electronics

MDPI

Article

A Framework for Understanding Unstructured Financial Documents Using RPA and Multimodal Approach

Seongkuk Cho [1,†], Jihoon Moon [2,†], Junhyeok Bae [1], Jiwon Kang [1] and Sangwook Lee [1,*]

[1] AI Unit, Shinhan Bank, Seoul 01056, Republic of Korea
[2] Department of AI and Big Data, Soonchunhyang University, Asan 31538, Republic of Korea
* Correspondence: swleee@shinhan.com
† These authors contributed equally to this work.

Abstract: The financial business process worldwide suffers from huge dependencies upon labor and written documents, thus making it tedious and time-consuming. In order to solve this problem, traditional robotic process automation (RPA) has recently been developed into a hyper-automation solution by combining computer vision (CV) and natural language processing (NLP) methods. These solutions are capable of image analysis, such as key information extraction and document classification. However, they could improve on text-rich document images and require much training data for processing multilingual documents. This study proposes a multimodal approach-based intelligent document processing framework that combines a pre-trained deep learning model with traditional RPA used in banks to automate business processes from real-world financial document images. The proposed framework can perform classification and key information extraction on a small amount of training data and analyze multilingual documents. In order to evaluate the effectiveness of the proposed framework, extensive experiments were conducted using Korean financial document images. The experimental results show the superiority of the multimodal approach for understanding financial documents and demonstrate that adequate labeling can improve performance by up to about 15%.

Keywords: intelligent document processing; visual-rich document understanding; optical character recognition; financial document analysis; key information extraction; image classification; RPA

Citation: Cho, S.; Moon, J.; Bae, J.; Kang, J.; Lee, S. A Framework for Understanding Unstructured Financial Documents Using RPA and Multimodal Approach. *Electronics* **2023**, *12*, 939. https://doi.org/10.3390/electronics12040939

Academic Editor: Taeshik Shon

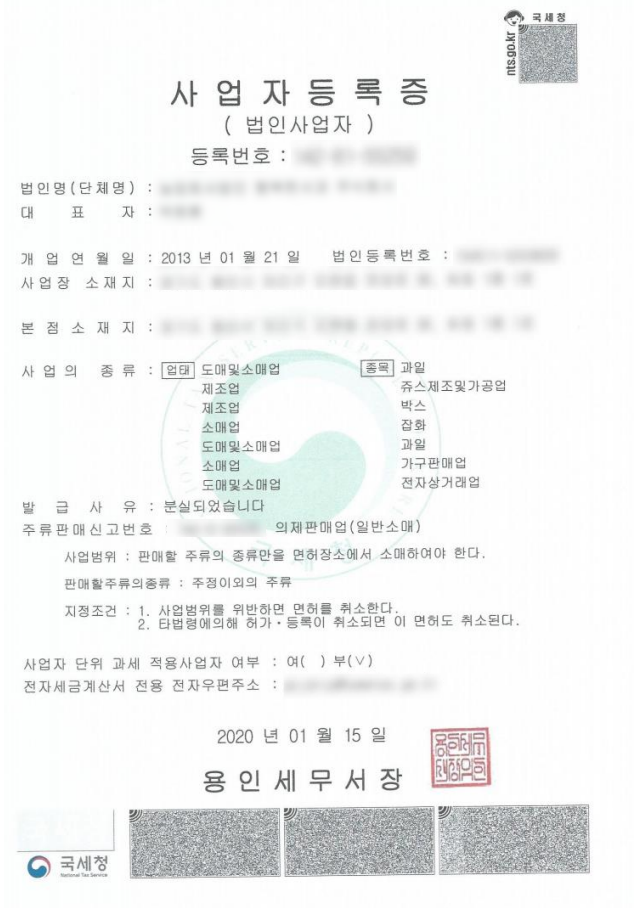
Received: 5 January 2023
Revised: 8 February 2023
Accepted: 10 February 2023
Published: 13 February 2023

1. Introduction

The financial industry is trying to automate business processes that are managed to attain better performance and efficiency [1–4]. For example, many banks have introduced intelligent document processing (IDP) frameworks to automate the process of manually recording business documents into back-office systems using robotic process automation (RPA) and artificial intelligence (AI) [5–7]. It was reported in a recent study that by automating only 29% of functions for a task using RPA, finance departments save more than 25,000 h of rework caused by human errors costing $878,000 per year for an organization with 40 full-time accounting staff [8]. The RPA is a system that communicates with other systems to capture and trigger data and can perform several repetitive tasks. Usually, RPA is used to capture an area containing information to be extracted within a structured document image, and optical character recognition (OCR) is applied to that area. After introducing these frameworks, business processing speed has improved, and scalability and flexibility have been increased through rapid response [9]. However, most banking occurs through the exchange of unstructured documents, such as salary transfer statements, trade transaction confirmation, and withholding tax receipts.

Figure 1 shows examples of unstructured financial documents. In Figure 1, all areas containing personal information have been obscured to effectively protect individuals' privacy through de-identification [10]. The information is presented in plain text but

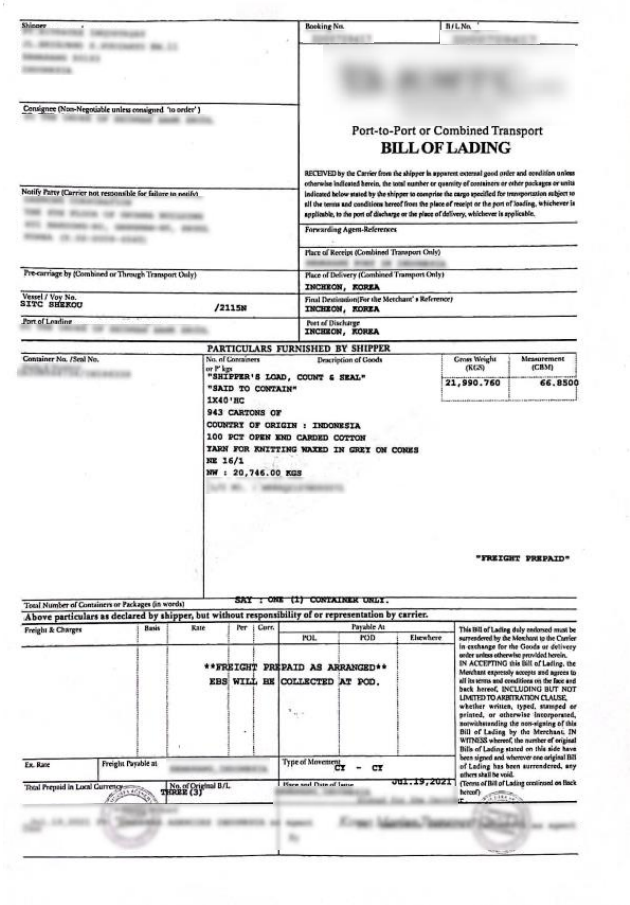
consists of multi-column layouts and a wide variety of tables/forms/figures/language. Additionally, financial documents are available multilingually, including in English and native languages. In these situations, relying only on RPA must be hard-coded and updated with changes to each format, which is costly. Figure 2 shows the typical workflow of banks in processing financial documents. In order to automate this workflow, it is essential to understand the document image. In summary, understanding unstructured and complex document images have considerable potential to reduce repetitive manual tasks and streamline business communication. With the development of AI technology recently, traditional RPA has developed into a hyper-automation solution by combining computer vision (CV) and natural language processing (NLP) methods [11,12]. Among these, key information extraction (KIE) is used to automatically extract essential information that must be entered into the back-office system. For example, invoice images have several critical information such as "total price," "amount of tax," and "amount of change." Most of these approaches consist of (1) conversion into a text blob using an OCR engine and then (2) applying NLP techniques such as named entity recognition (NER) [13].



