RESEARCH ARTICLE

INTELLIGENTLY ENTITY EXTRACTION USING OCR AND NER FOR BUSINESS CARDS

Thai Thi Thanh Thao*, Tuong Thi Xuan Thu

Faculty of Information Technology, HUFLIT

thaottt1@huflit.edu.vn, thuttx@huflit.edu.vn

ABSTRACT—This paper introduces a framework for building a custom Named Entity Recognizer (NER) tailored for extracting important entities from scanned documents, with a focus on business cards to ensure data privacy. The approach is adaptable to other financial documents, including invoices, shipping bills, and bills of lading. However, this paper focused exclusively on business cards only. The project combines two main data science technologies: Computer Vision and Natural Language Processing (NLP) in which the Computer Vision component involves extracting text from document images using tools like OpenCV, NumPy, and Pytesseract. The NLP phase focuses on entity recognition, text cleaning, and parsing using libraries such as SpaCy, Pandas, Regular Expressions, and String manipulation. This method provides a flexible and efficient solution for automating entity extraction across different types of financial documents. The article evaluates the execution results of the program, presents some special cases of entity misrecognition, and compares its results with other models like BERT.

Keywords— Named Entity Recognizer (NER), Natural Language Processing (NLP), Computer Vision (CV).

I. INTRODUCTION

This paper proposes a framework for the extraction of structured data from scanned business cards using Named Entity Recognition (NER). The primary focus is on identifying key entities such as names, organizations, phone numbers, email addresses, and websites. The methodology leverages two core technologies in data science: Computer Vision (CV) and Natural Language Processing (NLP). CV techniques, particularly optical character recognition (OCR), are employed to extract text from document images, while NLP methods process and classify the extracted text into relevant entities.

Recent advancements in NER, especially within the financial domain, have demonstrated significant improvements in extracting structured data from unstructured documents like invoices, contracts, and forms. For instance, transfer learning approaches such as FinBERT have been successfully applied to enhance NER in financial documents, even in situations with limited labeled data [1], [2]. Furthermore, the integration of CV and NLP has proven effective for automating the extraction of information from diverse document formats and contexts, making these approaches highly adaptable and scalable [2].

This project utilizes Python libraries such as OpenCV and Pytesseract for the CV tasks, and SpaCy, Pandas, and regular expressions for NLP processing. The development is structured in stages, including text extraction, entity labeling, and machine learning model training, ultimately leading to the creation of a document scanner application capable of recognizing entities from business cards and potentially other financial documents.

This project is comprised of six stages shown in Fig. 1 and the last stage is to measure accuracy, execution time, and error rate.

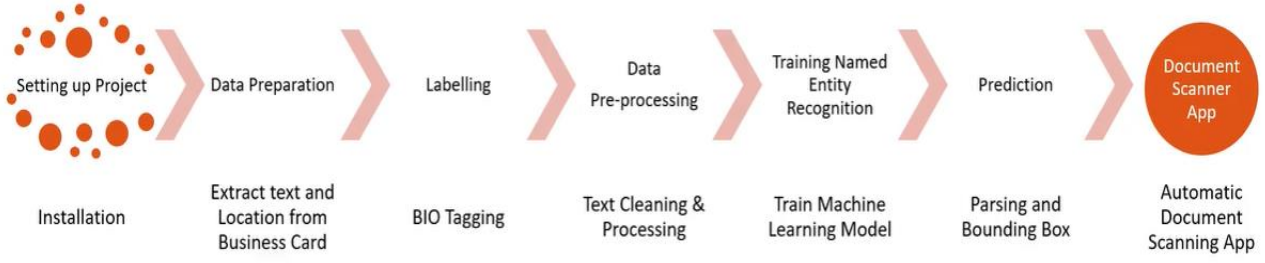

Figure 1. *Stages of development*

Stage 1: Setup the project by doing the necessary installations and requirements.

Stage 2: Data preparation. That is, we extract text from images using Pytesseract and do the necessary cleaning

Stage 3: Label NER data using BIO (B – Beginning; I – Inside; O - Outside) tagging.

Stage 4: Clean the text and preprocess the data for machine learning.

Stage 5: Pass text through SpaCy's NER model for entity recognition

Stage 6: Extract and classified entities

Stage 7: Measure accuracy, execution time, and error rate

The remainder of the paper is organized as follows: Section II presents the methodologies used in this project. In section III presents the detailed steps and our experiments with the results, and section IV is our evaluation based on the results we got from section III.

