*Science & Technology Development Journal – Engineering and Technology 2025, 9(1):2702-2714*

Open Access Full Text Article

*Research Article*

# NLP-Method for Identifying and Extracting Vietnamese Education Legal Documents Based on OCR and XML Techniques

**Nguyen Van Sinh *, Phan Anh Kiet , Nguyen Thanh Tuan , Nguyen Thi Thanh Sang , Le Thanh Son**

Use your smartphone to scan this QR code and download this article

**ABSTRACT**

Digital transformation in education requires intelligent systems for extracting, standardizing, and analyzing large volumes of legal documents in various formats. The structure of Vietnamese legal documents in education is complex, with different components, including national emblems, issuing agencies, symbols, dates of issuance, clauses, appendices, and sometimes signatures or handwritten notes. Besides, the difference has also come from their styles, such as constitutions, laws, decrees, circulars, decisions, etc. In addition, the decree also specifically regulates the document format (font, font size, margin, line spacing), arrangement of components, and rules for presenting appendices or attached documents. Therefore, processing these documents poses many challenges in information retrieval and legal data management. Natural language processing (NLP) plays a crucial role in such text and document processing. In this paper, we propose a novel approach integrating NLP, Optical Character Recognition (OCR), and image processing to process Vietnamese legal documents in the education section. This foundation serves a future optimal information retrieval system using Large Language Models (LLMs) and Generative Artificial Intelligence (GenAI). Our method includes the following steps: (1) collecting the legal document database in PDF format from the website of legal documents of the Ministry of Education and Training (MOET); (2) removing noise, segmenting different components, and converting into plain text via OCR and image processing techniques. (3) structuring the extracted text into XML format for our future system. Compared to the existing application, our system achieves an expected accuracy of over 99% with printed documents, including the ability to recognize handwritten text. This study is a step toward developing a technical solution for a data platform that enables the intelligent application of LLM and GenAI in an optimized database search engine for informed decision-making based on legal documents.

**Key words:** Artificial Intelligence, Legal Documents, Document Recognition, Natural Language Processing, Data Mining, Optical Character Recognition.

*International University - Vietnam National University of Ho Chi Minh City, Vietnam.*

**Correspondence**

**Nguyen Van Sinh** , International University - Vietnam National University of Ho Chi Minh City, Vietnam.

Email: nvsinh@hcmiu.edu.vn

**History**

- Received: 24-07-2025
- Revised: 16-10-2025
- Accepted: 22-12-2025
- Published Online: 31-12-2025

**DOI :**

**https://doi.org/10.32508/stdjet.v9i1.1536**

Check for updates



VNUHCM PRESS

## INTRODUCTION

The rapid development of information technology (IT) and artificial intelligence (AI) has created the premise for more effective digitization and exploitation of legal document data, especially in the context of strong digital transformation in the education sector. In Vietnam, the MOET issues a large number of legal documents yearly, such as circulars, decisions, instructions, etc., to regulate and manage educational activities nationwide. Due to their legal nature, they are typically formatted as non-editable PDF files or scanned versions. This format poses challenges for search engines, limiting not only accessing content but also the contextual linkages between documents.. Tools like Google or DeepSeek (deepseek.ai) are well-known, popular tools for searching information but the obtained results are not really extended connecting to the related information in the same context. They are more and more developed with AI support to adapt and provide full information that is more closer to the exact information of users.

In this research, our work builds upon and extends the OCR method developed by Nguyen et al.[1] for processing administrative documents, which utilizes a system combining Tesseract OCR with image processing techniques (OpenCV) to recognize printed documents. However, instead of just stopping at character extraction, the current study continues to process output documents according to legal structure, including recognizing document clauses, determining document types (circulars, decrees, decisions, etc.), and exporting to a standardized XML format. Besides, to continue our research project (Optimal Information Retrieval System from Education Document Database Based on Large Language Models and Content Generating Artificial Intelligence), a part of our contribution is proposed and successfully published in the research,[2] where the autoencoders (Topic

**Cite this article :** N V S, P A K, N T T, N T T S, L T S. **NLP-Method for Identifying and Extracting Vietnamese Education Legal Documents Based on OCR and XML Techniques**. *Sci. Tech. Dev. J. – Engineering and Technology* 2026; 9(1):2702-2714.

VAE) and Sentence-BERT (SBERT) are used as a tokenizer through specialized Vietnamese text preprocessing and a weighted integration mechanism that balances probabilistic modeling with contextual semantics. Unlike previous studies that only focus on character recognition, the model proposed in this paper focuses on the ability to integrate legal language into AI systems, thereby supporting the construction of a training database for deep learning models specialized for Vietnamese in the fields of education laws. In detail, we study and propose a comprehensive processing procedure to transform educational legal documents from PDF format to XML structured data based on OCR and Image processing to serve for future intelligent application of NLP. The method consists of three main steps: (1) First step is collecting the legal documents in PDF format by proposing an algorithm to search, crawl and download automatically legal documents from the website of MOET. (2) The next step is to process and convert files from PDF format (or scanned image files) to text using OCR and image processing techniques. This step is removing noise data and filtering the necessary information in the content of legal documents based on image processing techniques. Then, the content is automatically classified into the different components of a template form, following the structured format of the Vietnam Government. (3) The final step involves normalizing the text into a plain text format and converting it into a structured XML format, accurately reflecting legal components such as document type, issuing agency, clauses, items, and other relevant details. The XML data is then used to train specialized language models in the legal field, contributing to the construction of intelligent search systems, document classification, and policy decision support.

The remainder of this article is structured as follows: the next section is Related work, where we review the state-of-the-art (SOTA) methods with the same context. Section 3 is detailly of our proposed method. We present our implementation and the obtained results in Section 4. After that, the discussion, comparison, and evaluation of our proposed method are presented in section 5. The last section is the Conclusion and future work.