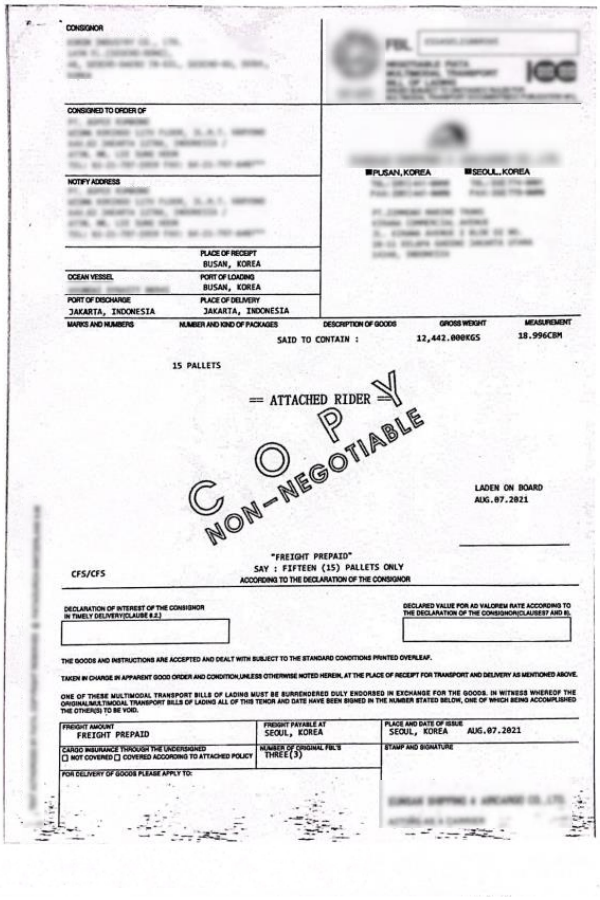
(**a**) Korean Business Registration #1

(**b**) Korean Business Registration #2

Figure 1. *Cont.*



(**c**) Bill of Loading #1



(**d**) Bill of Loading #2

Figure 1. Examples of unstructured financial documents.

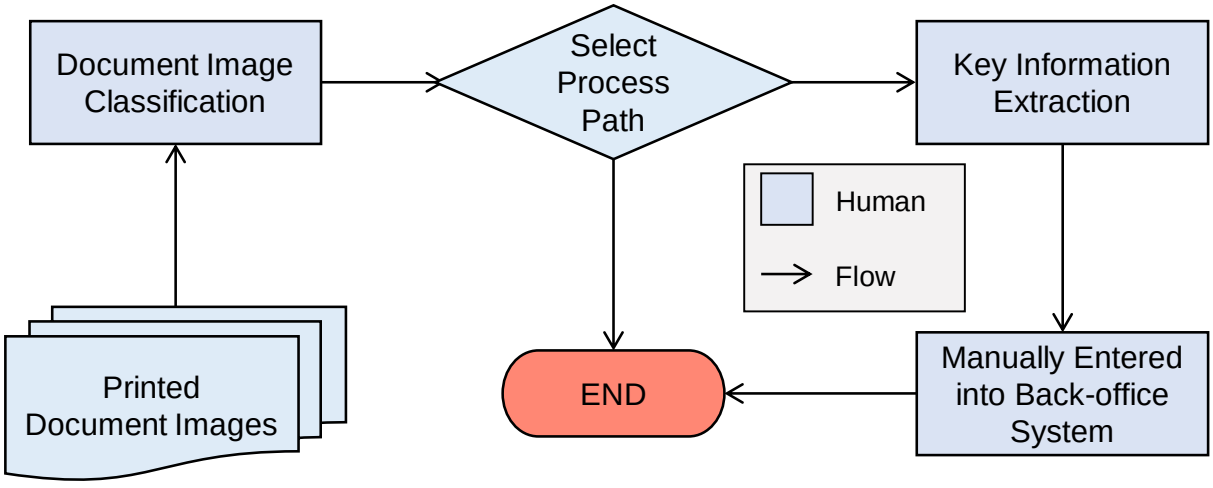


Figure 2. Original business process.

Currently, despite the KIE task's diverse and complex structure, it is possible to perform the KIE tasks with only a small amount of label data using a pre-trained deep learning model for visual document understanding [14]. These systems are capable of image analysis, such as KIE and document classification, but have some limitations in applying them to real-world financial documents. Additionally, they perform poorly on text-rich document images with a small percentage of entities extracted because they focus

only on language understanding. Also, multilingual document image analysis requires many training data. The problem addressed in this study is the difficulty in automating business processes in the financial industry that are handled manually, particularly the process of recording business documents into back-office systems. Neural architecture search methods with a lightweight transformer for text-to-image synthesis have been used to tackle these limitations [15,16]. In addition, RPA and AI have been introduced to automate tasks. However, these approaches have limitations when applied to unstructured financial documents, such as those containing multi-column layouts and a wide variety of tables, forms, figures, and languages.

In order to solve these problems, this paper presents an IDP framework that combines a pre-trained deep learning model with traditional RPA used in banks to extract key information from real-world Korean financial document images. It not only enables the detection of key information and their relationships but also the classification of types of document images. Furthermore, most financial documents at non-English banks are written in various languages. We use a multilingual model to accommodate these needs, pre-trained with large-scale real-world document images or natural language corpus. To the best of our knowledge, our framework is the first utilization of a multilingual, pre-trained model for visually rich document understanding of traditional bank systems such as databases and RPA. In addition, the main contributions of this study are summarized as follows:

1. We propose a new document image processing framework that combines a pre-trained deep learning model with traditional RPA for Knowledge Information Extraction (KIE) from real-world Korean financial document images.
2. We use RPA to take document images through a database, encode them, and compare the performance of our proposed method with various state-of-the-art language models such as XLM-RoBERTa and InfoXLM.
3. We conduct an ablation study which shows that labeling both "key-value" pairs increases the accuracy in all comparative experiments and is up to 10% better than labeling only values.
4. We compare the performance of KIE using four models: mBERT, XLM-RoBERTa, InfoXLM, and LayoutXLM, focusing on the effectiveness of assigning labels to "key-value" pairs in document understanding.
5. We also show a reduction in business processing time by over 30% in specific financial tasks.

This paper is organized as follows: Section 2 outlines the related work, Section 3 presents the overall structure of our system, Section 4 describes the experiment settings and results, and Section 5 concludes the paper and provides directions for future work.

2. Related Work

In this section, we discuss the significant studies reported on applying AI to the system of document image processing. In particular, pre-trained models for understanding multilingual document images have been applied in related systems. Therefore, we discuss two groups: the intelligent document processing (IDP) systems and pretraining techniques for multilingual document understanding.

2.1. Automation of Business Processes

With the advances in AI, the digitization that moves data from analog to digital has been studied in various fields, such as insurance, banking, healthcare, and manufacturing [14,17–19]. In particular, various attempts using deep learning or machine learning have been made for digitization. In most of these attempts, classification mechanisms or NLP have been employed for digitization. For instance, Baidya [17] proposed an automated business process with a support vector machine classifier and an RPA to classify unstructured documents. Roopesh et al. [18] proposed an intelligent system for automating interview processes using RPA and AI to help recruitment. The system first monitors mail

and attachment using RPA, then uses two deep learning models to classify resumes and extract essential information. Deep learning models take embedded text information extracted by OCR as input, classify it using bidirectional long short-term memory (LSTM) and extract essential information using NER based on the LSTM-CRF model. Guha et al. [19] proposed a multimodal binary classification approach based on state-of-the-art transfer learning techniques involving images and NLP models to classify insurance documents. Image and textual features were extracted simultaneously from the document image using VGG-16 and the pre-trained Bidirectional Encoder Representations from Transformer (BERT) model, respectively, through fine-tuning. Mungmeeprued et al. [20] proposed a deep learning solution to determine whether a page is a breaking point given a sequence of visually rich document images as input. This solution combines LayoutLM and residual network (ResNet) [21] to exploit the document pages' textual and visual features and achieve state-of-the-art (SOTA) in the task.

