

LNAI 13995

Ngoc Thanh Nguyen · Siridech Boonsang ·
Hamido Fujita · Bogumiła Hnatkowska ·
Tzung-Pei Hong · Kitsuchart Pasupa ·
Ali Selamat (Eds.)

Intelligent Information and Database Systems

15th Asian Conference, ACIIDS 2023
Phuket, Thailand, July 24–26, 2023
Proceedings, Part I

1
Part I

CIIDS
2023

 Springer

Ngoc Thanh Nguyen · Siridech Boonsang ·
Hamido Fujita · Bogumiła Hnatkowska ·
Tzung-Pei Hong · Kitsuchart Pasupa ·
Ali Selamat
Editors

Intelligent Information and Database Systems


15th Asian Conference, ACIIDS 2023
Phuket, Thailand, July 24–26, 2023
Proceedings, Part I

Editors

Ngoc Thanh Nguyen 
Wrocław University of Science
and Technology
Wrocław, Poland


Hamido Fujita 
Iwate Prefectural University Iwate
Iwate, Japan

Tzung-Pei Hong 
National University of Kaohsiung
Kaohsiung, Taiwan

Ali Selamat 
Malaysia Japan International Institute
of Technology
Kuala Lumpur, Malaysia

Siridech Boonsang 
King Mongkut's Institute of Technology
Ladkrabang
Bangkok, Thailand

Bogumiła Hnatkowska 
Wrocław University of Science
and Technology
Wrocław, Poland

Kitsuchart Pasupa 
King Mongkut's Institute of Technology
Ladkrabang
Bangkok, Thailand

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-981-99-5833-7 ISBN 978-981-99-5834-4 (eBook)
<https://doi.org/10.1007/978-981-99-5834-4>

LNCS Sublibrary: SL7 – Artificial Intelligence

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Singapore Pte Ltd. 2023

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Paper in this product is recyclable.

Contents – Part I

Case-Based Reasoning and Machine Comprehension

On the Improvement of the Reasoning Cycle in Case-Based Reasoning	3
<i>Fateh Boulmaiz, Patrick Reignier, and Stephane Ploix</i>	
Exploring Incompleteness in Case-Based Reasoning: A Strategy for Overcoming Challenge	17
<i>Fateh Boulmaiz, Patrick Reignier, and Stephane Ploix</i>	
Leveraging both Successes and Failures in Case-Based Reasoning for Optimal Solutions	31
<i>Fateh Boulmaiz, Patrick Reignier, and Stephane Ploix</i>	
Transfer Learning for Abnormal Behaviors Identification in Examination Room from Surveillance Videos: A Case Study in Vietnam	45
<i>Pham Thi-Ngoc-Diem, Lan Ngoc Ha, and Hai Thanh Nguyen</i>	
A Novel Question-Context Interaction Method for Machine Reading Comprehension	58
<i>Tuan-Anh Phan, Hoang Ngo, and Khac-Hoai Nam Bui</i>	
Granular Computing to Forecast Alzheimer’s Disease Distinctive Individual Development	70
<i>Andrzej W. Przybyszewski, Jerzy P. Nowacki, and Aldona Drabik</i>	

Computer Vision

AdVLO: Region Selection via Attention-Driven for Visual LiDAR Odometry	85
<i>Han Lam, Khoa Pho, and Atsuo Yoshitaka</i>	
Intelligent Retrieval System on Legal Information	97
<i>Hoang H. Le, Cong-Thanh Nguyen, Thinh P. Ngo, Phu V. Vinh, Binh T. Nguyen, Anh T. Huynh, and Hien D. Nguyen</i>	
VSNet: Vehicle State Classification for Drone Image with Mosaic Augmentation and Soft-Label Assignment	109
<i>Youlkyeong Lee, Jehwan Choi, and Kanghyun Jo</i>	

Creating High-Resolution Adversarial Images Against Convolutional
Neural Networks with the Noise Blowing-Up Method 121
Franck Leprévost, Ali Osman Topal, and Enea Mancellari

Faster Imputation Using Singular Value Decomposition for Sparse Data 135
*Phuc Nguyen, Linh G. H. Tran, Bao H. Le, Thuong H. T. Nguyen,
Thu Nguyen, Hien D. Nguyen, and Binh T. Nguyen*