## RELATED WORK

This section presents SOTA methods for NLP in general and specially for Vietnamese language based on OCR, Image processing and Machine learning techniques. NLP is increasingly proven to be an important research field and is widely applied in various fields of practice. Several applications, such as Chatbots, Voicebots, Google Translate, Text-to-speech, and Speech-to-text, are rapidly being developed. They are not only supportive in daily life but also in particular fields such as law, medicine, tourism, and administrative management. The process of digitizing administrative and legal documents is an important step in the digital transformation roadmap of organizations and government agencies. In particular, converting documents from image or PDF format to structured digital text plays a fundamental role in exploiting, storing, and analyzing data using natural language processing (NLP) tools or deep learning models.[3,4] The following researches provide us with an overview of digital document processing in practice.

### OCR and Image Processing

The research of Johan et al.[5,6] built a system for digitizing and evaluating student graduation projects at higher education institutions. The authors proposed a model using Tesseract OCR technique combined with cloud storage, in which academic documents are scanned and converted into searchable digital text and stored online. The system allows lecturers and grading boards to access and evaluate projects quickly. Although the scope of the research is academic and applied in an internal educational environment, the experimental results show that the system is capable of recognizing printed characters with stable accuracy. However, the model does not deeply handle issues related to legal structure, text segmentation or information extraction according to specialized data fields.

In another study, Pate et al.[7] conducted an experimental evaluation of the performance of Tesseract OCR, one of the most popular open–source tools for character recognition. The study demonstrated that the quality of the input (resolution, contrast) and the language used have a direct impact on the accuracy of the OCR results. Tesseract performs well with clear printed text, but is prone to errors with italics, handwriting, or special characters. The authors also emphasized that the ability to recognize accented language remains limited without a well-trained language model, which requires image preprocessing and appropriate language configuration when applied to particular languages.

Nguyen et al.[1] proposed an approach more suitable for the administrative documents in Vietnam. The authors developed a web application integrating Tesseract OCR and OpenCV, designed to digitize legal administrative documents in the national standard format of Decree 30/2020/ND-CP (the new decree, N$^o$

78/2025/NĐ-CP is updated on April 01, 2025).[8] The system performs steps such as converting PDF to image, filtering noise, increasing contrast, and segmenting the text structure to identify information components such as: national emblem, issuing agency, document number, date of issue, signature, seal, and main content. The results show that the system achieves an accuracy of over 91% with printed documents. However, the study did not organize output data in structured formats such as XML/CSV, and also did not target advanced NLP applications such as language model training.

Koichi Kise[9,10] studied in depth the technique of document image segmentation, an important step to increase the accuracy of text recognition from scanned images. By separating document images into homogeneous functional blocks (such as text paragraphs, tables, and images), the system can process the appropriate components according to each specific method. The authors applied image processing techniques to analyze the background, foreground, color and pixel intensity, thereby optimizing the ability to separate lines and columns in complex texts. This method created the premise for the following OCR steps to be more accurate, especially in documents with non-standard layouts.

In the field of image processing, Chung BW[11] provides a detailed practical guide on how to install and use the OpenCV library, one of the most powerful open source tools for image processing in Python. Although the article is inclined a technical guide than a theoretical study, it plays a fundamental role in document image processing systems. Many current OCR systems integrate OpenCV for steps such as preprocessing, noise filtering, and text region normalization.

In addition, image processing applications are also deployed in other contexts. Tran et al.[12] applied imaging processing techniques to construct a virtual museum, where artifacts are digitized in 3D and presented online. Although it is not directly related to OCR, this study shows the strong ability of image processing in digital transformation of traditional content. Similarly, Sinh et al.[13]built a system to visualize image data in a medical application, thereby supporting diagnosis and doctor-patient interaction. Although these two studies do not go into text processing, they affirm the wide application value of image processing technology, creating a basis for potential expansions in the legal and educational fields.

## Digitizing legal documents and natural language processing

The research of Blei et al.[14] laid the foundation for topic modeling through the Latent Dirichlet Allocation (LDA) model. This model assumes that each document is a probability distribution of topics, and each topic is a probability distribution of vocabulary. Thanks to the ability to mine latent topics from large text collections without manual labeling, LDA has become a popular method in legal text processing, supporting effective organization and data exploration. However, LDA assumes that words are not context-dependent and therefore does not capture semantic nuances, which are very important in Vietnamese legal documents. To overcome that weakness, Dieng et al.[15] proposed Topic VAE (Variational Autoencoder), an approach that uses neural networks to represent topics in the embedding space. Topic VAE combines the advantages of traditional topic modeling with the flexibility of deep learning models, allowing modeling of nonlinear relationships between words and topics. When applied to legal data, Topic VAE allows the extraction of complex hierarchical legal topics such as administrative regulations, student rights, output standards, etc. However, this model requires large training data and needs well-standardized input, especially with accented languages such as Vietnamese. In the research[16], Grootendorst developed a BERTopic, a topic model that leverages semantic representations from pre-trained language models such as BERT or SBERT to increase topic coherence. BERTopic applies dimensionality reduction and clustering techniques to group short documents by topic. This method is suitable for processing the legal documents divided by clauses, paragraphs, sentences, semantics grammar. However, BERTopic is mainly trained on English, and its application to Vietnamese requires appropriate embedding adjustments (e.g. using PhoBERT or multilingual Sentence-BERT), as well as a well-structured input data, which the current research is addressed by constructing XML from legal documents. Chalkidis et al.[16] pioneered the development of a specialized legal language model with LEGAL-BERT. This model was trained on a large English legal corpus (EU Legislation, US court decisions, etc.), showing superiority in text classification, legal Q&A, and entity recognition tasks. However, LEGAL-BERT[17] does not support Vietnamese and is difficult to apply directly due to differences in legal systems, language syntax, and legal expressions.

For the Vietnamese language, Nguyen and Tuan [18] developed PhoBERT, the first BERT model trained entirely for Vietnamese corpus. PhoBERT achieved high results in tasks such as text classification, entity labeling, and sentiment analysis. However, PhoBERT was not trained on a legal corpus, and therefore it is not clearly recognized specialized entities such as "Article", "Clause", "Circular", or the name of the issuing agency. Furthermore, this model requires input in the form of plain text with clear quality and structure. Therefore, the preprocessing and normalization steps from PDF to XML in this study play a very important role. Complementing PhoBERT, the vnCoreNLP tool by Vu et al. [19] provides Vietnamese language processing modules such as word segmentation, part-of-speech tagging, dependency parsing, and entity recognition. This is an effective toolkit for general texts, but is limited when applied to legal texts with complex structures, formal language, and many normative provisions.