2.2. Multilingual Pre-Trained Models

Multilingual pre-trained models [22–25] are pretraining of the transformers with a corpus composed of various languages and have been applied to various businesses to achieve many digital transformations. Devlin et al. [22] introduced a self-language understanding model called BERT and extended it to a multilingual version. The multilingual BERT (mBERT) was pre-trained on the top 104 languages with the most extensive Wikipedia corpus using a masked language modeling algorithm. Lample et al. [23] proposed two methods for extending pre-trained models with English corpus to multiple languages to train a cross-language model (XLM). One is an unsupervised method that only relies on monolingual data, and the other is a supervised method that leverages parallel data as a new cross-language model goal. Conneau et al. [24] proposed a multilingual modeling method without sacrificing per-language performance. Liu et al. [25] proposed a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages called the bidirectional auto-regressive transformer.

The main difference between our study and related work is that our study presents an IDP framework that combines a pre-trained deep learning model with traditional RPA used in banks to extract key information from real-world Korean financial document images. In contrast, related studies focus on digitizing various fields such as insurance, banking, healthcare, and manufacturing using deep learning or machine learning techniques like classification mechanisms or NLP. Additionally, our study utilizes a multilingual model to accommodate the needs of financial documents written in various languages. In contrast, related studies have proposed solutions for multilingual pre-trained models using English corpus to multiple languages to train a cross-language model. Additionally, our study proposes an ablation study to show the benefit of labeling both "key-value" pairs in document understanding and also demonstrates a reduction in business processing speed by over 30% in specific financial tasks.

3. Proposed Framework

In this section, the overall architecture of our proposed framework for financial document image understanding is introduced, which is illustrated in Figure 3. Document image understanding includes classification, key information/relation extraction, and layout analysis. We consider a transformer-based model for this study because transformer-based models use a self-attention mechanism which allows the model to weigh different parts of the input differently and focus on the essential parts for a given task; it also allows the model to handle variable-length input better and generalize better to new data. Transformer-based models are pre-trained on large amounts of data and fine-tuned on specific tasks, which allows them to perform well even with limited task-specific training data. In addition, they are trained to perform multiple tasks at once, allowing them to learn a more general representation of the input data and transfer knowledge across tasks.

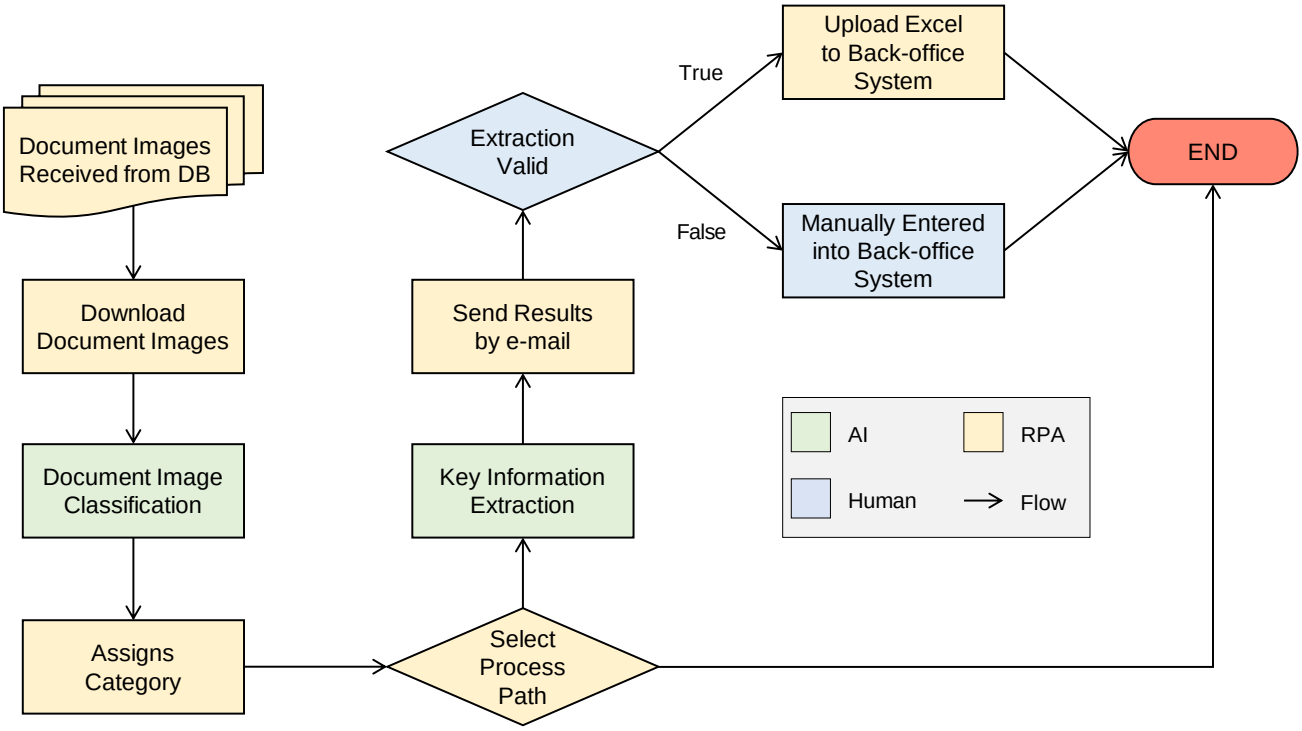


Figure 3. Overall architecture of the proposed framework for unstructured financial document image processing.

In this step, we describe how to perform two specific tasks (i.e., document classification and KIE) on unstructured multilingual document images using LayoutXLM, the SOTA transformer-based model. LayoutXLM is a multilingual version of LayoutLMv2 that is trained with 11 million multilingual document images from the IIT-CDIP dataset. It uses four types of embeddings for pretraining: text, visual, one-dimensional (1D) positional, and two-dimensional (2D) positional. LayoutXLM uses three strategies in the pretraining step: Masked visual-language modeling, Text-image alignment, and Text-image matching. This study fine-tunes the pre-trained LayoutXLM for two financial document image understanding tasks, including document image classification and KIE. For the KIE task, we fine-tune LayoutLM to predict {B, O} tags for each token and detect each entity type in the dataset using sequential labeling. For the document image classification task, we use the [CLS] token to predict the class labels using the representation of the token.

3.1. Overall Proposal

In this step, we explain the overall flow chart of a framework that combines AI with an automated click bot called RPA. Our proposal consists of three components: (i) RPA, (ii) AI, and (iii) humans. Figures 2 and 3 show that several manual works were changed to automation using AI and RPA. The framework first starts by loading the printed document by the user into the database through the scanner, and then the RPA downloads them as image files. A multimodal approach replaces manual document image classification. This is not only cost-effective compared with the actual business process, but it is possible to classify documents more accurately. We use fine-tuned LayoutXLM that can be classified for each document type as the multimodal approach.

Each document page is classified and then sent using RPA to the fine-tuned models for KIE based on class. These models have been fine-tuned according to the predefined process specifications of each document and use the same pre-trained LayoutXLM as the classification step. For example, the corporation's name, the representative name, opening date, registration number, and so on, corresponding to key information, are extracted in the business registration certificate. This key information is the predefined specification.

Therefore, the extracted information is converted into Excel with predefined classes, and the RPA sends an email and a notification of the completion of extraction to the user. Finally, the user processes additional manual validation and correction to ensure that the results are extracted correctly.

3.2. Unstructured Financial Document Understanding

In this step, we describe how to perform two specific tasks (i.e., document classification and KIE) on unstructured multilingual document images using a pre-trained model. One of the popular methods for achieving our goals is to fine-tune the pre-trained language model with multilingual text corpora, such as mBERT [22] and InfoLM [26]. Using the text sequences and coordinates extracted from the OCR engine, these methods can perform various tasks such as classification, KIE, and layout analysis on document images. Unlike these pre-trained language models, LayoutXLM [27] is not only pre-trained with multilingual document images rather than text corpus but uses layout information and image features for document understanding. LayoutXLM is a multilingual version of LayoutLMv2 [28,29]. This model builds an encoder using a multi-head self-attention layer stack and feed-forward network according to the transformer architecture [30]. LayoutXLM applies a spatial-aware self-attention mechanism [31] to efficiently model local invariance in the document layout. This model is trained with 11 million multilingual document images from the IIT-CDIP dataset.