Combination of Deep Learning and Ambiguity Rejection for Improving
Image-Based Disease Diagnosis 147
Thanh-An Pham and Van-Dung Hoang

Data Mining and Machine Learning

Towards Developing an Automated Chatbot for Predicting Legal Case
Outcomes: A Deep Learning Approach 163
*Shafiq Alam, Rohit Pande, Muhammad Sohaib Ayub,
and Muhammad Asad Khan*

Fuzzy-Based Factor Evaluation System for Momentum Overweight
Trading Strategy 175
Chi-Fang Chao, Mu-En Wu, and Ming-Hua Hsieh

Enhancing Abnormal-Behavior-Based Stock Trend Prediction Algorithm
with Cost-Sensitive Learning Using Genetic Algorithms 186
Chun-Hao Chen, Szu-Chi Wang, Mu-En Wu, and Kawuu W. Lin

Leveraging Natural Language Processing in Persuasive Marketing 197
Evrípides Christodoulou and Andreas Gregoriades

Direction of the Difference Between Bayesian Model Averaging
and the Best-Fit Model on Scarce-Data Low-Correlation Churn Prediction 210
Paul J. Darwen

Tree-Based Unified Temporal Erasable-Itemset Mining 224
Tzung-Pei Hong, Jia-Xiang Li, Yu-Chuan Tsai, and Wei-Ming Huang

Design Recovery of Data Model Hidden in JSON File 234
Bogumiła Hnatkowska

Accurate Lightweight Calibration Methods for Mobile Low-Cost
Particulate Matter Sensors 248
*Per-Martin Jørstad, Marek Wojcikowski, Tuan-Vu Cao,
Jean-Marie Lepioulfle, Krystian Wojtkiewicz, and Phuong Hoai Ha*



Intelligent Retrieval System on Legal Information

Hoang H. Le^{1,3,4}, Cong-Thanh Nguyen^{2,3}, Thinh P. Ngo^{1,3}, Phu V. Vinh^{1,3},
Binh T. Nguyen^{1,3,4}, Anh T. Huynh^{2,3}, and Hien D. Nguyen^{2,3(✉)}

¹ University of Science, Ho Chi Minh City, Vietnam

² University of Information Technology, Ho Chi Minh City, Vietnam
hiennd@uit.edu.vn

³ Vietnam National University, Ho Chi Minh City, Vietnam

⁴ AISIA Research Lab, Ho Chi Minh City, Vietnam

Abstract. Nowadays, intelligent retrieval systems in law are vital in facilitating legal research and providing access to vast legal information. These systems allow users to search for legal information more efficiently and accurately. This paper investigates retrieval systems, their technological advancements, and their impact on legal research. The experimental results show that the proposed method is emerging to apply for analysis queries of practical law cases and extract suitable information from legal documents. It also discusses the challenges associated with law retrieval systems and explores future research directions to improve them.

Keywords: Intelligent system · Information Retrieval · BERT · Natural language processing · Ontology

1 Introduction

After the Covid-19 pandemic, many workers find policies to support their benefits based on legal regulations [9, 21]. Vietnamese labor law, which has many related legal documents, plays a crucial role in safeguarding the rights and benefits of employees. It includes access to healthcare insurance [1], social insurance [2], and unemployment insurance [7, 8]. In this context, healthcare insurance refers to medical coverage provided by an employer, while social insurance encompasses benefits such as retirement and occupational accident insurance. Unemployment insurance provides temporary financial assistance to employees who have lost their jobs.

The law information retrieval system is designed to assist users in finding relevant legal information quickly and efficiently [22]. The system allows users to search a vast collection of legal documents and databases to retrieve relevant information. The system employs algorithms and techniques to provide accurate and relevant search results [25]. Modern legal retrieval systems use sophisticated technologies such as natural language processing to provide accurate and relevant results [4, 25].

This paper studies the techniques for legal text retrieval using two popular text retrieval models, BERT and TF-IDF/BM25. BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art language model that can capture the context and semantics of words in a sentence [5]. TF-IDF (Term Frequency-Inverse Document Frequency) and BM25 are traditional text retrieval model that assigns weights to words based on their frequency in a document, and inverse frequency in a corpus [18]. Those techniques are designed to extract legal information for inputted queries based on legal documents about health-care insurance [1], social insurance [2], and unemployment insurance [7, 8]. The experiments show that combining both models can further improve the accuracy and efficiency of legal text retrieval, especially for long and complex legal texts.

