributes defining a legal normative document include its nomenclature, document number, type, issuing agency, signatory, date of issuance, effective date, promulgation date, and the promulgation act. Besides, the current approaches for legal information retrieval are boosted by structured information derived from legal ontologies constructed from the built dataset. Experiments and results show that the enriched dataset and knowledge-infused models result in considerable improvements in retrieval performance across a variety of legal search tasks. The upgraded methodologies strengthen smart systems' capacity to grasp the semantics of legal texts and thus promote legal study. In the future, the incorporating knowledge graphs into extracting knowledge from legal documents in future work will be studied. By leveraging the structured information in the knowledge graph, the quality of the responses to user queries will be improved. Additionally, the use large language models will be explored to enhance the performance further. By leveraging the power of large language models, the system's ability is increased to reason and generate more helpful answers based on the information extracted from legal documents. Combining the knowledge graph and large language models, the designed system is emerging to provide accurate and relevant information to users and offer more comprehensive and insightful insights into the legal domain.

ACKNOWLEDGEMENTS

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2023-26-04.

REFERENCES

[1] F. Ryan, "Delivering legal services without lawyers," in *Digital Lawyering: Technology and Legal Practice in the 21st Century*, London, UK: Routledge, 2021, pp. 103–135, doi: 10.4324/9780429298219-4.

[2] S. Ramaswamy, R. Sreelekshmi, and G. Veena, "Complexity analysis of legal documents," in *International Conference on Artificial Intelligence on Textile and Apparel*, 2024, pp. 141–154, doi: 10.1007/978-981-99-8476-3_12.

[3] R. Sil, Alpana, and A. Roy, "A review on applications of artificial intelligence over indian legal system," *IETE Journal of Research*, vol. 69, no. 9, pp. 6029–6038, 2023, doi: 10.1080/03772063.2021.1987343.

[4] Y. Ren and X. Lu, "The lawyer system," in *A New Study on the Judicial Administrative System with Chinese Characteristics*, Singapore: Springer, 2020, pp. 393–429, doi: 10.1007/978-981-15-4182-7_10.

[5] K.-B. Ooi *et al.*, "The potential of generative artificial intelligence across disciplines: Perspectives and future directions," *Journal of Computer Information Systems*, pp. 1–32, 2023, doi: 10.1080/08874417.2023.2261010.

[6] J. Zhu, J. Wu, X. Luo, and J. Liu, "Semantic matching based legal information retrieval system for COVID-19 pandemic," *Artificial Intelligence and Law*, vol. 32, no. 2, pp. 397–426, 2024, doi: 10.1007/s10506-023-09354-x.

[7] T. S. Nguyen, L. M. Nguyen, S. Tojo, K. Satoh, and A. Shimazu, "Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts," *Artificial Intelligence and Law*, vol. 26, no. 2, pp. 169–199, 2018, doi: 10.1007/s10506-018-9225-1.

[8] N. A. K. Rosili, N. H. Zakaria, R. Hassan, S. Kasim, F. Z. C. Rose, and T. Sutikno, "A systematic literature review of machine learning methods in predicting court decisions," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 4, pp. 1091–1102, 2021, doi: 10.11591/ijai.v10.i4.pp1091-1102.

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, pp. 4171–4186.

[10] H. H. Le *et al.*, "Intelligent retrieval system on legal information," in *Asian Conference on Intelligent Information and Database Systems*, 2023, pp. 97–108, doi: 10.1007/978-981-99-5834-4_8.

[11] D. V. Dang, V. T. Pham, T. Cao, N. Do, H. Q. Ngo, and H. D. Nguyen, "A practical approach to leverage knowledge graphs for legal query," in *International Conference on Intelligent Systems and Data Science*, 2024, pp. 271–284, doi: 10.1007/978-981-99-7649-2_21.

[12] F. Sovrano, M. Palmirani, and F. Vitali, "Legal knowledge extraction for knowledge graph based question-answering," *Frontiers in Artificial Intelligence and Applications*, vol. 334, pp. 143–153, 2020, doi: 10.3233/FAIA200858.