As shown in Figure 4, LayoutXLM uses four types of embeddings for pretraining. In the pretraining step for document understanding, LayoutXLM uses four embeddings: text, visual, one-dimensional (1D) positional, and two-dimensional (2D) positional. Text embedding tokenizes an OCR text sequence, appending [CLS] to the beginning of the sequence and [SEP] to the end. Then, an additional [PAD] token is appended to the end so that the length is exactly the maximum sequence length L. Visual embedding uses the output function map of ResNeXt-FPN [32] with the document image as the input, and parameters are updated through backpropagation. 1D positional embedding is the output order of the OCR text sequence, which depends on the OCR engine used. The engine we used recognizes document images from top left to bottom right. The 2D positional embeddings are the coordinates of each OCR text sequence. Additionally, LayoutXLM was trained using these embeddings following three pretraining strategies:

1. Masked visual-language modeling that randomly masks and recovers some text tokens.
2. Text-image alignment that learns spatial position correspondence between image and text coordinates.
3. Text-image matching learns the correspondence between the image and text content.

This study fine-tunes the pre-trained LayoutXLM for two financial document image understanding tasks, including document image classification and KIE. We follow the general fine-tuning method [22,27–29] and update all parameters on financial document images. The KIE task aims to label words appearing in document images appropriately. For this, we fine-tune LayoutLM to predict {B, O} tags for each token and detect each entity type in the dataset using sequential labeling. Here indicating the {B, O} tag means classifying the labeled text of the document image as the first token divided by the tokenizer as B and the remaining tokens as O. For the document images classification task, we use the [CLS] token to predict the class labels using the representation of the token.

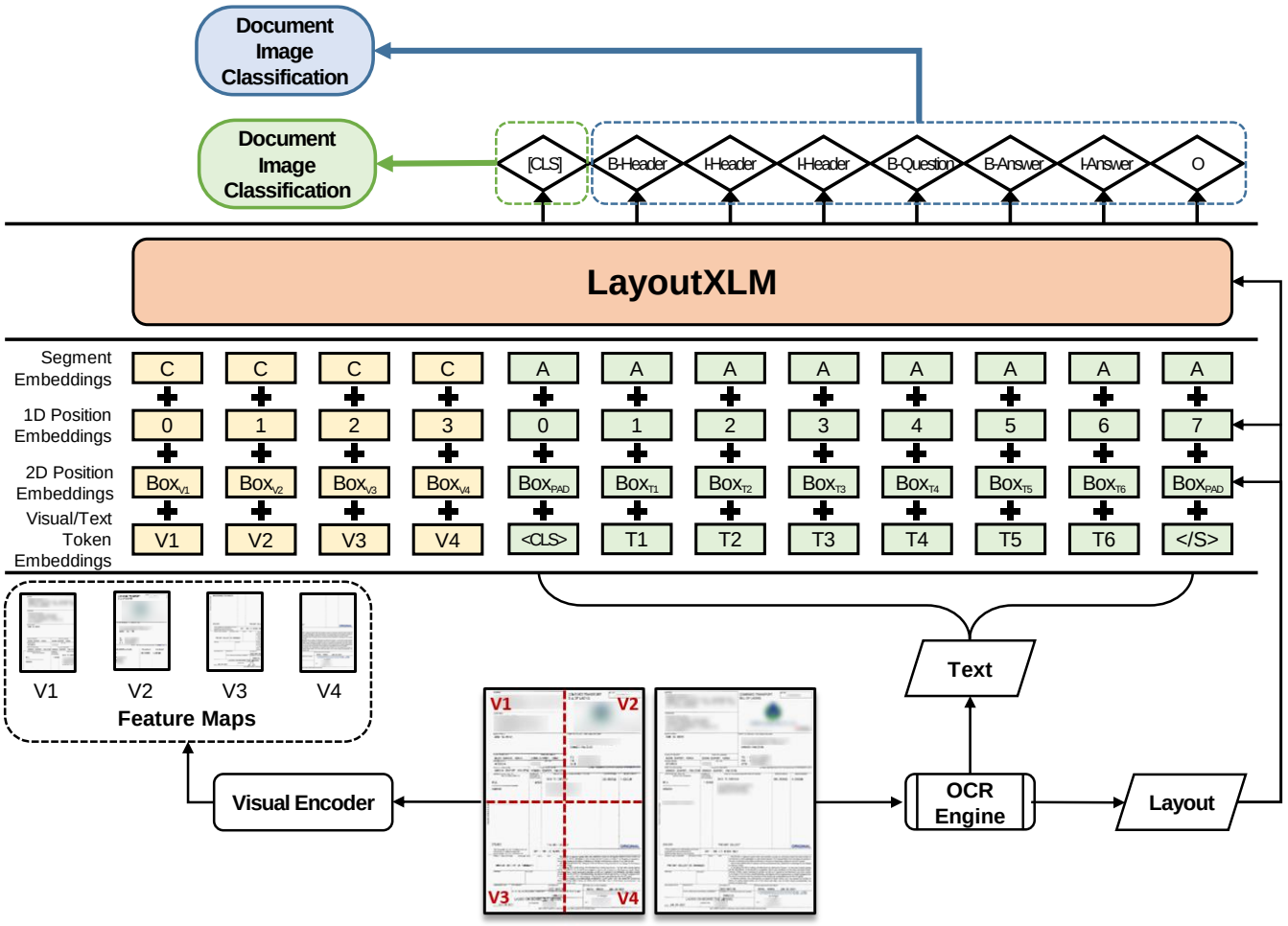


Figure 4. Structure of the unstructured document image understanding model (LayoutXLM).

4. Experiments

This section outlines the methodology and results of the experiments carried out for two objectives: (1) KIE and (2) Document Image Classification (DIC). The base models were fine-tuned using a 12-layer transformer encoder with 12 attention heads, having a hidden size of 768. The optical backbone network for LayoutXLM was built using ResNeXt-101-FPN. The proposed framework was evaluated through comparisons with several pre-trained language models, including Base size models of mBERT [23] and LayoutXLM [27], as the large models of these models have not been published. All experiments were conducted on an Intel (R) Broadwell 16-core vCPU, with 32GB DDR4 memory, a single NVIDIA A100 Tensor Core GPU, and using the Python 3.7 environment. The pre-trained models are available for download from the Hugging Face model hub [33]. They were trained using PyTorch 1.7.1 with CUDA 11.4.

To compare the performance of the different models for the two tasks, we evaluated the macro-averaged precision, macro-averaged recall, macro-averaged F1 score, and accuracy, which are widely used as evaluation metrics for NER and document classification. In the KIE task, the macro-averaged scores can avoid bias in accuracy because most entities are classified in the "other" class. The macro-averaged scores are computed using the arithmetic mean of all per-class scores. Equations (1)–(4) define these parameters as follows:

$$Precision = TP/(TP + FP) \quad (1)$$

$$Recall = TP/(TP + FN) \quad (2)$$

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (3)$$

$$F1\text{-}score = 2 \times ((Precision \times Recall)/(Precision + Recall)) \quad (4)$$

In these equations, *TP* and *TN* represent True Positive and True Negative, and *FP* and *FB* represent False Positive and False Negative, respectively.

4.1. Dataset

This study used business document images collected via our currently deployed active products between 1 March and 31 August 2022. The documents were received as images from the scanner, and the text that appeared in the document was extracted using the NAVER CLOVA OCR engine [34,35]. For labeling of the KIE task, the entities were labeled to extract in each document similar to CORD [36] and FUNSD [37], and this was worked by two people and inspected by one person. Most financial documents require various document images, such as semi-structured or unstructured. The dataset must contain key-value pairs for humans to extract data from banks manually. Therefore, we used two datasets for KIE: (1) Business Registration Certificate (BRC) and (2) Bill of Loading (BL). These financial data contain sufficient key-value pairs and satisfy the dataset mentioned above conditions.