The following section presents related work for designing information retrieval on law documents. Section 3 proposes extracting information from multiple legal documents using BERT and TF-IDF/BM25. It also gives the metric to evaluate the effectiveness of methods. Section 4 shows the experimental results of the proposed methods and the combination method. The last section concludes this study and gives future works.

2 Related Work

Numerous labor law restrictions have an impact on workers. So, it is essential to process legal paperwork correctly. Presently, several techniques have made substantial progress in the intelligent processing of legal documents. Those techniques include extracting, classifying, and question-answering [10, 24].

Ontology is an effective method to organize the knowledge base of legal documents [6, 16]. Legal-Onto is an ontology for organizing legal documents [15, 17]. That ontology was built based on the foundation of the relational, intellectual model, Rela-model [14]. Legal-Onto was applied to represent the Land Law [15], and road traffic law [17] in Vietnam. Another ontology for a legal informatics document is also presented in [6]. This ontology can be used to represent the structure of a legal resource, legal temporal events, legal activities that have an impact on the document, and the semantic organization of the legal document. However, those methods have yet to be mentioned as the solution to analyze queries as law cases in practice.

An ontology that specifies domain knowledge and a database of document repositories can be combined using a technique created by the study in [13]. This method applies a model of domain knowledge called the categorized key phrase-based ontology to a number of information retrieval tasks. However, this model's graph-based measure has yet to be used to evaluate the semantic relevance of legal documents.

The authors of [11] created a system for analyzing Vietnamese legal text by fusing the benefits of standard information retrieval methods, pretrained masked language models (BERT), and legal domain knowledge. It was also advised to use a novel data augmentation technique based on knowledge of the legal field and legal textual entailment. Nevertheless, that method fails to convey the meaning of the legal instrument accurately.

3 Methodology

This section presents the main problem of information retrieval systems for legal documents, the data collection process, data processing, and evaluation metrics.

3.1 Problem Formulation

In this research, the general idea is to identify a list of legal documents that includes the intended meaning of the questions. Specifically, the goal is to develop a broad approach to generate a set of legal documents that can automatically extract the relevant information from a given question. The resulting list of legal documents will provide a comprehensive framework for accurately extracting the questions' intended meaning and improving the system's overall performance.

The model employed various evaluation metrics and techniques to evaluate the proposed approach's effectiveness. Using multiple evaluation approaches, the system aims to provide a comprehensive assessment of the proposed methodology and its ability to extract the intended meaning of questions accurately.

3.2 Data Collection

Questions Data. The data are published and belong to the Vietnam Social Insurance Portal¹. There is a section on the Portal for users to submit questions about social insurance, view answers, and view other questions and answers.

Data are crawled from people's questions and official answers from the source: including question id, sender, submission date, field, question name, question content, and response content. For each question, data have been extracted from the law contained in the response. First, the method has been used semi-automatic to withdraw, in the answer, the sentences that start with an article, clause, or point. Then, we use the manual method to make the extracted sentence more accurate. After that, we separate that sentence into the corresponding four components: Article, Clause, Point, and Law document (Code, Resolution, Joint Circular, Circular, Decree, Decision, etc.).

Currently, the data crawled has about 2000 records. In it, there are questions and answers about VSSID applications, administrative procedures, and legal documents. However, with the VSSID application and some administrative procedures that are not related to legal documents, they are not extracted from legal documents.

Legal Documents. The legal documents were collected from Thuvienphapluat². We use a web scraping tool that automates retrieving legal documents from

¹ baohiemxahoi.gov.vn - This website provides information on social insurance policies, social insurance duties and procedures related to social insurance.

² thuvienphapluat.vn - Thuvienphapluat is a Vietnamese website that provides online legal documents of Vietnam and related legal documents.

the website. The collected data were evaluated for their quality and completeness, they were also compared with question data obtained from other sources to assess the reliability of the Law Library as a data source for legal research.

After collecting the legal documents from that webpage, we further process the data by splitting the documents into smaller segments based on the hierarchical structure of the legal text. Specifically, we will identify and extract articles within each document and assign a unique ID to each segment. This process will enable us to organize the legal data in a structured format that can be easily queried and analyzed (Table 1).