[13] E. Filtz, "Building and processing a knowledge-graph for legal data," in *14th International Conference on Semantic Web (ESWC 2017)*, 2017, pp. 184–194, doi: 10.1007/978-3-319-58451-5_13.

[14] H. Q. Ngo, H. D. Nguyen, and N. A. L. -Khac, "Ontology knowledge map approach towards building linked data for vietnamese legal applications," *Vietnam Journal of Computer Science*, vol. 11, no. 2, pp. 323–342, 2024, doi: 10.1142/S2196888824500015.

[15] The National Assembly, *Labor code*. Hanoi, Vietnam: Constitution of Socialist Republic of Vietnam, Law no. 45/2019/QH14, 2019.

[16] G. Governatori, T. B. -Capon, B. Verheij, M. Araszkiewicz, E. Francesconi, and M. Grabmair, "Thirty years of artificial intelligence and law: the first decade," *Artificial Intelligence and Law*, vol. 30, no. 4, pp. 481–519, 2022, doi: 10.1007/s10506-022-09329-4.

[17] G. Sartor *et al.*, "Thirty years of artificial intelligence and law: the second decade," *Artificial Intelligence and Law*, vol. 30, no. 4, pp. 521–557, 2022, doi: 10.1007/s10506-022-09326-7.

[18] S. Villata *et al.*, "Thirty years of artificial intelligence and law: the third decade," *Artificial Intelligence and Law*, vol. 30, no. 4, pp. 561–591, 2022, doi: 10.1007/s10506-022-09327-6.

[19] The National Assembly, *Law on employment*. Hanoi, Vietnam: Constitution of Socialist Republic of Vietnam, Law no. 38/2013/QH13, 2013.

[20] R. K. Venkatesh and K. Raghuveer, "Legal documents clustering and summarization using hierarchical latent dirichlet allocation," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 2, no. 1, pp. 27–35, 2013, doi: 10.11591/ij-ai.v2i1.1186.

[21] H. N. Van, D. Nguyen, P. M. Nguyen, and M. L. Nguyen, "Miko team: Deep learning approach for legal question answering in ALQAC 2022," in *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*, 2022, pp. 1–5, doi: 10.1109/KSE56063.2022.9953780.

[22] M. Yoshioka, Y. Suzuki, and Y. Aoki, "HUKB at the COLIEE 2022 statute law task," in *JSAI International Symposium on Artificial Intelligence*, 2023, pp. 109–124, doi: 10.1007/978-3-031-29168-5_8.

[23] N. T.-H. Nguyen, P. P.-D. Ha, L. T. Nguyen, K. V. Nguyen, and N. L.-T. Nguyen, "SPBERTQA: A two-stage question answering system based on sentence transformers for medical texts," in *15th International Conference on Knowledge Science, Engineering and Management*, 2022, pp. 371–382, doi: 10.1007/978-3-031-10986-7_30.

[24] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019, pp. 3982–3992, doi: 10.18653/v1/d19-1410.

[25] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[26] B. Min *et al.*, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023, doi: 10.1145/3605943.

[27] G. Balloccu, L. Boratto, G. Fenu, F. M. Malloci, and M. Marras, "Explainable recommender systems with knowledge graphs and language models," in *European Conference on Information Retrieval*, 2024, pp. 352–357, doi: 10.1007/978-3-031-56069-9_46.

[28] S. Maroudas, S. Legkas, P. Malakasiotis, and I. Chalkidis, "Legal-tech open diaries: Lesson learned on how to develop and deploy light-weight models in the era of humongous language models," in *Natural Legal Language Processing Workshop 2022*, 2022, pp. 88–110, doi: 10.18653/v1/2022.nllp-1.8.

[29] H. Q. Ngo, H. D. Nguyen, and N. A. L. -Khac, "Building legal knowledge map repository with NLP toolkits," in *Conference on Information Technology and its Applications*, 2023, pp. 25–36, doi: 10.1007/978-3-031-36886-8_3.

[30] Vietnamese Government, "Vietnamese legal library," *Socialist Republic of Vietnam*. Accessed: Mar. 27, 2024. [Online]. Available: https://thuvienphapluat.vn

[31] D. K. -Loukas, *Learning scrapy*. Birmingham, USA: Packt Publishing, 2016.