In the next research, Tewari [20] recently presented the LegalPro-BERT model, a refined version of BERT Large for legal clause classification. The model is applied to well-structured legal documents with specific legal topic labels. This demonstrates that effective legal NLP models require standardized, structured (e.g., XML/JSON) and well-labeled legal training data, which is also the output data target in the current study.

The current research demonstrates that the application of OCR technology, image processing, and language models in the fields of text digitization and legal text analysis is widespread and varied in practice. However, when examining the context of Vietnamese legal documents in the field of education specifically, significant gaps remain. For example, in the research [1,5] , authors focused on character recognition and basic administrative structures, but did not incorporate specialized legal recognition such as clauses, document types, or issuing agencies. This limitation affects the ability of organizational data in in-depth analysis. One more issue is that, although many studies are developing NLP models for legal documents (such as Legal-BERT [16] , LegalPro-BERT [20] ), these models are mainly built on English corpus, no model has been fully trained on Vietnamese legal documents, especially in the field of education, which has its context, semantic, structure and terminology. Besides, most legal documents still exist in PDF format. To process the content in each of them, there is a need to develop an application that automatically converts them into structured data (XML/CSV). After that, the NLP model is applied for processing. Models such as PhoBERT [16] , BERTopic, [15] or vnCoreNLP [19] all require clean, consistent, and well-structured input data. Last but not least, the problem is that a study has proposed a closed process from PDF data collection to a process based on OCR, [21,22] analyzing the legal structure, and converting it into an XML structure. Especially applicable to documents in the Vietnamese education legal system, where there is a large volume of documents, constant changes, and it is subject to strict management by public authorities.

## PROPOSED METHOD

### Overview

In this study, we propose a sequential multi-stages system to identify, extract and structure text data from legal documents into XML structure files. The process starts by converting the input PDF file into image sequences corresponding to each page of the document. Then, OCR and SVM techniques are applied to determine and extract text content from each image contenting both printed and handwritten text. The obtained text data will be processed to standardize and remove noise data, thereby unifying the structure. Finally, the system labels the data in XML format for structured representation, serving the purposes of later search and analysis. The algorithms are implemented in Python with a modular architecture, supporting flexible expansion (see Figure 1).

### Data collection and image conversion

This section presents our step to collect data from the website of Vietnam Ministry of Education and Training (MOET), an official site provides legal documents in education field (https://moet.gov.vn). The legal document data have been collected through an automatic collection algorithm using Python source code. After collecting, the scanned PDF file is processed through a multi-step process. First, each PDF page is converted to a raster image (in PNG or JPG format) with high resolution to ensure the input quality for the next processing steps. The next step is then applied image processing techniques such as contrast enhancement, noise reduction, and grayscale image conversion to highlight text areas. This step plays an important role in improving the efficiency of the OCR process in the later steps. The algorithm (Algorithm 1) presents steps to automatically collect legal documents data files on the website of MOET as follows:

**Algorithm 1. DataCollection():**

1: **Input**: base_webpage_url of the MOET website
2: **Output**: Set of PDF files downloaded
3: **Set** current_page_number = 1

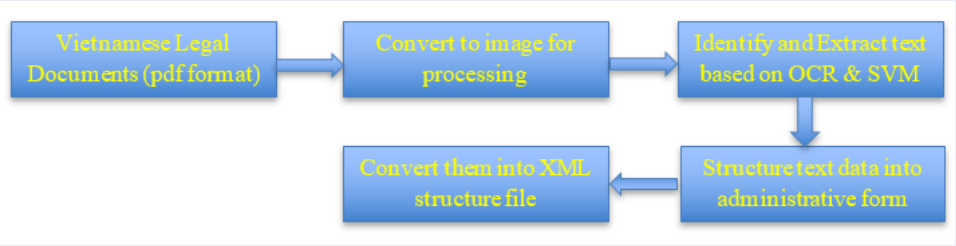


**Figure 1**: Our proposed method for legal text processing and tagging [Source: Sinh et.al.]

4: **Set** base_site_url = "https://moet.gov.vn/van-ban/?Page="
5: **while** current_page_number is not NULL do:
6: **Visit**: base_site_url + current_page_number
7: // Example: https://moet.gov.vn/van-ban/?Page=1
8: Parse the HTML content of the current page
9: **For each** element in the HTML content:
10: **If** the element text contains .pdf file:
11: Extract the 'href' link of the PDF file from the ele-ment
12: // Example href: "/TW/Pages/vbpq-van-ban-goc.aspx?ItemID=..."
13: Download the PDF file
14: **End if**
15: **End for**
16: **If** "next page" link is found in the page:
17: current_page_number++
18: **End if**
19: **End while** (when current_page_number = NULL)
20: Return: Set of PDF files downloaded

## Extraction of text data based on OCR

To effectively extract text from scanned legal documents, the system employs an OCR pipeline that begins with region detection and classification. After preprocessing and segmenting the document into text-containing regions, each region is analyzed to determine whether it contains printed or handwritten text. A lightweight Support Vector Machine (SVM) classifier is applied[20], using visual features such as stroke curvature, pixel density, and alignment patterns to distinguish between printed and handwritten segments: (1) For printed text, the system utilizes Tesseract OCR (v4.1.1), which supports Vietnamese and is optimized for high-accuracy recognition of structured, machine-printed characters. (2) For handwritten text, such as signatures, annotations, or filled-in forms, the system integrates a Convolutional Recurrent Neural Network (CRNN) model, which combines CNN for feature extraction, Bi-LSTM for sequence modeling, and CTC Loss for decoding unsegmented text.

This two-ways recognition strategy ensures that each text region is processed with the most suitable OCR model, significantly improving overall recognition accuracy for heterogeneous legal document formats (see Figure 2).

### *Image processing*

The process begins by converting each PDF page to a high-resolution PNG or JPG to ensure quality input for next steps. The image is then passed through preprocessing techniques to increase contrast, remove noise, and convert to grayscale. Next, the system applies adaptive thresholding to separate the text from the background. Dilation is then used to thicken the text and connect the broken parts. Finally, contour detection is used to identify text areas, serving the subsequent character recognition and content classification steps.