3.3 Data Processing

Due to the legal text data and questions collected from the website, many unwanted components exist. We eliminated irrelevant parts and special characters (e.g., new-line and extra white space) to improve the model’s accuracy. But there are specific difficulties in processing raw data. As a result, we have to manually handle and verify the data to ensure its accuracy. This extra step is necessary to avoid errors arising from incomplete or inaccurate data, which could impact our analysis and decision-making process.

After collecting data and filtering information, we obtained a reliable dataset for further analysis and research. The dataset comprises 23 legal documents appropriate to healthcare, social, and unemployment insurance, with 1094 articles and 618 questions.

3.4 Vocabulary Frequency

To gain insights into the field of Vietnamese labor law, WordClouds are used to represent the most commonly occurring words within our dataset visually (Fig. 1 and Fig. 3). Additionally, Fig. 2 and Fig. 4 present the top 10 words with the highest frequencies, all relevant to the legal domain of Vietnamese labor law.

3.5 Feature Extraction and Modeling

Prior to the training phase, we examine **TF-IDF** and **BM25** [20] as the initial methods to tackle the problem. They are statistical methods for estimating the relevance of a given query matching to documents based on the frequency of word occurrences.

Then, we used **word segmentation** from **VnCoreNLP** [23] to improve the performance of **TF-IDF** and **BM25** [20] by segmenting Vietnamese words into meaningful phrases.

Table 1. Different examples for legal texts.

ID	Articles
law1_1	<p>“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.”</p> <p>“Article 1. Scope of regulation This Law provides for the social insurance regime and policies, the rights and responsibilities of employees, employers, relevant agencies, organizations, and individuals involved in social insurance, representative organizations of labor collectives and employers, social insurance agencies, social insurance funds, social insurance implementation procedures, and state management of social insurance.”</p>
law1_2	<p>“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...”</p> <p>“Article 2. Applicable subjects 1. Employees who are Vietnamese citizens are mandatory participants in social insurance, including a) Workers under an indefinite-term labor contract, a definite-term labor contract, a seasonal labor contract, or a specific job with a duration of 03 months or more, but less than 12 months, including labor contracts signed between the employer and a legal representative of a person under 15 years of age, as regulated by labor law;...”</p>
law23_8	<p>Điều 8. Trách nhiệm của Tổ kiểm soát các cấp 1. Tổ kiểm soát cấp huyện 1.1. Kiểm soát việc tuân thủ quy trình thực hiện, hồ Sơ đính kèm, thông tin của người tham gia và thành viên HGD được cập nhật vào phần mềm HGD đảm bảo đầy đủ, chính xác từ đề nghị của cán bộ sổ, thẻ...</p> <p>“Article 8. Responsibilities of Control Boards at all levels 1. Responsibilities of Control Boards at district level 1.1. Control the compliance with the implementation procedures, the attached documents, information of the participants and members of the Household Management Board updated in the Household Management software, ensuring their completeness and accuracy based on the proposal of the bookkeepers, card issuers,...”</p>

As previously stated, each article belonging to a certain law document is granted a unique ID. When we feed a query into the pipeline, it is analyzed, and our objective is to guess which article is linked to the query. Each one may be related to more than one article or be assigned multiple labels simultaneously. This part will introduce fundamental approaches for solving our task. Examples are shown in Table 2. At outperform modeling, our core architecture implements SBERT [19], a sentence embedding model that employs PhoBERT [12], a large-scale monolingual language pre-trained model for Vietnamese. The model is

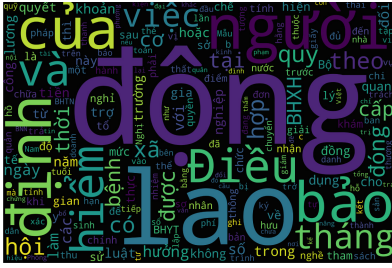


Fig. 1. Wordcloud Legal corpus.

Word	English	Frequency
Lao	Labor (Part-of)	6131
ng	Labor (Part-of)	5540
nh	Regulations (Part-of)	4681
ca	Of	4632
ngi	People	4349
him	insurance (Part-of)	3840
vị	And	3822
bo	insurance (Part-of)	3740
theo	follow	3234
quy	Regulations (Part-of)	3217

Fig. 2. Word Frequency Legal corpus.