We trained and evaluated our approach using Korean business registration certificate images as a fine-tuning dataset for the KIE task. This dataset consists of 252 scanned document images and fully annotated 18,326 entities. We divided the BRC dataset into five superclasses: 10,176 business information, 3189 tax information, 2541 document information, 570 alcohol sales information, and 1850 others. Each was divided into ten subclasses: Title, B/L number*, Shipper address*, Corporation's name, Consignee address*, Notify party information*, Place of receipt*, Place of delivery*, Port of loading*, Port of discharge*. This form is also organized as a list of interlinked semantic entities, with a few exceptions. We label the interlinked semantic entity pairs as "key" and "value" and indicate "*" in the previous sentence. Each piece of information varies depending on the type of business (individual or corporate), and alcohol sales information may be included.

We trained and evaluated our proposal for the KIE task in scanned financial document images consisting of multilingual, including English, Korean, Chinese, and so on. This dataset consists of 230 BL document images and fully annotated 60,610 entities. Table 1 shows the BL dataset statistics. The number of superclasses is five, including 888 document information, 2922 shipper information, 4989 consignee information, 3092 place, and 48,719 others. These five superclasses were divided into 18 subclasses: Title, Issue date, Issuer, Reason for issuance*, Tax status, Tax information*, Email*, Corporation name*, Representative name*, Birthday*, Opening date*, Location*, Business type*, Name of co-representative*, Registration number*, Report number*, Business scope*, Alcoholic beverage type*, Designation conditions*. Also, interlinked semantic entities of this dataset indicated "*" in the previous sentence, the same as BRC.

Table 1. Statistics of datasets for key information extraction.

BL				BRC			
No.	**Superclass**	**# Subclasses**	**Ratio**	**No.**	**Superclass**	**# Subclasses**	**Ratio**
1	Document	3	0.011	1	Document	5	0.138
2	Shipper	3	0.048	2	Tax	5	0.174
3	Consignee	4	0.082	3	Business	16	0.556
4	Place	8	0.051	4	Alcohol sales	8	0.031
5	Other	1	0.803	5	Other	1	0.101
Total		18	1.0	Total		35	1.0

As fine-tuning datasets for the DIC task, we use Korean financial documents on the 998 SWT and 1340 SD datasets, which are multilingual documents. We defined SWT dataset into four classes: defined contribution (DC) retirement pension, defined benefit

(DB) retirement pension, a combination of DB and DC (DB&DC), and other images. The SD dataset was defined into five classes as follows: commercial invoice, insurance policy, Bill of Landing (BL), packing list, and other images.

4.2. Key Information Extraction

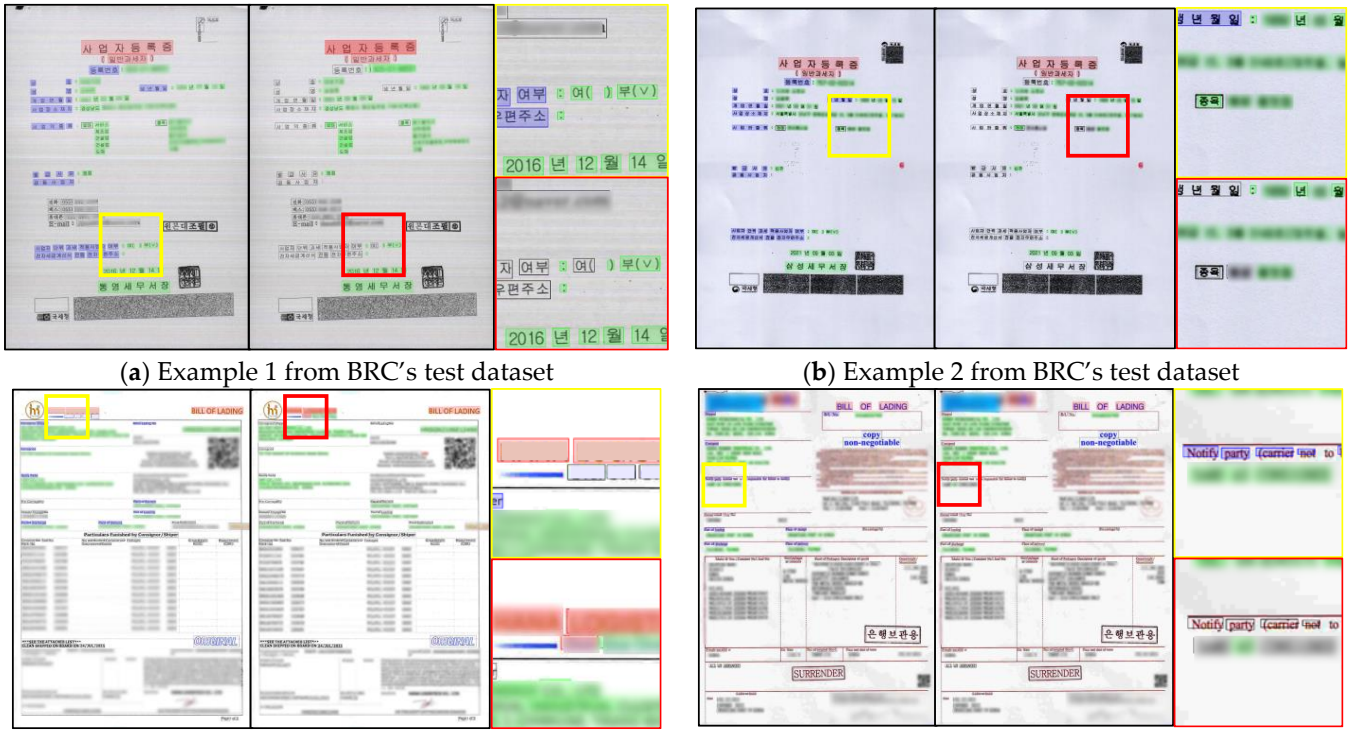
Some transformer-based models, such as mBERT, XLM-RoBERTa, InfoXLM, and LayoutXLM, are specifically designed for natural language processing tasks and have been pre-trained on a massive amount of text data. mBERT was trained on a corpus of 109GB of text data in 104 different languages. XLM-RoBERTa was pre-trained on a 2.5TB text corpus in more than 100 languages. InfoXLM was similarly pre-trained on a large text corpus in over 100 languages. LayoutXLM was trained on a dataset of 30 million document images in 53 languages. This allows them to perform well on any NLP task with fine-tuning with a smaller amount of task-specific data. XLM-RoBERTa is a variant of RoBERTa, a variant of BERT, that has been pre-trained on a diverse set of languages, resulting in better performance on cross-lingual and multilingual NLP tasks. InfoXLM is a multilingual model that is pre-trained on the text and structured data, allowing it to perform well on text-based and structured data-based NLP tasks.

To compare the performance of KIE in the BRC and BL datasets, we considered four models: mBERT, XLM-RoBERTa, InfoXLM, and LayoutXLM. Each experiment was conducted using the open pre-trained models with reference to [22,24,27,38]. Additionally, we conducted an ablation study in which the "key" entities were converted to "other" entities on BRC and BL because related studies and benchmark types are divided into two types: (1) when both keys and values are labeled, and (2) when only values are labeled. Also, because the output of the framework is predefined, we only need to be able to extract key information from within the document image. However, LayoutXLM used in the framework can understand not only the text but also the layout. Therefore, we conducted an ablation study to understand how much the layout information affects the performance.