### *Region classification using SVM*

After image preprocessing, the system detects and segments text areas. Techniques such as contour detection or lightweight object detection models are used to accurately determine the location of each block of text. For each detected area, the system extracts image features including stroke curvature, pixel density, and alignment direction. In order to optimize character recognition accuracy, these regions are classified into two groups: printed and handwritten, using an SVM (Support Vector Machine) classifier. SVM distinguishes based on features such as stroke curvature, pixel density and alignment - following the approach of [SVM]. Separating these two types of text allows the system to apply the corresponding OCR algorithms: Tesseract OCR for printed text and CRNN for handwritten text.

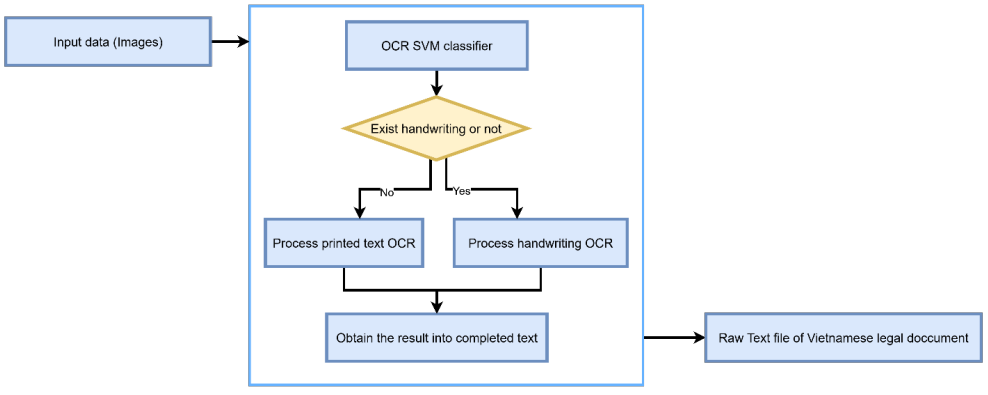


**Figure 2: The recognition and extraction of legal text components from legaldocuments. [Source: Sinh et.al.]**

### *Handwriting recognition CRNN integrated with OCR*

For text regions identified as handwritten, the system uses a deep learning model CRNN combined with OCR techniques to extract content. This model is especially suitable for handwritten text with irregularities with spacing and character shape - factors that often cause difficulties for traditional OCR methods such as Tesseract. In this study, the CRNN model is integrated from existing open source code, pre-trained for Vietnamese handwriting recognition.[23,24] After the recognition process is completed, the output will be normalized and merged with the text extracted from the printed text to reproduce the full content of the legal document in the form of digitized text.

### *Structured Data into an Administrative Form*

Once the recognition process is complete, all text results are reassembled based on their original positions in the source image. This arrangement ensures that the final text is as close in structure and order as possible to the original document, which is good for subsequent processing steps such as clause extraction, legal analysis, or data archiving. This structuring step is essential for converting raw text into machine-readable and semantically tagged data that can be visualized and edited via a web-based interface.

The structured content is automatically filled into a dynamic web data form. Each field is then corresponded to one of the XML tags in the last step. This interface enables legal professionals to ensure that the extracted data accurately reflects the document's original structure and intent. This approach facilitates better readability, efficient querying, and interoperability with legal databases and information systems, laying the foundation for advanced legal analytics and digital archiving. The below algorithm (Algorithm 2) describes how to convert a pdf file to a text file:

**Algorithm 2. OCR_Processing(image_file)**

1: **Input**: image_file (it is scanned from PDF legal document)
2: **Output**: Full legal text
3: Initialize text_output
4: **for** each page_image **do**
5: regions = DetectTextRegions(page_image) // usingSVM
6: **for** each region in regions **do**
7: features = ExtractVisualFeatures(region) //usingOCR
8: **if** IsPrintedText(features) **then**
9: text = PrintedText
10: **else**
11: text = HandwritingText
12: **end if**
13: Append text to text_output in proper order
14: **end for**
15: **end** for
16: **Return** text_output

### Convert into XML structure

The final step is to convert text file into its XML structure. The label tagging is an important step in the process of digitizing legal documents, allowing the extraction of meaningful components from raw data (in .txt format) and reorganizing them into a structured format, specifically XML. The goal of this step is to identify, classify, and represent information units such as document numbers, items, abstracts, and legal contexts to serve data mining, semantic search, and automated legal analysis tasks. This transformation is driven by a keyword-based

mapping algorithm that identifies and segments critical components of Vietnamese legal documents. The algorithm is first scanned the text and extracted segments that match predefined structural patterns (e.g., "Số:", "Điều", "Về việc", or date expressions like "ngày... tháng... năm..."). Each of these extracted segments is then aligned with a corresponding XML tag, including but not limited to following items: <so_hieu>, <quoc_hieu>, <co_quan_ban_hanh>, <ngay_ban_hanh>, <trich_yeu>, <dieu>, and <noi_dung_chinh> (see Table 1, its structure is followed the decree No 78/2025/NĐ-CP).

The conversion process ensures that the document's original legal structure is preserved while enhancing its usability for further computational processing such as search, indexing, and data mining. Once the matching is complete, the tagged content is serialized into a well-formed XML document. The next algorithm (Algorithm 3) is described as follows:

**Algorithm 3. XMLconvert():**

1: **Input**: Plain legal text file
2: **Output**: Structured XML file
3: Extract all components of the legal document
4: **For each** component in the legal document
5: **If** keywords match extracted segments
6: Map matched components to corresponding XMLtags
7: **End if**
8: Convert the structured content into a valid XMLdocument
9: **End for**