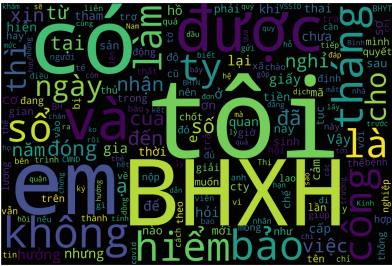


Fig. 3. Wordcloud User Questions.

Word	English	Frequency
tôi	I	1234
cứ	Have	1111
c	Get	994
em	I	803
BHXX	Social insurance (acronym)	792
vị	And	754
bo	Insurance (Part- of)	703
tháng	Month	653
him	Insurance (Part- of)	628
lị	is	620

Fig. 4. Word Frequency User Questions.

trained throughout two stages, as follows: (Fig. 5 depicts the flow of our model in further detail.)

- Stage 1:** Queries and article documents are transformed into vectors through TF-IDF [18] or BM25 [20] algorithm to retrieve relevant and irrelevant article documents to each query which is computed by cosine similarity between vectors. It can be understood that negative samples for each question are generated by selecting the top-n article documents from **Article Retriever TF-IDF/BM25** layer that did not contain the correct answer. The following phase is the SBERT method. First, at **Word Embedding** layer, those samples are embedded by sentence transformer based on PhoBERT [12]. The goal of the next mean **Pooling** layer is to produce fixed-size sentence embeddings. Finally, the **Cosine Similarity Loss** is used to determine how similar two embeddings are. It enables our architecture to be fine-tuned to recognize the similarity between the negative and positive samples (The positive sample includes query and article documents target). This technique, namely contrastive learning, is commonly utilized when there is a shortage of data and the need to heighten the contrast between the positive and negative ones. After training, our system can make reasonably accurate predictions.

Table 2. A query sample and its corresponding ground-truth labels as well as prediction of our outperforming system.

Query	Ground-truth labels	Prediction
<p>“Bố em làm việc cơ quan và tham gia bảo hiểm 35 năm.Nay được quyết định nghỉ hưu. Vậy khi nào bố em được lãnh tiền bảo hiểm và lãnh thế nào.có được lãnh BHXH 1 lần không?”</p> <p>English translation: <i>“My father worked in a company and participated in social insurance for 35 years. Now he has decided to retire. So when will my father receive the insurance money, and how to receive it? Can he get one-time social insurance?”</i></p>	law1_59	law1_59 law1_13
<p>“Minh đã nghỉ làm ở công ty cũ từ đầu năm 2021, đến hiện tại công ty cũ chưa chốt sổ bảo hiểm cho mình. Hiện tại mình đã đi làm và đóng bảo hiểm tại công ty mới, nhưng trên VSSID chỉ hiện quá trình đóng bảo hiểm tại công ty mới mà không hiện quá trình ở công ty cũ. Cho mình hỏi làm thế nào để lấy lại quá trình trên vssid được ạ. Trong trường hợp công ty cũ làm sai không chốt sổ bảo hiểm cho mình thì quá trình cũ của mình có được tính hay không?”</p> <p>English translation: <i>“I quit my job at my old company in 2021, and until now, the old company has not closed my social insurance. Currently, I have worked and paid insurance at the new company, but on VSSID, only the insurance payment process at the new company is displayed, but not the process at the old company. Please tell me how to get the process back on VSSID. Also, in case the old company makes a mistake and does not close the insurance book for me, will my old process be counted or not?”</i></p>	law13_46.96 law1_21 law3_48	law13_46.96 law1_21 law3_48

- (b) **Stage 2:** Negative samples for each question are generated once again by selecting the top-n documents, which are embedded from **SBERT** [19] trained in **stage 1**. This phase quickly and effectively provides our model more embedded negative data samples, which are of high quality as well. Furthermore, our model gets strengthened, reinforced, and fine-tuned due to learning the contrastive information in those data.