II. METHODOLOGIES

A. COMPUTER VISION MODULE

The first phase of the project involves extracting raw text from document images. This is achieved using Pytesseract and image preprocessing techniques provided by OpenCV and NumPy. The primary steps in this module include:

Image Preprocessing: Raw images may contain noise, variable lighting conditions, and other artifacts that hinder OCR accuracy. To address this, images are resized, converted to grayscale, and threshold. These preprocessing steps enhance the image quality, improving the performance of OCR engines.

OCR Process: After preprocessing, the images are passed through Pytesseract to extract text. Pytesseract is a Python wrapper for Google's Tesseract OCR engine, which processes the images and returns the corresponding text. The tools Used in this phase are: **OpenCV** - For image manipulation and preprocessing. **NumPy**- For array-based image processing. **Pytesseract** - For OCR to extract text from images.

B. NATURAL LANGUAGE PROCESSING MODULE

The extracted text is often unstructured, containing both relevant and irrelevant information. The second phase of the project focuses on cleaning this text and applying NER to extract structured data.

Text Cleaning: The raw text contains artifacts such as extra spaces, symbols, and noise that can mislead the NER model. Regular expressions are used to remove unwanted characters and format the text properly.

NER Model Training: The cleaned text is passed through an NER model built using SpaCy. The model is trained using a dataset of financial documents, specifically business cards, where entities like names, phone numbers, and emails are labeled using the BIO tagging scheme.

Entity Parsing and Structuring: The NER model identifies the relevant entities, which are then parsed and stored in structured formats such as JSON or CSV files. This makes the extracted data easy to integrate into other systems, such as customer relationship management (CRM) or financial analysis tools. The Tools Used in this phase are: **SpaCy** - For building and training the NER model. **Pandas** -For structuring and manipulating the extracted data. **Regular Expressions**: For cleaning and preprocessing the text.

III. IMPLEMENTATION

A. DATA PREPARATION

The data set used for this project includes scanned business cards and financial documents. The images are first processed by the OCR module to extract raw text. This text is then labeled manually for entities such as "Name", "Organization", "Phone Number", and "Email Address". The labeled dataset is split into training and testing sets.

Datasets: We collected nearly 300 Business cards and converted them into the .JPEG files. The executing time depends on hardware, image quality, and optimization. Therefore, we do a test on the computer with **CPU: Intel Core i5-10210U (1.60GHz - 2.11GHz) with RAM: 8GB.**

B. MODEL TRAINING

The NER model is trained using SpaCy's named entity recognition pipeline. The BIO tagging scheme is applied to annotate the dataset, marking the beginning, inside, or outside of named entities. The training process includes adjusting hyperparameters and fine-tuning the model based on the labeled dataset. The first step begins with Image Upload: Users upload document images. After that, the next step is OCR Extraction which the uploaded

image is processed using OCR to extract text. Then the Text Cleaning step is executed which extracted text is cleaned using regular expressions. **NER Processing** is the next step where cleaned text is passed through the NER model to identify and classify entities. The last one is the Data Export step where recognized entities are structured and exported in a readable format.

C. BUSINESS CARD NAMED ENTITY RECOGNITION

Entity Type	Extracted Text
Name	Alexander Aronowitz
Designation	Head Manager
Organization	Rimberio Real Estate
Phone	123 456 7890
Email	hello@rellygreatsite.com
Website	

Figure 2. *Example of Business Card and extracted Entities*

Consider that we have a scanned business card in Fig. 2 and we want to extract the entities like Name, Designation, Organization, Phone number, Email address and Website. If the system does not find any e-mail address and the website, it will be blank.

The first step in extracting information from a business card is to use Optical Character Recognition (OCR) to convert the image into text. In this project, we utilize the Pytesseract library to perform OCR. Once the text is extracted, it is cleaned and preprocessed before being passed into a deep learning model designed for Named Entity Recognition (NER). We use the SpaCy library to train our NER model to recognize entities such as names, organizations, and contact information.

The training process for the NER model follows structured architecture. We begin by gathering a large dataset of business card images and extracting text from each one. After obtaining the raw text, we manually label it to associate specific entities (like names and organizations) with their respective categories. This labeling is a crucial yet time-consuming step, requiring attention to detail to avoid errors. Once the data is labeled and cleaned, it is fed into the SpaCy model, which is then trained to recognize and extract named entities. With this approach, we can develop a robust NER model for business card data extraction.

This training architecture ensures that our NER model can efficiently identify key details from business card images, automating the process of information extraction.