## IMPLEMENTATION AND RESULTS

This section presents our implementation for collecting, processing and converting legal documents into structured XML files. We develop a web-app based on Flask framework. The data collection step is performed based on Algorithm 1. We deployed a dataset of 300 legal documents in the education sector downloaded from the Portal of MOET. These documents are mainly Circulars, Decisions, Decrees, and guidance documents in PDF format. The next algorithm (Algorithm 2) is implemented to identify, extract data components (on the legal document) and transformed them into a data-form segmented based on structure of administrative form. This step is based on image processing technique as OpenCV and NLP as Tesseract OCR (v4.1.1 is supported for Vietnamese). For handwritten regions (such as signatures or annotations), the method integrates an open-source AI model based on CRNN (Convolutional Recurrent Neural Network), which combines CNN, LSTM, and CTC [24] to recognize handwritten strings without requiring character segmentation. The choice between conventional OCR and handwritten recognition is automatically made based on image features and confidence from Tesseract OCR. In the last algorithm (Algorithm 3), we convert them into a structured XML file. We use a Laptop (GIGABYTE G5 MF5, Core I5-13500H, RAM 16GB, GPU RTX 4050, Window 11) with enough configuration and many open-source frameworks, APIs and source code to build our web application successfully. The Web-UI shown in Figure 3 is as follows:

During the testing of the text recognition algorithm, the model was applied to the entire PDF legal documents that have been converted into the images. The application performed recognition on individual words and automatically classified between printed and handwritten letters based on the right appropriate OCR model. The obtained results showed that the application operated stably with a ability to extract Vietnamese content with very high accuracy. Most words, including accented words and special characters, were recognized correctly. In particular, word-by-word processing helps to minimize recognition errors in noisy image regions, and the application of confidence classification helps to optimize the selection of the appropriate OCR algorithm for each specific case. Based on the obtained output results, the overall accuracy is almost very high. However, to ensure objectivity and scientific, it should be noted that the accuracy of the application on the test set is approximately 100%, with insignificant deviations and does not affect the overall content of the document.

## DISCUSSION AND EVALUATION

This section discusses and evaluates our method for extracting the text in the images and converting them into XML tags. To measure the accuracy, we used the following formulas to compute character error rate (CER) and word error rate (WER) as presented in [1], the formula *Accuracy = 1 - (E/C)*, where *E*: number of error characters and *C*: total characters in the document. Otherwise, we can also use the formula to compute for both CER and WER:

$$CER = WER = \frac{S + D + 1}{N} \quad (1)$$

Where:

- S: number of characters (words) that are incorrectly recognized and replaced with another word.

**Table 1: Mappingthe components of a text file into the XML Tags of an XML file [Source: Sinh et.al.]**

| Components | Description | XML Tags |
|---|---|---|
| 1. 1. | National Title (e.g. Socialist Republic of Vietnam) | <quoc_hieu> |
| 1. 2. | Issuing organization | <co_quan_ban_hanh> |
| 1. 3. | Document ID | <ky_hieu_van_ban> |
| 1. 4. | Date of issuance | <ngay_ban_hanh> |
| 1. 5. | Type of document | <loai_van_ban> |
| 1. 6. | Abstract or summary | <trich_yeu_noi_dung> |
| 1. 7. | Main content | <noi_dung_chinh> |
| 1. 8. | Received organization (Recipient) | <noi_nhan> |
| 1. 9. | Full name and Signature of competent person | <nguoi_ky_van_ban> |

- D: number of characters (words) that were not recognized and are thus missing in the OCR output.
- I: number of characters (words) that are incorrectly added in the OCR output.
- N: total number of characters (words) in the ground truth document.

We also use computer processing time to evaluate the efficiency of the application. Depending on the number of pages in each document, the scanning and generating process for a set of images; also the status of documents (with or without noise), the obtained results are show in detail in Table 2 and Table 3.

In general, we tested on 300 legal documents and computed the accuracy of CER and WER, obtaining the ratio at 1.28% and 2.38% respectively. If the documents are close to recent years, the accuracy is higher. On the contrary, the accuracy is reducing if they are far from now (see Table 4).

Additionally, we compared the accuracy of CER and WER using the two models: Tesseract for Printed/Digital text and CRNN for Handwriting on 300 documents. The results obtained are presented in Table 5.

Experimental results show that the proposed application has achieved high accuracy in both tasks: recognizing printed legal documents from PDF files and analyzing to covert output legal documents into XML structured files. To compared with previous research works, our application not only reproduces equivalent results, but also has significant improvements in the NLP field.

In previous research,[1] the OCR system was deployed on Vietnamese administrative documents using Tesseract combined with OpenCV, resulting in a recognition result of printed documents approached the accuracy from 91% to 93%. However, their application is only proposed to process administrative