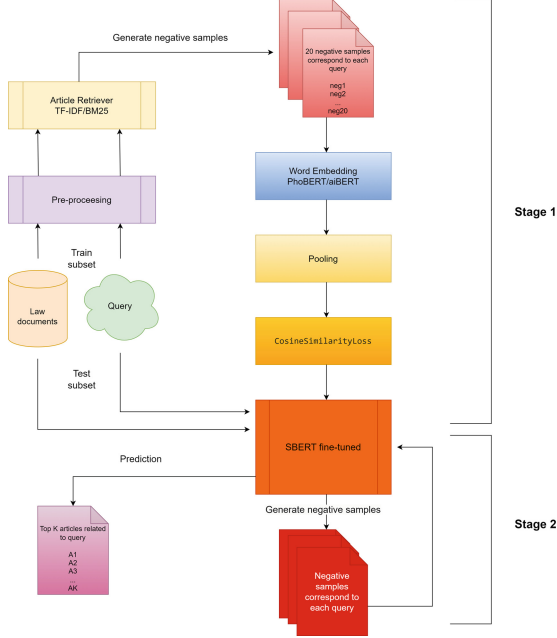


Fig. 5. TF-IDF/BM25 + SBERT system (Explained at **Feature extraction and modeling** subsection of Sect. 3.).

3.6 Metric Evaluation

We evaluate the performance of different approaches using $Top_K@acc$ as the metric. Accuracy is calculated as the proportion of questions with all correct labels in the **Top** K documents returned by our methods. L_K is a collection containing k labels, or IDs of Article, which our system predicts are most related to the query, l_q is the query's actual collection of labels.

$$Top_K@acc = \frac{1}{n} \sum_1^n \begin{cases} 1, & l_q \subseteq L_K \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

4 Experiments

4.1 Experimental Design

The following experimental is designed to compare the performance of different methods. In the word embedding phase, the proposed model utilizes **vinai/phobert-base** [12], **fptai/vibert** [3] with a maximum sequence length of 256. SBERT is fine-tuned using three epochs and a batch size of 16. On the local machine, we run our tests on a single NVIDIA RTX 3060.

TF-IDF [18] is a numerical statistic that indicates how important a lexical unit is in a document or text in a dataset. Specifically, the TF-IDF weight is composed of two terms: the first term computes the normalized term frequency (TF) is defined as the number of times a word appears in a document divided by the total number of words in that document; the second the term is the Inverse Document Frequency (IDF). Practically, TF-IDF can be computed as follows:

$$\begin{aligned} tfidf(t, d, D) &= tf(t, d) \times idf(t, D) \\ tf(t, D) &= \log(1 + freq(t, d)) \\ idf(t, D) &= \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right), \end{aligned}$$

where t is a unigram or bigram term in a document d from a collection of documents D . $freq(t, d)$ measures how many times t a term appears in d .

BM25 [20] is an improved variant of TF-IDF. BM25 has two new parameters: k , which helps balance the relevance of term frequency and IDF, and b , which modifies the weight of document length normalization. Suggested values are $k = [1.2, 2.0]$ and $b = 0.75$.

$$BM25(D, Q) = \sum_{i=1}^n \log\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right) \times \frac{f(q_i, D) \cdot (k + 1)}{f(q_i, D) + k \cdot (1 - b + b \cdot \frac{|D|}{avgdl})}$$

where $f(q_i, D)$ is the frequency with which q_i appears in document D . $|D|$ is the number of words in D , and $avgdl$ is the average document length in the text collection from which documents are drawn; N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i .

SBERT [19] is a framework for computing sentence embeddings using BERT [5] models. Sentence-BERT modifies the original BERT model using a Siamese or triplet network structure. These are networks that share weights and can encode multiple inputs at once. **PhoBERT** [12] and **viBERT** [3], pretrained BERT embeddings for the Vietnamese language, has been developed by VinAI and FPTAI, respectively.

Cosine similarity can be computed by the following equation:

$$similarity(q, d_i) = \frac{q \cdot d_i}{\|q\| \|d_i\|}, \quad (2)$$

where q is the vector form of the query and d_i is vector form of the article i from a collection of law documents D ($d \in D$).

CosineSimilarityLoss

$$similarityLoss = \frac{1}{N} * \sum (label - \tanh(similarity(e_i, e_j)))^2 \quad (3)$$

where N is the number of pairs, the *label* is the target similarity (-1 or 1), *tanh* is the hyperbolic tangent function, *similarity* is the cosine similarity function, and e_i is the vector form of the article i .