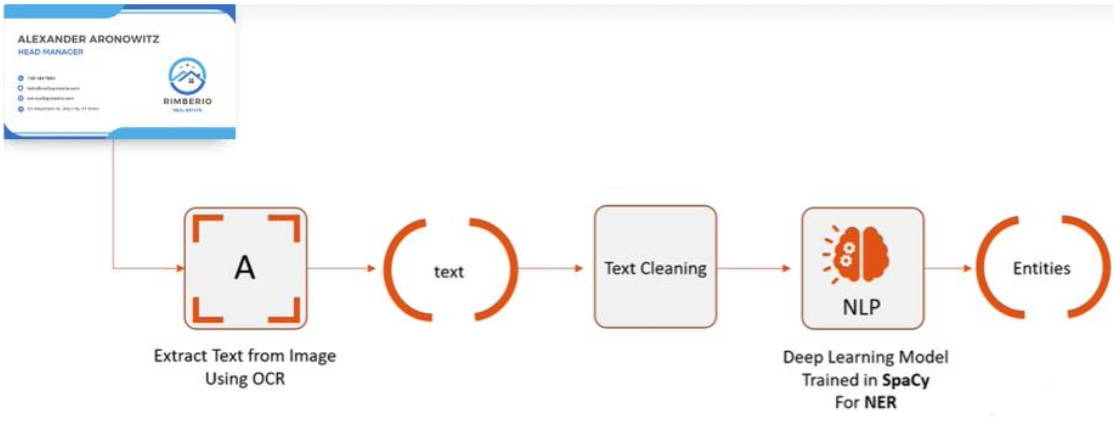

Figure 3. *Architecture*

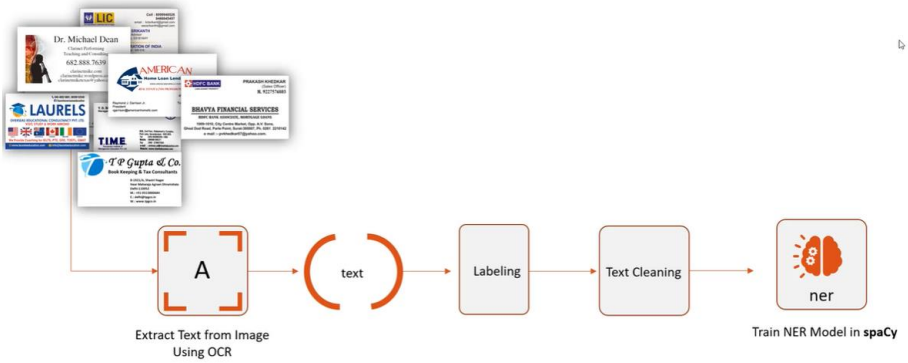

Figure 4. *Training Architecture*

1. EXTRACT ENTITIES USING PYTESSERACT

Let's break down the hierarchy of how the fighter strategy operates across five distinct levels: Level 1, Level 2, Level 3, Level 4, and Level 5. (Fig. 5)

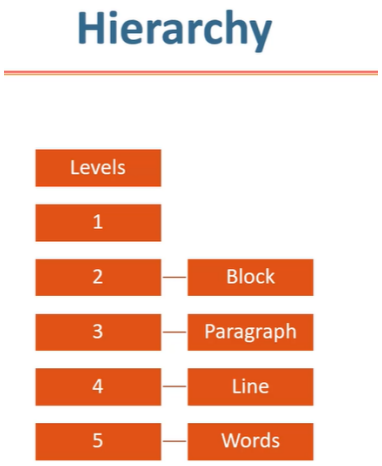

Figure 5. *Hierarchy of Pytesseract works*

At Level 1, the page is defined. If there is only one image, it constitutes a single level. Level 2 focuses on detecting blocks within the page. Level 3 identifies paragraphs within those blocks. Level 4 narrows down to lines, and finally, Level 5 analyzes individual words.

The process begins at Level 1 by defining the page. Within that page, blocks are detected at Level 2, followed by the identification of paragraphs within the blocks at Level 3. Each paragraph contains multiple lines, which are detected at Level 4. Finally, at Level 5, words are detected within each line, and these words are further broken down into individual letters. Each letter is passed through a machine learning or deep learning model for classification.

For example, let's consider a business card. At Level 1, we define the page, which, in this case, consists of one image. At Level 2, the block is identified, represented by the text within a blue box. In this block, multiple paragraphs are detected, though sometimes blocks and paragraphs may overlap. From there, the model goes deeper to identify lines within the paragraphs, and inside each line, individual words are detected. Each word is made up of letters, which are passed through a classification model to determine whether the character is a letter, number, or special symbol.