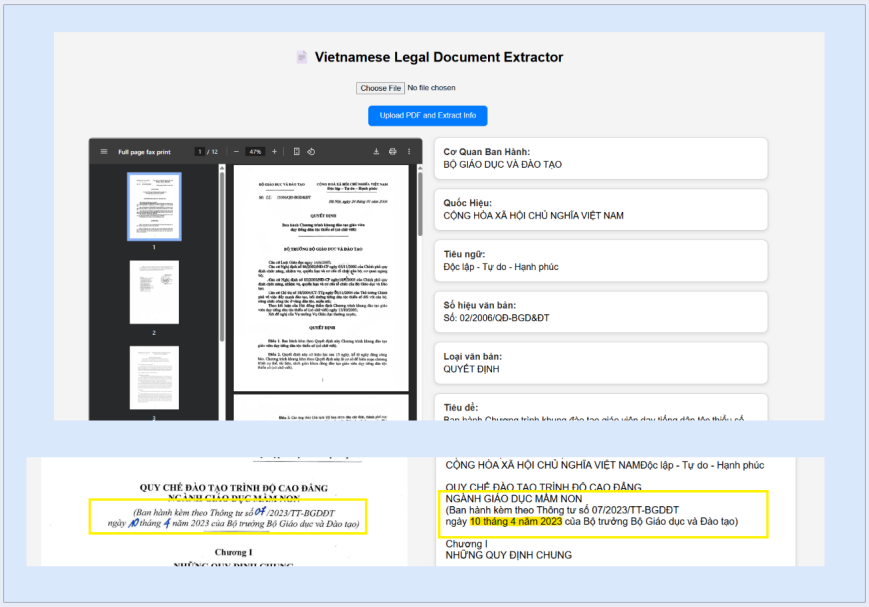


**Figure 3: The web-app forimplementing all steps of our proposed method [Source: Sinh et.al.]**

fields without generating to the structured XML data. In comparison, the application in this research not only improves the accuracy (to 98.7%), but also organizes legal content in the form of hierarchical XML, which is more suitable in topic classification or training legal language models. Similarly, the research in [5] and [6], the academic document digitization system [23] also used Tesseract OCR, mainly for the purpose of archiving and evaluating student projects. Although the character recognition results were quite stable, that system did not address the specifics of legal language, and did not process handwritten text at all. Meanwhile, our application now has expanded scope to the handwritten signature area, with a recognition accuracy met 91.3%. This point proved the efficient integration of CRNN model. Another related study in [22], the author proposed a general text digitization system consisting of the following steps: scanning, indexing, quality checking, and electronic storage. However, their system still processes text in an unstructured form, without performing content analysis in legal format or extracting the detail components. In comparison, the highlight of this study is the combination of parsing legal text and generating XML to serve the training of specific NLP models. In addition, in the field of legal language modeling research, works such as LEGAL-BERT, [16] PhoBERT [16] or LegalPro-BERT [19] all emphasize the role of structured data in training legal AI models. However, most of the above studies have not solved the step of preprocessing input data from PDF into structured data sets. Therefore, our contribution in this research is to create an automatic, closed pipeline from original text files to XML ready for use in NLP. Finally, while many systems only verify a few types of sample documents, the current system has been successfully applied on 300 actual legal documents of MOET with different document types (Circulars, Decisions, Official Dispatches, etc.), that help to ensure generality and high scalability in practical application.

## CONCLUSION

In this study, we proposed a method and successfully built an application to digitize Vietnamese administrative and legal documents based on image processing, computer vision, and artificial intelligence techniques. The application is developed based on our proposed method, using Flask framework, integrating the OpenCV library for image preprocessing and Tesseract OCR for recognizing printed text, combined with an open source CRNN model for handwriting recognition. The application is deployed on a web-app, that helps users upload PDF documents, process and extract text and put into a structured format (XML). During the testing process, the application achieved an average accuracy of over 93% for handwriting and 99.7% for printed text, with a processing time of 9 to 12 seconds per document page, including the recognition and structure analysis steps. This result demonstrates the feasibility, high speed and reliable accuracy of the system, especially in the

**Table 2: Measurethe error ratio of words and characters performed in our proposed method [Source: Sinh et.al.]**

| Input Doc_ID | Quality Status | Application | WER | CER |
|---|---|---|---|---|
| 287/QĐ-BGDĐT.pdf | Low quality, containing printed and handwritten text, making it hard to read. | Ground-truth | 0 | 0 |
| | | Our Application | 0 | 0.0035 |
| | | VietOCR [25] | 0.3645 | 0.306 |
| | | SodaPDF [26] | 0.1888 | 0.1620 |
| | | ABBYY FineReader [27] | 0.0467 | 0.0345 |
| | | Omnipage [28] | 0.1981 | 0.1751 |
| 06_2016_TT-BGDĐT.pdf | The handwriting is of moderate quality, making it hard to recognize. | Ground-truth | 0 | 0 |
| | | Our Application | 0 | 0.0012 |
| | | VietOCR [25] | 0.4756 | 0.2954 |
| | | SodaPDF [26] | 0.8716 | 0.3155 |
| | | ABBYY FineReader [27] | 0.3761 | 0.1436 |
| | | Omnipage [28] | 0.3291 | 0.0956 |
| 12/2023/TT-BGDĐT.pdf | Good quality, minor smudges affect readability. | Ground-truth | 0 | 0 |
| | | Our Application | 0 | 0.0016 |
| | | VietOCR [25] | 0.1359 | 0.1534 |
| | | SodaPDF [26] | 0.0894 | 0.0557 |
| | | ABBYY FineReader [27] | 0.0931 | 0.0557 |
| | | Omnipage [28] | 0.0279 | 0.0016 |
| 04/2018/TTBGDĐT.pdf | High quality, some handwritten text is included. | Ground-truth | 0 | 0 |
| | | Our Application | 0 | 0.0001 |
| | | VietOCR [25] | 0.6355 | 0.2742 |
| | | SodaPDF [26] | 0.2314 | 0.1944 |
| | | ABBYY FineReader [27] | 0.1384 | 0.11 |
| | | Omnipage [28] | 0.1378 | 0.1092 |

context of legal document digitization, where the requirements for content integrity and legal structure are mandatory. The web application based on OCR does not reduce performance, but also enhances user experience to edit and export data. The handwriting processing still needs to be improved in future work for the optimal search engine of our project. With high accuracy, fast processing speed, and easy scalability and integration, the proposed application becomes a potential tool applied in state agencies, administrative organizations, and enterprises in the process of digital transformation for processing legal documents. We will go on to upgrade the deep learning model for handwriting, expand the data set, and improve the ability to recognize different types of documents.

## ACKNOWLEDGMENT

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2025-28-03. We would like to thank for the fund. This research is also supported by The Central Interdisciplinary Laboratory in Electronics and Information Technology (AI and Cooperation

**Table 3: Comparison between our application and existing applications in practice [Source: Sinh et.al.]**

| Tool/Application | Support processing Vietnamese | Printed and handwriting Recognition | XML Structured Output | License | Platform |
|---|---|---|---|---|---|
| Our application | Yes | Yes | Yes | Free | Web-app |
| VietOCR [24] | Yes | No | No | Free | Web-app |
| SodaPDF [25] | Yes | Not clear | No | Trial | Web-app |
| ABBYY FineReader [26] | Yes | Not clear | No | Trial | Desktop |
| Omnipage Docudirect [27] | No | Not clear | No | Trial | Desktop |

**Table 4: Comparingthe accuracy of CER and WER among the number of documents [Source: Sinh et.al.]**

| Years | Quantity of documents | CER | WER |
|---|---|---|---|
| 2010–2015 | 62 | 2.48% | 4.07% |
| 2016–2020 | 98 | 1.21% | 2.33% |
| 2021–2023 | 108 | 0.88% | 1.72% |
| 2024–2025 | 32 | 0.71% | 1.39% |
| Overall | 300 | 1.28% | 2.38% |

**Table 5: Comparing the accuracy of CER and WER using different models [Source: Sinh et.al.]**

| Model | Scope | CER | WER |
|---|---|---|---|
| Tesseract | Printed / Digital text | 0.94% | 1.79% |
| CRNN | Handwriting | 2.61% | 4.88% |

Robot), International University - VNU-HCM. We would like to thank for supporting the machines in experiments.