Table 2 shows that LayoutXLM exhibited the best accuracy and macro-averaged score performance on all datasets. In the BL dataset, LayoutXLM achieves the best high macro-F1 score of 0.8874 and shows the high-performances difference from other models. This indicates that our proposed method is more effective for text-rich document images. Figure 5 shows two sampled forms from the BRC and BL datasets. In each image, the first image shows the output of a model trained with key-value labels, the second image shows the output of a model trained without key labels, and the third image shows a more detailed result. In the third image, the red text box means Header, the blue means Key, the green text box means Value, and the transparent text box means Other. These results show that a model trained without "key" labels misclassified some "value" labels as "other" labels.

Table 2. Performance of key information extraction with key-value pair.

Dataset	Modality	Model	Macro-Precision	Macro-Recall	Macro-F1	F1 of Value
BRC	Text only	mBERT_{BASE}	0.9860	0.9920	0.9888	0.9825
		$\text{XLM} - \text{RoBERTa}_{BASE}$	0.9819	0.9919	0.9867	0.9830
		InfoXLM_{BASE}	0.9684	0.9869	0.9773	0.9654
	Multimodal	LayoutXLM_{BASE}	0.9838	0.9949	0.9891	0.9889
BL	Text only	mBERT_{BASE}	0.8037	0.8561	0.8283	0.7442
		$\text{XLM} - \text{RoBERTa}_{BASE}$	0.8205	0.8626	0.8293	0.7385
		InfoXLM_{BASE}	0.8980	0.8920	0.8508	0.7676
	Multimodal	LayoutXLM_{BASE}	0.8475	0.9349	0.8874	0.8353



(**a**) Example 1 from BRC's test dataset

(**b**) Example 2 from BRC's test dataset

(**c**) Example 1 from BL's test dataset

(**d**) Example 2 from BL's test dataset

Figure 5. Examples of key information extraction results. Red, green, and blue text boxes represent the header, key, and value, respectively.

In addition, LayoutXLM achieved the best performance in all datasets in the ablation study, as shown in Table 3. In most cases of text-only pre-trained models on the BL dataset, the macro-averaged F1-score decreased by about 10%, and the recognition results of the "value" labels significantly declined for all models. However, LayoutXLM had a similar macro-averaged F1 score compared to other models despite excluding "key" labels. The recognition results of only the "value" labels decreased by about 1.5%. This indicates that if the document contains key-value pairs, such as the BRC and BL datasets, assigning labels to "key-value" pairs is effective for document understanding. Also, it shows that our proposed method is better even for document images with only "value" labels. We concluded that LayoutXLM could perform well on text-based and layout-based document understanding tasks because it is pre-trained on text and layout information.

Table 3. Performance of key information extraction without key labels.

Dataset	Modality	Model	Macro-Precision	Macro-Recall	Macro-F1	F1 of Value
BRC	Text only	mBERT_{BASE}	0.9748	0.9842	0.9765	0.9769
		$\text{XLM} - \text{RoBERTa}_{BASE}$	0.9665	0.9790	0.9725	0.9728
		InfoXLM_{BASE}	0.9703	0.9803	0.9751	0.9769
	Multimodal	LayoutXLM_{BASE}	0.9698	0.9837	0.9794	0.9773
BL	Text only	mBERT_{BASE}	0.6451	0.7290	0.6788	0.6555
		$\text{XLM} - \text{RoBERTa}_{BASE}$	0.6769	0.7675	0.7153	0.6783
		InfoXLM_{BASE}	0.6863	0.7565	0.7145	0.6918
	Multimodal	LayoutXLM_{BASE}	0.8143	0.8992	0.8504	0.8170

4.3. Document Image Classification

To compare document image classification performance, we considered three pre-trained models on the same KIE task and the CNN model. For a fair comparison, we selected VGG-16 [39] and InceptionResNet-V2 [40], which are widely used as CNN models in comparative experiments [28,29]. Table 4 and Figure 6 show the proposed method and the comparative experimental results. LayoutXLM achieved the highest classification performance in the SWT dataset but the second-highest classification performance in the SD dataset. From the perspective of large model size, transformer-based models are much larger than classical DNN models like VGG-16 and Inception-ResNet-V2, which allows them to model more complex patterns and relationships in the data. The SWT dataset is visually similar because each document image contains a table. Consequently, CNN using only visual features showed an insignificant performance. However, CNN performed better in the SD datasets with very different visual shapes for each image.

Table 4. Performance of document classification.

Dataset	Modality	Model	Accuracy	# Params.	Latency (ms)
SWT	Image only	VGG-16	0.8711	134 M	70.60
		Inception-ResNet-V2	0.8806	56 M	34.08
	Text only	mBERT_{BASE}	0.8912	110 M	14.10
		$\text{XLM} - \text{RoBERTa}_{BASE}$	0.8998	125 M	15.02
		InfoXLM_{BASE}	0.9007	110 M	14.48
	Multimodal	LayoutXLM_{BASE}	0.9829	200 M	38.02
SD	Image only	VGG-16	0.9335	134 M	70.60
		Inception-ResNet-V2	0.9637	56 M	34.08
	Text only	mBERT_{BASE}	0.8882	110 M	14.10
		$\text{XLM} - \text{RoBERTa}_{BASE}$	0.9041	125 M	15.02
		InfoXLM_{BASE}	0.9295	110 M	14.48
	Multimodal	LayoutXLM_{BASE}	0.9584	200 M	38.02

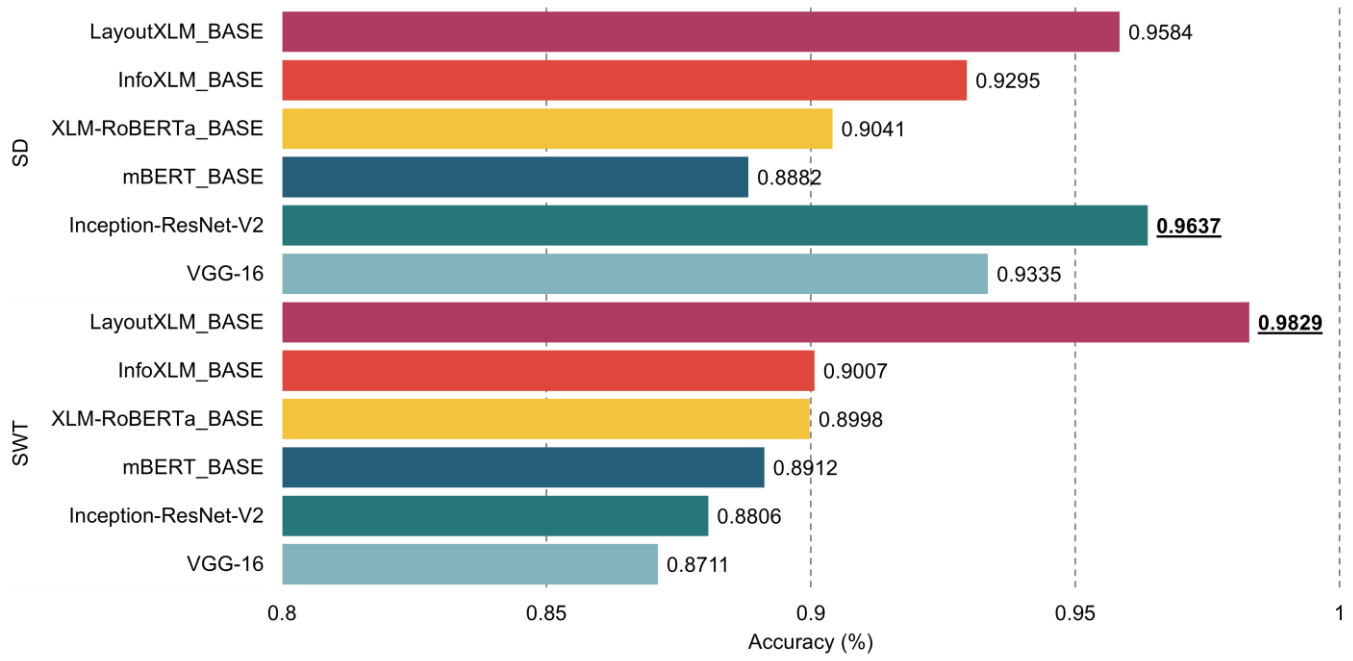


Figure 6. Performance comparison for accuracy. The values indicated in bold and underlined represent the best performance.