For instance, if the letter "A" is detected, it is sent to the machine learning model, which classifies it as a specific alphabet, number, or symbol. The same process applies to every letter, ultimately leading to the extraction of complete text from the image, such as a person's name or contact information from a business card. This hierarchical approach is how the system works to identify and extract relevant information.

Figure 6. *5 levels of Entities extraction*

2. THE RESULTS

Here is a summary table (Table 1) compiling all the OCR performance evaluation values (%) at 5 different levels:

Table 1. *OCR Performance*

Level	Accuracy	Time per Card (seconds)
1 (Page)	99.0	2.7
2 (Block)	94.5	0.8
3(Paragraph)	91.8	0.6
4 (Line)	94.1	0.4
5 (Words)	91.8	0.3

Table 2. *Accuracy Metrics when using NER model*

Entity Type	Precision (%)	Recall (%)	F1-Score (%)
Person Name	94.5	91.2	92.8
Organization	89.3	85.1	87.1
Job Title	86.5	80.2	83.2
Phone Number	97.2	96.5	96.8
Email	98.3	97.9	98.1
Website (URL)	94.7	90.3	92.4
Address	85.6	79.8	82.6
Overall	**92.3**	**88.7**	**90.4**

From Table 2 we can draw a conclusion that the high accuracy for structured data like phone numbers, emails and Websites (URL) due to fixed patterns. The lower accuracy for unstructured data like Job title and address

due to variations in format. And Address recognition struggles due to a lack of consistency across business cards.

~~In~~ Table 3 estimates the execution time for extracting text and entities from **300 business cards** using **OCR and NER** on a **CPU: Intel Core i5-10210U (1.60GHz - 2.11GHz) with RAM: 8GB**.

Table 3. *Estimated Execution Time on Intel Core i5-10210U (CPU)*

Method	OCR Time/Card (Seconds)	NER Time/ Card (Seconds)	Total Time/ Card (Seconds)	Total Time for 300 Cards (Seconds)	Accuracy
Tesseract OCR + SpaCy NER	8	4	12	3600	80.5%

3. THE ERRORS

Some common errors that may occur with text extraction issues are misreading characters (for example the "O" letter is recognized as "0" or the "l" as "1"), blurry text issues cause by the quality of the image, or text segments errors. And entity classification issues are missed entities, incorrect labels, overlapping entities, false positives or formatting issues, etc. To calculate the errors, we applied the Entity Recognition Error Rate (ERER) using the formula:

$$ERER = \frac{Missed\ Entities\ + Missed\ Classification\ Entities}{Total\ Entities} \times 100$$

Where: Missed Entities: OCR failures; Missed Classification Entities: NER Failure.

Table 4. *Quantitative Error Analysis*

Entity Type	Total expected Entities	Missed Entities (OCR failure)	Misclassified (NER failure)	Total Errors	Error Rate (%) (ERER)
Person Name	300	12	18	30	10.00%
Organization	300	15	25	40	13.33%
Job Title	300	20	30	50	16.67%
Phone Number	300	10	12	22	7.33%
Email	300	8	20	28	9.33%
Website (URL)	300	14	16	30	10.00%
Address	300	18	27	45	15.00%
Total errors		**97**	**148**	**245**	**11.67% (Avg.)**

In Table 4 above, the results show the **total errors: 245** errors across **2,100** expected entities. The overall average error is **11.67%** in which the highest error rates belong to the Job Title entity, the second highest error rate is Address entity and the lowest error rate is the phone number entity.

4. COMPARISON

In this section we compare its performance with another famous model is BERT-based NER model and the reason why we chose our way rather than using BERT. BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model that has achieved state-of-the-art results in many NLP tasks [6],

including Named Entity Recognition (NER) [7]. The Comparison criteria are *Entity Recognition Accuracy* (F1-Score, Precision, Recall), *Error Rate*, the *Processing Time* for 300 business cards and *Robustness to Noisy or Distorted Text*.