4.2 Results

We conduct different experiments to compare these approaches in the collected legislation dataset. This dataset mainly focuses on a particular subject, the Labor Code of Viet Nam that we have recently formed. Table 3 displays the results of eight different methods, where each notation stands for every distinct approach as follows. **TIWS** employs TF-IDF to retrieve the top K most relevant article documents to the query, where each sentence can be tokenized as tokens by Word Segmentation from Undersea library. Similarly, **BMWS** is the approach where BM25 replaces TF-IDF and combines with Word Segmentation.

For **TPS1**, TF-IDF/BM25 is used to generate negative samples, and then they are embedded by PhoBERT and fed into the model for training. After the training process at Stage 1 is completed, one can get PhoBERT fine-tuned. This fine-tuned one is used to embed article documents and queries in the testing dataset and to predict which documents are most relevant to the question using cosine similarity. By the inheritance of PhoBERT fine-tuned from the previous **TPS1**, **TPS2** continuously re-embeds all of the article documents and queries from the training subset, and negative samples are generated based on the similarity between article and query embedding vectors. After that, they are fed into the model for training the second time, and PhoBERT fine-tuned-v2 is created. Finally, this fine-tuned-v2 can embed and predict the relevant article documents and queries efficiently.

TvS1 is similar to the **TPS1**. The only change between this model and **TPS1** is that viBERT has replaced PhoBERT. Finally, **TvS2** is analogous to the **TPS2**. Again, the only change between this model and **TPS2** is that viBERT has replaced PhoBERT.

As illustrated in Table 3, the proposed model combining TF-IDF/BM25 and SBERT achieves outperformance on the dataset. Given that TF-IDF + SBERT and BM25+ SBERT have virtually identical scores, we only display one representative on the table. The score increases when k in $Top_K@acc$ increases.

Table 3. Evaluation results on our dataset using several techniques with varying $Top_K@acc$.

Methods	$Top_5@acc$	$Top_{10}@acc$	$Top_{20}@acc$	$Top_{50}@acc$	$Top_{100}@acc$
TF-IDF	0.0356	0.0777	0.1472	0.3867	0.4919
BM25	0.0307	0.0728	0.1634	0.3155	0.4660
TIWS	0.0518	0.0728	0.1408	0.3608	0.4854
BMWS	0.0388	0.0841	0.2023	0.3414	0.4693
TPS1	0.4345	0.4935	0.5575	0.6812	0.7540
TPS2	0.5518	0.5939	0.6359	0.6618	0.6942
TvS1	0.2335	0.2875	0.3620	0.5475	0.7691
TvS2	0.2573	0.3020	0.3921	0.5021	0.5523

5 Conclusion and Future Work

This study proposed a method to design intelligent retrieval systems on law documents based on NLP approaches, TF-IDF, BM25 and BERT. Although each technique did not work well on legal domain, the performance of model is enhanced when combining TF-IDF/BM25 and BERT. Thus, the combination model is emerging to serve as a foundation for future text retrieval models geared at legal inquiries. Ultimately, the efforts will result in a more robust and efficient legal retrieval system.

In the future work, the proposed model will continue to improve and experiment with new techniques to increase its performance. It can be incorporated a knowledge graph [15] to assist in the search for relevant documents and improve the overall efficiency of the system.