We conducted experiments to compare the performance of four transformer-based models, namely mBERT, XLM-RoBERTa, InfoXLM, and LayoutXLM, and two CNN models, such as VGG-16, InceptionResNet-V2, on two NLP tasks: KIE and Document Image Classification. The results showed that LayoutXLM performed the best on the KIE task, with the highest accuracy and macro-averaged score on all datasets, even after an ablation study was conducted. LayoutXLM was pre-trained on both text and layout information, which allowed it to perform well on both text-based and layout-based document understanding tasks. The results also showed that LayoutXLM had the highest performance on the SWT dataset in the document image classification task but the second-highest performance in the SD dataset. We concluded that the use of LayoutXLM can lead to improved performance in NLP tasks.

In this study, we did not calculate the number of floating-point operations (FLOPs), an important metric when comparing the efficiency of different models in Table 4. The number of FLOPs required to run a model can vary depending on the computing power available, and as technology improves, the number of FLOPs required to run a model can decrease. However, if the focus of a study is not on comparing the efficiency of different models, then calculating FLOPs may not be necessary. The efficiency of a model can be sufficiently known from the number of parameters and the measured latency. Therefore, if a study focuses on other aspects, it might be okay to avoid counting FLOPs. However, if the study focuses on comparing the efficiency of different models, FLOPs calculation would be helpful.

5. Discussion

The study presents a novel approach for the KIE task on scanned document images, using Korean BRC and BL as fine-tuning datasets. The dataset used for the study consists of business document images collected between 1 March and 31 August 2022. The text in the document images was extracted using the NAVER CLOVA OCR engine, and the entities were labeled for the KIE task. The BRC dataset comprises 252 scanned document images and 18,326 fully annotated entities, divided into five superclasses and ten subclasses. The BL dataset consists of 230 scanned document images and 60,610 fully annotated entities, divided into five superclasses and 18 subclasses. The fine-tuning datasets for the DIC task were the Korean financial document images on 998 SWT and 1340 SD datasets. An ablation study was conducted to compare the performance of the four models, namely mBERT, XLM-RoBERTa, InfoXLM, and LayoutXLM, on the BRC and BL datasets, both with and without "key" entities labeled.

The results indicate that LayoutXLM outperforms other models in terms of accuracy and macro-averaged score performance on all datasets, particularly on the BL dataset, where it achieved a macro-F1 score of 0.8874. This suggests that the proposed method is effective for text-rich document images. The study conducted an ablation study converting "key" entities to "other" entities on the BRC and BL datasets to understand the impact of layout information. In most cases, text-only models on the BL dataset resulted in a 10% decrease in macro-averaged F1 score and a significant decline in recognition of "value" labels. However, LayoutXLM maintained a similar macro-averaged F1 score despite excluding "key" labels and only had a 1.5% decrease in recognition of "value" labels. This suggests that assigning labels to "key-value" pairs is beneficial for document understanding, especially in documents with key-value pairs like the BRC and BL datasets. LayoutXLM also achieved the best performance in the DIC task on two more datasets, 998 SWT and 1340 SD.

Strengths of the proposed solution include:

1. The proposed NLP solution utilizes state-of-the-art transformer-based models pre-trained on massive text data, making it highly effective for fine-tuning with task-specific data. This provides enhanced results in NLP tasks, thanks to the vast amount of text data used in pretraining the models.

2. Using XLM-RoBERTa and InfoXLM, pre-trained on diverse languages, improves performance in cross-lingual and multilingual NLP tasks. The wide range of languages used in pretraining helps the models understand and process different linguistic structures, resulting in better results in NLP tasks.
3. InfoXLM stands out due to its ability to incorporate structured data in its pretraining process, making it well-suited for text-based and structured data-based NLP tasks. This approach, which involves fine-tuning a dataset for the DIC task, is particularly effective in categorizing financial document images.
4. The ablation study extensively analyzes the performance of four different models (i.e., mBERT, XLM-RoBERTa, InfoXLM, and LayoutXLM) with and without labeled "key" entities, emphasizing the importance of these labels for the KIE task. The study results show that the presence of labeled entities significantly improves the performance of the models in KIE tasks.
5. Using a fine-tuning dataset for the DIC task guarantees that the proposed solution is optimal for categorizing financial document images. The fine-tuning process allows the models to adapt to the specific characteristics and requirements of financial document images, resulting in more accurate categorization.
6. Comparing multiple pre-trained models to comprehensively evaluate each model's performance in the KIE and DIC tasks. This enables selecting the best-performing model for each task, ensuring that the optimal model is used for each specific NLP task.

The limitations of this study are as follows:

1. The study was confined to four pre-trained models, namely mBERT, XLM-RoBERTa, InfoXLM, and LayoutXLM, for the KIE task, and the findings may vary for different models. Therefore, it would be helpful to test various models to determine if they perform better on the KIE task.
2. The study only evaluated the performance of pre-trained models, and examining models trained from scratch would provide valuable insights. The ablation analysis was restricted to replacing "key" entities with "other" entities, but examining the effect of further modifications could yield informative results.
3. The study did not consider the efficiency of the models in terms of FLOPs, a crucial metric in model comparison. The models' performance was only tested on the KIE task, and evaluating their performance on additional NLP tasks would be informative.
4. The scalability and performance of the models on larger datasets were not considered in the study. The document image classification experiment was limited to two datasets, and including a more significant number of datasets would provide a more robust evaluation.
5. The study's findings were limited to Korean language company registration certificates and financial document images and may not be generalizable to other document images or languages. The conclusions were based on a limited number of studies, and further validation through repetition is necessary.
6. The interpretability of the models should have been analyzed, which is a critical factor when interpreting the results. The models' resilience to noise or data corruption was also not evaluated, and it would be valuable to assess their performance in such scenarios.

Future research directions include:

1. We will expand the study to include other types of document images and languages to increase the generalizability of the results. This can include different documents, such as legal or medical documents, as well as non-Korean languages.
2. We will evaluate other pre-trained models to see if they perform better or worse than the models used in this study. This can include testing new models as they are released and comparing their results to the four models used in the study.

3. We will consider other information extraction tasks to see if the proposed method is effective for other tasks. This can include evaluating the method's performance on tasks such as named entity recognition or relationship extraction.
4. We will investigate the use of the proposed method in real-world applications, such as document management systems or financial data analysis. This can help determine the practicality and utility of the method in industry settings.
5. We will investigate combining the proposed method with OCR/data validation to enhance system performance. We will also analyze model interpretability and assess resilience to noise/data corruption for accurate results interpretation.

6. Conclusions

This study proposed a multimodal approach-based IDP framework that combines a pre-trained deep learning model with traditional RPA used in banks for KIE from real-world Korean financial document images. First, we used RPA to take the document image through the database and encode it. Then, document image classification and KIE tasks were performed using the fine-tuned LayoutXLM model. The proposed method's performance was compared during the experiment by applying various state-of-the-art language models such as XLM-RoBERTa and InfoXLM. We conducted an ablation study, which involves labeling document images in which all "key-value" pairs of information to be extracted are written. The experimental results demonstrated that the multimodal approach-based VrDU pre-trained model outperformed the language models in terms of the accuracy of the KIE and document classification, which require automation in banks. The ablation study showed that labeling both "key-value" pairs increased the accuracy in all comparative experiments and was up to 10% better than labeling only values.

Although our method has a superior document comprehension level, there are cases where the token classification performance is unsatisfactory because there are few Korean document images in the pre-trained data. In future studies, we will collect any financial document images and pre-train them to solve this problem. We also plan to address the approaches and challenges to applying this system to languages other than Korean.