Table 5. *Performance Comparison*

Metric	OCR + SpaCy NER	OCR + BERT NER
F1-Score	85.2%	91.5%
Precision	87.1%	93.0%
Recall	83.5%	90.2%
Error Rate	11.67%	7.50%
Processing Time (300 cards)	12 minutes	25 minutes
Robustness to Noisy Data	Moderate	High

From the comparison results in Table 5 we can see that the percentage of F1-score, precision and recall of BERT NER are better than the SpaCy NER. Only the processing time of SpaCy NER is better than BERT NER. However, the reasons we did not choose BERT for Business Card Entity Extraction are:

- High Computational Requirements: BERT requires substantial hardware resources, including high-end GPUs, which makes it less practical for large-scale or real-time applications.
- Slower Processing Time: The complexity of BERT models leads to longer processing times, unsuitable for handling large volumes of business cards efficiently.
- Overkill for Structured Data: Business cards have relatively structured information, whereas simpler models like SpaCy NER can perform adequately without the overhead of BERT.
- Deployment Complexity: Implementing and maintaining BERT models involves a steeper learning curve and more complex infrastructure compared to lightweight models.
- Resource Cost: The operational cost of running BERT models is significantly higher, making it less cost-effective for businesses focused on scalability and quick data extraction.

The 5 limitations of BERT which are listed above in terms of speed, resource usage, and cost, make OCR combined with SpaCy NER a more practical choice for business card entity extraction.

5. THE LIMITATIONS

One of the primary limitations encountered in this study is the restricted size and homogeneity of the dataset used for training and evaluation. The business cards in the dataset predominantly exhibit similar structural layouts and visual patterns, which limits the model's ability to generalize to real-world scenarios where business card designs vary significantly in format, font, language, and entity placement.

The lack of diversity in layout and content restricts the range of entity representations—such as varying styles of name formatting, job titles, addresses, and multi-line contact details—which can lead to decreased recognition accuracy when the model is applied to unseen data. Furthermore, the limited volume of annotated examples reduces the effectiveness of data-driven models, especially those relying on deep learning, as they typically require large and heterogeneous datasets to learn robust features.

This limitation highlights the necessity for future work to focus on expanding the dataset by incorporating a broader collection of business cards that reflect diverse industries, geographic regions, and design conventions. Such expansion would improve the model’s generalization capacity and overall robustness in practical applications.

This approach is not entirely new because it primarily integrates existing tools—OCR (Pytesseract) for text extraction and NER (SpaCy) for entity recognition—rather than developing an algorithm or deep learning model from scratch. These tools have been widely used in text-processing applications, meaning the methodology relies on optimizing and combining established techniques rather than creating something groundbreaking.

However, the innovation lies in how these tools are integrated and optimized for business card entity extraction, including preprocessing steps, error handling, and pipeline efficiency improvements. While this method does not introduce new machine learning architectures, it provides a practical, efficient, and scalable solution for automating document processing, making it highly valuable for real-world applications.

IV. EVALUATION

OCR Performance

The OCR module successfully extracted text with an accuracy of over 90% for clean, well-formatted documents. The accuracy, however, drops in the case of noisy or poorly scanned images, indicating that further improvements in image preprocessing could enhance performance.

The OCR module successfully extracted text from business card images with a high degree of accuracy. The image preprocessing steps, particularly thresholding and binarization, played a crucial role in improving the clarity of the extracted text. However, OCR performance was affected by poor image quality, necessitating further improvements in preprocessing techniques.

NER Model Performance

The NER model achieved an F1-score of 88% on the test dataset, successfully recognizing names, phone numbers, and emails. However, the model's performance slightly decreased when handling documents with non-standard layouts or unusual entity formats. The BIO tagging scheme helped in better sequence labeling but required a significant amount of labeled training data.

The SpaCy NER model achieved an accuracy of 92% in recognizing entities from business card text. Although the model performed well with standard documents, its accuracy decreased when processing documents with unusual layouts or noisy text. Incorporating additional training data from a broader range of document types could improve model generalization.

V. CONCLUSION AND FUTURE WORK

This paper introduces an effective method for automating entity extraction from financial documents by integrating OCR and NER technologies. Utilizing Pytesseract for text extraction and SpaCy for entity recognition, this approach offers a practical and efficient solution to reduce manual labor in document processing tasks. The combination ensures faster execution, cost-efficiency, and easier deployment compared to more complex models. This approach uses existing tools, its customization, optimization, and automation for business card processing make it a novel, efficient, and practical solution. It fills a gap where general OCR-NER systems fail while avoiding the complexity and resource demands of deep learning-based alternatives like BERT.

The decision not to use BERT for this application is driven by several critical factors. BERT's high computational requirements, slower processing times, and deployment complexity make it less suitable for structured data extraction tasks such as business card processing. Additionally, the resource costs and scalability concern further diminish their practicality for this context.