## ABBREVIATIONS

AI: Artificial Intelligence
LLM: Large Language Model
GenAI: Generative Artificial Intelligence
OCR: Optical Character Recognition
NLP: Natural Language Processing
XML: eXtensible Markup Language
MOET: Ministry of Education and Training
API: Application Programming Interface
SVM: Support Vector Machine
CRNN: Convolutional Recurrent Neural Network
CNN: Convolutional Neural Network
CTC: Connectionist Temporal Classification
LSTM: Long Short-Term Memory
HTML: HyperText Markup Language
CSS: Cascading Style Sheets
JSON: JavaScript Object Notation
CSV: Comma-Separated Values
RNN: Recurrent Neural Network
DL: Deep Learning
ML: Machine Learning
PyTorch: Python Torch (Deep Learning Framework)
OpenCV: Open Source Computer Vision Library
TF: TensorFlow
CER: Character Error Rate
WER: Word Error Rate
VNU-HCM: Vietnam National University Ho Chi Minh City
SOTA: The State of the Art

## CONFLIC OF INTEREST

We declare that: in this paper, there is no conflict of interests

## AUTHORS' CONTRIBUTIONS

Nguyen Van Sinh initialized concepts and directions. Phan Anh Kiet and Nguyen Thanh Tuan conceived experiments. Nguyen Thi Thanh Sang and Le Thanh Son conducted experiments and analyzed results. Nguyen Van Sinh and Nguyen Thanh Tuan provided critical updates and suggestions that significantly enhanced the scope and direction of the research. Nguyen Van Sinh wrote the paper with important. All authors (Nguyen Van Sinh, Phan Anh Kiet, Nguyen Thanh Tuan, Nguyen Thi Thanh Sang and Le Thanh Son) reviewed and approved the final manuscript.

## REFERENCES


1. Van Sinh N, Dung NA, Lam PQS, 2021 8th NAFOSTED Conference on Information and Computer Science (NICS). Digitalization of administrative documents a digital transformation step in practice.; 2021.
2. Tuan TN, Sang NTT, Kiet PA, Son TL, Sinh NV, accepted in ICCCI . A Hybrid Ensemble Framework for Topic Extraction in Vietnamese Legal Documents; 2025.
3. Siebel TM. Digital Transformation: Survive and Thrive in an Era of Mass Extinction. 2019;.
4. S Z, S S, A U. Digital Transformation in Business. In: and others, editor. International Scientific Conference "Digital Transformation of the Economy: Challenges, Trends, New Opportunities, Lecture Notes in Networks and Systems; 2019. Available from: https://api.semanticscholar.org/CorpusID:201132370.
5. Johan M, Tan R, Suteja B, Afiany N. Document Digitalization and Scoring System of Students Final Project. Jurnal Teknik Informatika Dan Sistem Informasi. 2020;6(3). Available from: https://www.doi.org/10.28932/jutisi.v6i3.3126.
6. Johan M, Tan R, Suteja B, Afiany N. Document digitalization through use of cloud computing technology. International Journal of Engineering Applied Sciences and Technology. 2020;4(10):260–262.
7. Pate C, Patel D. Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study. Interna- national Journal of Computer Applications. 2012;55(10).
8. Vietnam Government. "Format of the administrative document", Number 30/2020/ND-CP, March 23, 2020. The decree No 78/2025/NĐ-CP. Updated on April 01, 2025.
9. Kise K. Page Segmentation Techniques in Document Analysis. Handbook of Document Image Processing and Recognition. 2014;p. 135–75.
10. Kise K. Page Segmentation Techniques in Document Analysis. 2014;Available from: https://www.doi.org/10.1007/978-0-85729-859-1_5.
11. Chung BW. Getting Started with Processing and OpenCV. Pro Processing for Images and Computer Vision with Open. 2017;CV:1–37. Available from: https://www.doi.org/10.1007/978-1-4842-2775-61.
12. Tran MK, Van Sinh N, To NT, Maleszka M. Processing and Visualizing the 3D Models in Digital Heritage. In: and others, editor. 13th In- ternational Conference on Computational Collective Intelligence (ICCCI 2021, Rank B). 13th In- ternational Conference on Computational Collective Intelligence (ICCCI 2021, Rank B). vol. 12876. Springer; 2021. p. 613–625.
13. Van Sinh N, Ha TM, Truong LS, Lecture notes printed computer science 11814. Visualization of Medical Images Data Based on Geometric Modeling. Springer; 2019.
14. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research. 2003;3:993–1022.
15. Adji B. Dieng, Francisco J. R. Ruiz, David M. Blei; Topic Modeling in Embedding Spaces. Transactions of the Association for Computational Linguistics. 2020;8:439–53. Available from: https://www.doi.org/10.1162/tacl_a_00325.
16. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure; 2022. Available from: https://www.doi.org/10.48550/arXiv.2203.05794.
17. Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The Muppets straight out of Law School, Online. Association for Computational Linguistics; 2020.
18. Nguyen, DQ, Tuan Nguyen, A.: PhoBERT: Pre-trained language models for Vietnamese namese. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Compu- tational Linguistics: EMNLP 2020. pp. 1037–1042. Association for Computational Linguistics; 2020.
19. Vu T, Nguyen DQ, Nguyen DQ, Dras M, Johnson M. VnCoreNLP: A Vietnamese natural language processing toolkit. In: Liu Y, Paek T, Patwardhan M, editors. {P}roceedings of the 2018 {C}onference of the {N}orth {A}merican {C}hapter of the {A}ssociation for {C}omputational {L}inguistics: {D}emonstrations. New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 56–60. Available from: https://www.doi.org/10.18653/v1/N18-5012.
20. Tewari, A.: Legalpro-bert: Classification of legal provisions by fine-tuning bert large language model; 2024. Available from: https://arxiv.org/abs/2404.10097.
21. Memon J, Sami M, Khan RA, Uddin M. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). IEEE Access : Practical Innovations, Open Solutions. 2020;8:142642–68. Available from: https://www.doi.org/10.1109/ACCESS.2020.3012542.
22. Smith RW. History of the Tesseract OCR engine: what worked and what didn't. Document Recognition and Retrieval XX. 2013;865802. Available from: https://www.doi.org/10.1117/12.2010051.
23. Zhang R, Yang Y, Wang W; 2020. Available from: https://www.doi.org/10.1051/matecconf/202030902014.
24. Shilaskar S, Iramani H. CTC-CNN-Bidirectional LSTM based Lip Reading System. In: 2024 {I}nternational {C}onference on {E}merging {S}mart {C}omputing and {I}nformatics ({ESCI}); 2024. p. 1–6. Available from: https://www.doi.org/10.1109/ESCI59607.2024.10497275.
25. VietOCR; 2025. Available from: https://sourceforge.net/projects/vietocr/.
26. SodaPDF; 2025. Available from: https://www.sodapdf.com/pdf-tools.
27. ABBYY; 2025. Available from: https://pdf.abbyy.com.
28. Omnipage. Available from: https://www.tungstenautomation.com/products/omnipage.

