The Smart Document Processing with Artificial Intelligence

Raenu Kolandaisamy, Heshalini Rajagopal
Institute of Computer Science and Digital Innovation, UCSI University, 56000 Kuala Lumpur, Malaysia

Indraah Kolandaisamy
School of Business Management, University Utara Malaysia 06010 Sintok, Kedah

Glaret Shirley Sinnappan
Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia
E-mail: raenu@ucsiuniversity.edu.my

Abstract

This study focuses on the challenges and potential for Intelligent Document Processing (IDP) with Artificial Intelligence (AI) to manage unstructured data. A large amount of data in many different forms, such as information from the IoT, cybersecurity and more are produced during this modern Digital Age which has been widely distributed through a wide range of unformatted formats. IDP utilizes AI technologies like Machine Learning (ML), Natural Language Processing (NLP), and Computer Vision (CV), with the aim of converting unstructured data into structured, usable information. The IDP, are explored in the findings and discussion section that emphasizes its capability to automatically perform redundant tasks, reduce operating costs as well as improve employee productivity.

Keywords: Artificial Intelligence (AI), Natural Language Processing (NLP), Document Processing, Computer Vision (CV)

1. Introduction

In today's world, everything is connected to a data source, and digitization has made it possible to record everything. The current electronic era has an abundance of data available in various forms including the Internet of Things (IoT) data, cybersecurity data, smart city data, business data, smartphone data, social media data, health data, COVID-19 data, etc. This data can be structured, semi-structured, or unstructured [1].

With the emergence of new communication platforms and diverse applications such as social media, mobile applications, and digital marketing, the data delivered lacks a typical format or predefined schemas like the standard data and is unmanageable with the relational database models. Data is generated in various forms like text, audio, videos, emails, and images, which are categorized as unstructured data. Unfortunately, such data lacks structure and standardization, which leads to difficulty for most organizations to edit, search, and analyze [2].

Based on Forbes statistics, it reported that 95% of business organizations struggle to analyze unstructured data due to they do not have the required expertise to deal with unstructured data [2]. In a business company, 80% to 90% of data is in the form of unstructured format (Sukrutha, 2023). Business companies face significant challenges in effectively processing such unstructured data. In addition, the process of unstructured data analysis is always time-consuming and costly for any organization [2].

However, all organizations are required to extract valuable information from unstructured data to access functional information. Thus, automated analysis is crucial in helping the organization manage and access useful information in a structured manner. By automating the unstructured data kept in digital formats, organizations are allowed to quickly gain insight into their businesses, increase their competitive edge, improve productivity, and drive innovations. Artificial Intelligence-based (AI-based) technologies are critical for organizations that would like to adapt to automation solutions [2].

©The 2024 International Conference on Artificial Life and Robotics (ICAROB2024), J:COM HorutoHall, Oita, Japan, 2024

Intelligent Document Processing (IDP) is the conversion process of unstructured data such as email, images, and PDF documents to functional data. Intelligent Document Processing is one of the processes carried out by artificial intelligence (AI) and AI technologies are one of the best automation processes of the next generation. It involves Computer Vision (CV), Deep Learning (DL), Machine Learning (ML), etc. for data extraction [3].

The field of Artificial Intelligence (AI) is widely renowned as one of the most exciting and promising areas in the technology landscape. Analysts believe that AI will play a vital part in any business sector and industry. Besides that, the potential for value creation across the economy with machine learning (ML), a major subset of AI, is vast. It not only affects consumers, but it also transforms the way businesses operate aTsyapewtheoxlet . [4].

The increasing demand for the effective use of unstructured data necessitates the development of an AI-based unstructured document processing framework. This framework can aid organizations in extracting valuable insights from unstructured data automatically. Therefore, this research area is considered a significant and an important research topic [2].

Extracting insights from unstructured data can be challenging due to the lack of a predefined schema or format. Therefore, it is more challenging to analyze and process. Besides that, it also required the users to have a certain level of understanding to utilize the AI technologies and techniques. Thus, the main research objectives of this study are listed as below:

- To review the advantages and disadvantages of IDP with AI-based techniques for unstructured data.
- To understand the challenges with IDP to manage unstructured data.

Intelligent Document Processing (IDP) is a powerful technology that can transform unstructured and semi-structured data into a structured format. By leveraging advanced AI technologies like machine learning (ML) and natural language processing (NLP), IDP can capture, classify, and extract even the most complex and challenging data. This extracted data can then be validated and automatically entered into existing applications using Robotic Process Automation (RPA) technology [5].

According to [4], although advanced analytics and ML have been shown to provide a competitive advantage, many enterprises still struggle with their adoption. The lack of a clear AI strategy, talent, and data are major barriers, along with soft factors like leadership, decision- making, and company culture. According to [6], it is shown that there are some potential issues and challenges with using machine learning, such as collecting sufficient and relevant training data. Besides that, another challenge is the development methods to automate document processing using machine learning, particularly for complex documents. Moreover, it requires efficient and effective technology and techniques to process and analyze fast streams of unstructured data. The complexity of unstructured data poses challenges for information extraction, especially for streaming data. Structural variation and differences in data formats are the critical obstacles associated with unstructured big data [7]. To fully leverage the IDP and gain access to valuable insights, organizations must go beyond simply knowing about it and actively incorporate it into their management and utilization of data.