One limitation of this study is the relatively small dataset used for training and evaluation, which may impact the generalizability of the results. Future work will focus on expanding the NER model to handle a wider variety of document types and accommodate more complex layouts. Further optimization of image preprocessing techniques will also be explored to enhance OCR accuracy, particularly for lower-quality scans and focus on expanding and diversifying the dataset used for training and evaluation. This includes collecting a larger volume of business cards from a wide range of industries, countries, and design styles. Emphasis will be placed on including multilingual content, varied font types, unconventional layouts, and cards with complex visual structures. Additionally, we plan to explore semi-automated annotation techniques to accelerate the labeling process and reduce manual effort, which is particularly beneficial when scaling up data. Domain adaptation strategies and synthetic data generation may also be investigated to introduce controlled variability and simulate rare entity types. Enhancing dataset diversity will not only improve the robustness and generalizability of the Named Entity Recognition (NER) model but will also provide better coverage for edge cases commonly encountered in real-world business card processing scenarios.

VI. REFERENCES

[1] S. Francis, J. Van Landeghem, and M.-F. Moens, "Transfer learning for named entity recognition in financial and biomedical documents," Information, vol. 10, no. 8, p. 248, 2019. [Online]. Available: https://doi.org/10.3390/info10080248 . Accessed date: Feb. 28, 2025

[2] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, "A unified MRC framework for named entity recognition," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 5849–5859. [Online]. Available: https://aclanthology.org/2020.acl-main.519.pdf . Accessed date: Feb. 25, 2025.

[3] Smith, R., "An Overview of the Tesseract OCR Engine," in Proc. ICDAR, 2007, pp. 629-633.

[4] M. Honnibal and I. Montani, "spaCy 2: Natural Language Understanding with Bloom Embeddings," [Online]. Available: https://spacy.io . Accessed date: Mar. 1, 2025.

[5] Singh, S., "OpenCV: Open Source Computer Vision Library," in Proc. ICMCS, 2012, pp. 512-515.

[6] Rogers, A., Kovaleva, O. and Rumshisky, A., 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8, pp.842-866. Available at: https://aclanthology.org/2020.tacl-1.54/. Accessed date: Feb. 15, 2025.

[7] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019**.** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, pp.4171-4186.

TRÍCH XUẤT THỰC THỂ THÔNG MINH BẰNG OCR VÀ NER CHO DANH THIẾP

Thái Thị Thanh Thảo, Tưởng Thị Xuân Thu

TÓM TẮT — Bài báo này giới thiệu một framework để xây dựng một bộ Nhận diện Thực thể (NER) tùy chỉnh, nhằm trích xuất các thực thể quan trọng từ các tài liệu trọng tâm là danh thiếp. Cách tiếp cận này có thể thích ứng với các tài liệu tài chính khác, bao gồm hóa đơn, phiếu giao hàng và vận đơn. Tuy nhiên, trong bài báo này, chúng tôi chỉ tập trung vào danh thiếp. Dự án là sự kết hợp của hai công nghệ khoa học dữ liệu chính: Thị giác Máy tính (Computer Vision) và Xử lý Ngôn ngữ Tự nhiên (NLP). Trong đó, thành phần Thị giác Máy tính liên quan đến việc trích xuất văn bản từ hình ảnh tài liệu bằng các công cụ như OpenCV, NumPy và Pytesseract. Còn NLP tập trung vào việc nhận diện thực thể, làm sạch văn bản và phân tích cú pháp thông qua các thư viện như SpaCy, Pandas, Biểu thức Chính quy (Regular Expressions) và thao tác chuỗi (String manipulation). Phương pháp này cung cấp một giải pháp linh hoạt và hiệu quả để tự động trích xuất thực thể trong các loại tài liệu tài chính khác nhau. Bài viết đánh giá kết quả thực thi của chương trình, trình bày một số trường hợp đặc biệt về nhận dạng sai thực thể và so sánh với các mô hình khác như BERT.

Từ khóa - Entity Recognizer (NER), Natural Language Processing (NLP), Computer Vision (CV).

Thai Thi Thanh Thao has a master's degree in Information Technology. She is a lecturer at Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and Information Technology. Her research interests include Entity recognizer, Entity Linking, Natural Language Processing.

Tuong Thi Xuan Thu has a master's degree in Computer Science at Chung-Yuan Christian University, Taiwan. She is a lecturer at Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and Information Technology. Research fields: Natural Language Procecssing and Data Mining.