[32] "Beautiful soup documentation," *Crummy*. Accessed: Mar. 27, 2024. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[33] H. Su *et al.*, "One embedder, any task: Instruction-finetuned text embeddings," in *Findings of the Association for Computational Linguistics*: ACL 2023, 2023, pp. 1102–1121, doi: 10.18653/v1/2023.findings-acl.71.

[34] S. Qaiser and R. Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[35] M.-Y. Kim, J. Rabelo, K. Okeke, and R. Goebel, "Legal information retrieval and entailment based on BM25, transformer and semantic thesaurus methods," *The Review of Socionetwork Strategies*, vol. 16, no. 1, pp. 157–174, 2022, doi: 10.1007/s12626-022-00103-1.

[36] V. Anh *et al.*, "Underthesea," *GitHub*. 2024. Accessed: Mar. 27, 2024. [Online]. Available: https://github.com/undertheseanlp/underthesea

BIOGRAPHIES OF AUTHORS

Vuong Pham received a B.S. degree in mathematics and informatics from the University of Sciences, VNU-HCM, Vietnam, in 2003 and an M.S. degree in Information Technology from the University of Science, VNU-HCM, Vietnam, in 2008. He is a Ph.D. student at the University of Science, VNU-HCM, Vietnam. From 2003 – 2006, he was a lecturer at the University of Sciences, VNU-HCM, Vietnam. From 2006 – 2019, he was a lecturer at the University of Information Technology, VNU-HCM, Vietnam. He is currently a Vice Director of the Institute of Data Science and Artificial Intelligence at Sai Gon University, Vietnam. His research interests include artificial intelligence, software engineering, and game development. He received the Best Paper Award at ICOCO 2022, Best Student Paper Awards at KEOD 2023 and KSE 2020, and Best Presentation Award at KSE 2021. He can be contacted at email: vuong.pham@sgu.edu.vn.

Hoang Huy Le He received his B.Sc. (Mathematics and Computer Science) from the University of Science, Viet Nam National University, Ho Chi Minh City, in 2023. His research includes computer vision, natural language processing, missing data, and economics. He can be contacted at email: hoangle7910@gmail.com.

Thinh Phu Ngo holds a Bachelor of Mathematics and Computer Science from the University of Science, Viet Nam National University Ho Chi Minh City. His research areas of interest include artificial intelligence, blockchain, and legal. He can be contacted at email: thinhngow@gmail.com.

Binh Nguyen is Head of the Department of Computer Science at the Faculty of Mathematics and Computer Science, VNU HCM – University of Science. He has over ten years of experience in AI and data science. He defended his Ph.D. thesis with the highest honors at Ecole Polytechnique (Paris, France) in 2012. Up to now, he has had over 100 publications and four patents filed in USA and Canada. He also had substantial experience in building research and development teams to help the company or the startup deliver AI products. He can be contacted at email: ngtbinh@hcmus.edu.vn.

Diem Nguyen received the B.S. and M.Sc. in Computer Science from the University of Information Technology, VNU-HCM, Vietnam, in 2011 and 2015. She is a lecturer at the Faculty of Computer Science, University of Information Technology, VNU-HCM, Vietnam. Her research interests include artificial intelligence, knowledge representation, and intelligent systems. She can be contacted at email: diemntn@uit.edu.vn.

Hien D. Nguyen received his B.S. and M.S. degrees from the University of Sciences, VNU-HCM, Vietnam, in 2008 and 2011, respectively. He received his Ph.D. degree from the University of Information Technology, VNU-HCM, in 2020. He is a senior lecturer at the Faculty of Computer Science, University of Information Technology, VNU-HCM, Vietnam. His research interests include knowledge representation, knowledge-based systems, and knowledge engineering, especially intelligent and expert systems. He received the Best Paper Awards at CITA 2023, SOMET 2022, and ICOCO 2022, Best Student Paper Awards at KEOD 2023 and KSE 2020, and Incentive Prizes of the Technological Creation Awards of Binh Duong province in 2021 and VIFOTEC in 2016. He can be contacted at email: hiennd@uit.edu.vn.