Author Contributions: Conceptualization, S.C. and J.M.; Project administration, J.B. and J.K.; Supervision, S.L.; Writing—Original draft, S.C.; Writing—Review & editing, J.M. and S.L.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Shinhan Bank and also supported by the Soonchunhyang University Research Fund. The views reflected in this article are the authors' views. They do not necessarily represent the viewpoints of Shinhan Bank or the other unit's members.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We would like to express our sincere gratitude to the four reviewers for their insightful and valuable feedback, which has helped us improve our work. We also offer special thanks to members of the AI Unit, Shinhan Bank, Republic of Korea.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Maqsood, H.; Maqsood, M.; Yasmin, S.; Mehmood, I.; Moon, J.; Rho, S. Analyzing the Stock Exchange Markets of EU Nations: A Case Study of Brexit Social Media Sentiment. *Systems* **2022**, *10*, 24. [CrossRef]
2. Yasir, M.; Ansari, Y.; Latif, K.; Maqsood, H.; Habib, A.; Moon, J.; Rho, S. Machine learning–assisted efficient demand forecasting using endogenous and exogenous indicators for the textile industry. *Int. J. Logist. Res. Appl.* **2022**. [CrossRef]
3. Jabeen, A.; Yasir, M.; Ansari, Y.; Yasmin, S.; Moon, J.; Rho, S. An Empirical Study of Macroeconomic Factors and Stock Returns in the Context of Economic Uncertainty News Sentiment Using Machine Learning. *Complexity* **2022**, *2022*, 4646733. [CrossRef]
4. Ansari, Y.; Yasmin, S.; Naz, S.; Zaffar, H.; Ali, Z.; Moon, J.; Rho, S. A Deep Reinforcement Learning-Based Decision Support System for Automated Stock Market Trading. *IEEE Access* **2022**, *10*, 127469–127501. [CrossRef]
5. Anagoste, S. Robotic Automation Process—The next major revolution in terms of back office operations improvement. In Proceedings of the 11th International Conference on Business Excellence, Bucharest, Romania, 30–31 March 2017; pp. 676–686.

6. Zhang, X.; Wen, Z. Thoughts on the development of artificial intelligence combined with RPA. *J. Phys. Conf. Ser.* **2021**, *1883*, 012151. [CrossRef]
7. Ling, X.; Gao, M.; Wang, D. Intelligent document processing based on RPA and machine learning. In Proceedings of the 2020 Chinese Automation Congress, Shanghai, China, 6–8 November 2020; pp. 1349–1353.
8. Kaya, C.T.; Turkyilmaz, M.; Birol, B. Impact of RPA technologies on accounting systems. *J. Acc. Financ.* **2019**, *82*, 235–250.
9. Ribeiro, J.; Lima, R.; Eckhardt, T.; Paiva, S. Robotic Process Automation and Artificial Intelligence in Industry 4.0—A Literature review. *Procedia Comput. Sci.* **2021**, *181*, 51–58. [CrossRef]
10. Lee, J.; Jeong, J.; Jung, S.; Moon, J.; Rho, S. Verification of De-Identification Techniques for Personal Information Using Tree-Based Methods with Shapley Values. *J. Pers. Med.* **2022**, *12*, 190. [CrossRef]
11. Lievano-Martínez, F.A.; Fernández-Ledesma, J.D.; Burgos, D.; Branch-Bedoya, J.W.; Jimenez-Builes, J.A. Intelligent Process Automation: An Application in Manufacturing Industry. *Sustainability* **2022**, *14*, 8804. [CrossRef]
12. Martínez-Rojas, A.; Sánchez-Oliva, J.; López-Carnicer, J.; Jiménez-Ramírez, A. AIRPA: An Architecture to Support the Execution and Maintenance of AI-Powered RPA Robots. In Proceedings of the International Conference on Business Process Management, Rome, Italy, 6–10 September 2021; pp. 38–48.
13. Lima, R.; Paiva, S.; Ribeiro, J. Artificial Intelligence Optimization Strategies for Invoice Management: A Preliminary Study. In *Communication and Intelligent Systems*; Lecture Notes in Networks and Systems; Springer: Singapore, 2021; Volume 204.
14. Lu, J.; Wu, S.; Xiang, Z.; Cheng, H. Intelligent document-filling system on mobile devices by document classification and electronization. *Comp. Intell.* **2020**, *36*, 1463–1479. [CrossRef]
15. Li, W.; Wen, S.; Shi, K.; Yang, Y.; Huang, T. Neural architecture search with a lightweight transformer for text-to-image synthesis. *IEEE Trans. Netw. Sci. Eng.* **2022**, *9*, 1567–1576. [CrossRef]
16. Lyu, B.; Yang, Y.; Wen, S.; Huang, T.; Li, K. Neural Architecture Search for Portrait Parsing. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *53*, 1158–1169. [CrossRef] [PubMed]
17. Baidya, A. Document Analysis and Classification: A Robotic Process Automation (RPA) and Machine Learning Approach. In Proceedings of the 2021 4th International Conference on Information and Computer Technologies, Kahului, HI, USA, 11–14 March 2021; pp. 33–37.
18. Roopesh, N.; Babu, C.N. Robotic process automation for resume processing system. 2021 Int. Conf. Recent Trends Electron. Inform. *Commun. Technol. (RTEICT)* **2021**, *2021*, 180–184.
19. Guha, A.; Alahmadi, A.; Samanta, D.; Khan, M.Z.; Alahmadi, A.H. A multi-modal approach to digital document stream seg-mentation for title insurance domain. *IEEE Access* **2022**, *10*, 11341–11353. [CrossRef]
20. Mungmeeprued, T.; Ma, Y.; Mehta, N.; Lipani, A. Tab this folder of documents: Page stream segmentation of business documents. In Proceedings of the 22nd ACM Symposium on Document Engineering, San Jose, CA, USA, 20–23 September 2022; pp. 1–10.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
22. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
23. Lample, G.; Conneau, A. Cross-lingual Language Model Pretraining. *arXiv* **2019**, arXiv:1901.07291.
24. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv* **2020**, arXiv:cs.CL/1911.02116.
25. Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation. *arXiv* **2020**, arXiv:2001.08210. [CrossRef]
26. Colombo, P.J.A.; Clavel, C.; Piantanida, P. InfoLM: A new metric to evaluate summarization & Data2Text generation. *Proc. AAAI* **2022**, *36*, 10554–10562.
27. Xu, Y.; Lv, T.; Cui, L.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Wei, F. LayoutXLM: Multi-Modal Pre-Training for Multilingual Visually-Rich Document Understanding. *arXiv* **2021**, arXiv:2104.08836.
28. Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; Zhou, M. LayoutLM: Pre-training of text and layout for document image understanding. *arXiv* **2019**, arXiv:1912.13318.
29. Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; et al. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Virtual Event, 1–6 August 2021; Volume 1, pp. 2579–2582.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
31. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 4–19.
32. Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
33. Hugging Face. Available online: https://huggingface.co/ (accessed on 30 January 2023).

34. Baek, Y.; Lee, B.; Han, D.; Yun, S.; Lee, H. Character region awareness for text detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9365–9374.
35. Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D.; Yun, S.; Oh, S.J.; Lee, H. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4714–4722.
36. Park, S.; Shin, S.; Lee, B.; Lee, J.; Surh, J.; Seo, M.; Lee, H. Cord: A consolidated receipt dataset for post ocr parsing. In Proceedings of the Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
37. Jaume, G.; Ekenel, H.K.; Thiran, J.-P. FUNSD: A dataset for form understanding in noisy scanned documents. *Workshop Doc. Intell. NeurIPS* **2019**, *2019*, 1–6.
38. Chi, Z.; Dong, L.; Wei, F.; Yang, N.; Singhal, S.; Wang, W.; Song, X.; Mao, X.L.; Huang, H.; Zhou, M. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. *arXiv* **2020**, arXiv:2007.07834.
39. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
40. Szegedy, C.; Ioffe, S.; Vanhoucke, V. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv* **2016**, arXiv:1602.07261. [CrossRef]