References

1. National Assembly: Labor on Employment 2013, No. 38/2013/QH13 (2013)
2. National Assembly: Labor Code 2019, No. 45/2019/QH14 (2019)
3. Bui, T.V., Tran, O.T., Le-Hong, P.: Improving sequence tagging for Vietnamese text using transformer-based neural models. CoRR abs/2006.15994 (2020). <https://arxiv.org/abs/2006.15994>
4. Dale, R.: Law and word order: NLP in legal tech. *Nat. Lang. Eng.* **25**(1), 211–217 (2019)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, June 2019
6. Fernández-Barrera, M., Sartor, G.: The legal theory perspective: doctrinal conceptual systems vs. computational ontologies. In: Sartor, G., Casanovas, P., Bia-siotti, M., Fernández-Barrera, M. (eds.) *Approaches to Legal Ontologies*. Law, Governance and Technology Series, vol. 1, pp. 15–47. Springer, Dordrecht (2011). https://doi.org/10.1007/978-94-007-0120-5_2
7. Vietnam Government: Decree on Detailing Unemployment Insurance of the Law on employment - No. 28/2015/ND-CP (2015)
8. Vietnam Government: Decree on detailing and guiding the implementation of a number of articles of the labour code regarding working conditions and labour relations, No. 145/2020/ND-CP (2020)
9. Le, T.A.T., Vodden, K., Wu, J., Atiweh, G.: Policy responses to the covid-19 pandemic in Vietnam. *Int. J. Environ. Res. Public Health* **18**(2), 559 (2021)
10. Mirończuk, M.M.: The BigGrams: the semi-supervised information extraction system from html: an improvement in the wrapper induction. *Knowl. Inf. Syst.* **54**(3), 711–776 (2018)
11. Ngo, H., Nguyen, T., Nguyen, D., et al.: AimeLaw at ALQAC 2021: enriching neural network models with legal-domain knowledge. In: 2021 13th International Conference on Knowledge and Systems Engineering (KSE). IEEE (2021)
12. Nguyen, D.Q., Nguyen, A.T.: PhoBERT: pre-trained language models for Vietnamese. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1037–1042 (2020)

13. Nguyen, H., Tran, T.V., Pham, X.T., Huynh, A.: Ontology-based integration of knowledge base for building an intelligent searching chatbot. *Sens. Mater.* **33**(9), 3101–3123 (2021)
14. Nguyen, H.D., Pham, V.T., Le, T.T., Tran, D.H.: A mathematical approach for representing knowledge about relations and its application. In: *Proceedings of 7th International Conference on Knowledge and Systems Engineering (KSE 2015)*, pp. 324–327. IEEE (2015)
15. Nguyen, T.H., Nguyen, H.D., Pham, V.T., Tran, D.A., Selamat, A.: Legal-Onto: an ontology-based model for representing the knowledge of a legal document. In: *Proceedings of 17th Evaluation of Novel Approaches to Software Engineering (ENASE 2022)*, Online streaming, pp. 426–434 (2022)
16. de Oliveira Rodrigues, C.M., de Freitas, F.L.G., Barreiros, E.F.S., de Azevedo, R.R., de Almeida Filho, A.T.: Legal ontologies over time: a systematic mapping study. *Exp. Syst. Appl.* **130**, 12–30 (2019)
17. Pham, V.T., Nguyen, H.D., Le, T., et al.: Ontology-based solution for building an intelligent searching system on traffic law documents. In: *Proceedings of 15th International Conference on Agents and Artificial Intelligence (ICAART 2023)*, Lisbon, Portugal, pp. 217–224 (2023)
18. Qaiser, S., Ali, R.: Text mining: use of TF-IDF to examine the relevance of words to documents. *Int. J. Comput. Appl.* **181**(1), 25–29 (2018)
19. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics* (2019). <https://arxiv.org/abs/1908.10084>
20. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. *Foundations Trends Inf. Retrieval* **3**, 333–389 (2009). <https://doi.org/10.1561/15000000019>
21. Trinh, N.T.T.: Impact of the Covid-19 on the labor market in Vietnam. *Int. J. Health Sci.* **6**, 6355–6367 (2022)
22. Villata, S., Araszkievicz, M., Ashley, K., et al.: Thirty years of artificial intelligence and law: the third decade. *Artif. Intell. Law* **30**, 561–591 (2022)
23. Vu, T., Nguyen, D.Q., Nguyen, D.Q., Dras, M., Johnson, M.: VnCoreNLP: a Vietnamese natural language processing toolkit. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 56–60. Association for Computational Linguistics, New Orleans, Louisiana, June 2018. <https://doi.org/10.18653/v1/N18-5012>. <https://aclanthology.org/N18-5012>
24. Zhao, G., Liu, Y., Erdun, E.: Review on intelligent processing technologies of legal documents. In: Sun, X., Zhang, X., Xia, Z., Bertino, E. (eds.) *Artificial Intelligence and Security: 8th International Conference, ICAIS 2022, Qinghai, China, 15–20 July 2022, Proceedings, Part I*, pp. 684–695. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-06794-5_55
25. Zhong, H., Xiao, C., Tu, C., et al.: How does NLP benefit legal system: a summary of legal artificial intelligence. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (COLING 2020)*, pp. 5218–5230. Association for Computational Linguistics (2020)