*Tạp chí Phát triển Khoa học và Công nghệ – Engineering and Technology 2026, 9(1):2702-2714*

Open Access Full Text Article

*Bài nghiên cứu*

# Phương pháp xử lý ngôn ngữ tự nhiên (NLP) để nhận diện và trích xuất các văn bản pháp luật về giáo dục của Việt Nam dựa trên kỹ thuật OCR và XML

**Nguyễn Văn Sinh *, Phan Anh Kiệt , Nguyễn Thanh Tuấn , Nguyễn Thị Thanh Sang , Lê Thanh Sơn**

Use your smartphone to scan this QR code and download this article

**TÓM TẮT**

Chuyển đổi số trong giáo dục đòi hỏi các hệ thống thông minh để trích xuất, chuẩn hóa và phân tích một khối lượng lớn các văn bản pháp luật ở nhiều định dạng khác nhau. Cấu trúc của các văn bản pháp luật Việt Nam trong lĩnh vực giáo dục rất phức tạp, với nhiều thành phần khác nhau, bao gồm Quốc hiệu, Cơ quan ban hành, Ký hiệu văn bản, Ngày ban hành, Điều khoản, Phụ lục, và đôi khi cả Chữ ký hoặc Ghi chú viết tay. Bên cạnh đó, sự khác biệt còn đến từ loại văn bản như hiến pháp, luật, nghị định, thông tư, quyết định, v.v. Ngoài ra, nghị định còn quy định cụ thể về định dạng văn bản (phông chữ, cỡ chữ, lề, khoảng cách dòng), cách bố trí các thành phần và quy tắc trình bày phụ lục hoặc văn bản đính kèm. Do đó, việc xử lý các văn bản này đặt ra nhiều thách thức trong việc truy xuất thông tin và quản lý dữ liệu pháp lý. Xử lý ngôn ngữ tự nhiên (NLP) đóng vai trò quan trọng trong việc xử lý văn bản và tài liệu như vậy. Trong bài báo này, chúng tôi đề xuất một phương pháp mới tích hợp NLP, nhận dạng ký tự quang học (OCR) và xử lý ảnh để xử lý các văn bản pháp luật Việt Nam trong lĩnh vực giáo dục. Nghiên cứu nền tảng này sẽ phục vụ cho một hệ thống truy xuất thông tin tối ưu trong tương lai sử dụng Mô hình ngôn ngữ lớn (LLM) và Trí tuệ nhân tạo tạo sinh (GenAI). Phương pháp của chúng tôi bao gồm các bước sau: (1) thu thập cơ sở dữ liệu văn bản pháp luật ở định dạng PDF từ trang web văn bản pháp luật của Bộ Giáo dục và Đào tạo (MOET); (2) loại bỏ nhiễu, phân loại các thành phần khác nhau và chuyển đổi thành văn bản thuần túy thông qua kỹ thuật OCR và xử lý hình ảnh; (3) chuyển cấu trúc văn bản được trích xuất qua định dạng XML cho hệ thống tương lai của chúng tôi. So với ứng dụng hiện có, phương pháp của chúng tôi đạt được độ chính xác mong muốn trên 99% với các tài liệu in, bao gồm cả khả năng nhận dạng văn bản viết tay. Nghiên cứu này là một bước tiến hướng tới việc phát triển giải pháp kỹ thuật cơ bản cho xử lý dữ liệu, cho phép áp dụng LLM và GenAI trong công cụ tìm kiếm cơ sở dữ liệu được tối ưu để đưa ra quyết định dựa trên văn bản pháp luật.

**Từ khoá:** Trí tuệ nhân tạo, Văn bản pháp lý, Nhận dạng tài liệu, Xử lý ngôn ngữ tự nhiên, Khai phá dữ liệu, Nhận dạng ký tự quang học

*Khoa Công nghệ Thông tin, Trường Đại học Quốc tế, ĐHQG-HCM, Việt Nam*

**Liên hệ**

**Nguyễn Văn Sinh** , Khoa Công nghệ Thông tin, Trường Đại học Quốc tế, ĐHQG-HCM, Việt Nam

Email: nvsinh@hcmiu.edu.vn

**Lịch sử**

- Ngày nhận: 24-07-2025
- Ngày sửa đổi: 16-10-2025
- Ngày chấp nhận: 22-12-2025
- Ngày đăng: 31-12-2025

**DOI :**

**https://doi.org/10.32508/stdjet.v9i1.1536**

Check for updates



VNUHCM PRESS

**Trích dẫn bài báo này:** N V S, P A K, N T T, N T T S, L T S. **Phương pháp xử lý ngôn ngữ tự nhiên (NLP) để nhận diện và trích xuất các văn bản pháp luật về giáo dục của Việt Nam dựa trên kỹ thuật OCR và XML**. *Sci. Tech. Dev. J. - Eng. Tech.* 2026; 9(1):2702-2714.