Extracting insights from unstructured data can be challenging due to the lack of a predefined schema or format. Therefore, it is more challenging to analyze and process. Besides that, it also required the users to have a certain level of understanding to utilize the AI technologies and techniques. Thus, the main research objectives of this study are listed as below:

- To review the advantages and disadvantages of IDP with AI-based techniques for unstructured data.
- To understand the challenges with IDP to manage unstructured data.

2. Literature Review

Intelligent Document Processing (IDP) is one of the most popular technologies in this modern era with a high demand focused on the automation of the way humans work with different types of documents on a daily basis. Besides, Digital transformation is also one of the trending topics nowadays as Robotic Process Automation and other AI technologies such as ML work toward automating the ordinary operations performed by users. These AI technologies work together to process documents intelligently with limited human involvement [8].

The possibilities of electronically exchanging structured information continue to rise significantly. Nevertheless, a large amount of business processes continue to center on the sending, receiving, and processing of unstructured documents. According to Prasad (2020), only 20 % of data in enterprises is structured and the rest of it is unstructured data with most of it locked in documents. Companies with the adaptation of an automated document processing solution reported a reduction of 50% in their staffing and labor cost [9], while processing time is reduced to less than 20 % compared to manual processing time. It shows that

automation of business document processing is the key element for businesses to be more cost-effective and cost efficient.

Robotic Process Automation (RPA) refers to the automation of service tasks to perform repetitive human tasks precisely with the assistance of AI. The developer set the task instruction with the use of some form of screen recording and defining variables. Logging into applications, copying and pasting data, opening emails, filling forms are some of the examples of tasks. Different from the traditional automation method, RPA mainly emphasizes element identification instead of screen coordinates or XPath selections which results in better intelligent interaction with the user interface [9].

An alternative to the manual processing of documents is necessary to limit human involvement, minimize error rates and manage the time and cost efficiently and effectively. In this case, the research of automatic document processing has become one of the trending topics in recent years. RPA, also known as digital employee, has become a powerful tool to facilitate the automatic processing of business documents. RPA shows its effectiveness by enabling simple configuration of robots to handle high-volume and repetitive tasks automatically based on predefined rules which not only limits the human errors, but also improves the efficiency and cost reduction [10].

AI-based approaches are capable of extracting useful information from unstructured documents automatically. AI-based approaches are used in numerous applications such as invoice digitization, health record extraction, metadata extraction, machine translation, text summarization, insurance claims processing, and contract analysis [2].

Natural Language Processing (NLP) is one of the subfields of Artificial Intelligence that focuses on empowering computers to understand and process human languages, bringing them closer to human-like language understanding. Named Entity Recognition (NER) is an important component of NLP systems, which involves the process of automatic identification of named entities in a given text or document. Named entities refer to real-world objects such as people, location, date, and time. Recognizing and extracting real named entities is vital for various research areas, including Question Answering and Summarization Systems, Information Extraction, Machine Learning, Semantic Web Search, Video Annotation, and etc [11] According to [12], significant improvements have been made in extracting information from free text using NLP techniques based on AI. NLP uses syntactic rules to automatically scan and extract named entities like names, locations, organizations, dates, and invoice numbers from unstructured text using NER. The main steps in NER involve text pre-processing, extracting NER features, then training and classification. In text pre-processing, the given text will be formatted to be understandable for machine learning algorithms. Then, features are extracted to create a numerical representation because the NLP model could not work on the text data. Lastly, the NER model will classify and categorize words and phrases. [2].

Computer Vision (CV) is also a component of AI that enables systems to identify and extract useful information from scanned document images and other visual inputs. The goal of computer vision is to extract meaningful information from images with the development of technology that enables computers or systems to understand and recognize digital images [13].

Machine Learning (ML) is a type of data-driven AI that provides the ability to learn about a system without explicit programming [14]. AI Algorithms and Machine Learning (ML) approaches have been effectively used in real-world scenarios which include the digital services, industry and commerce throughout the years. ML acts as the teacher of the machines which “teach” the efficient data handling skills by simulating the learning concept of rational beings. It can be achieved with Al algorithms that reflect diverse rational paradigms including connectionist, genetics, statistics and probabilities, based on cases, etc. The integration of Al algorithms and the ML approach provide a possibility for the exploration and extraction of information to perform tasks such as classification, association, optimization, grouping, prediction and pattern identification. [9]

Since a large amount of enterprise data these days is usually unstructured data which has no predefined schema or format, it is more complicated and challenging for users to collect, process and analyze the data without proper structure [2]. In this case, Intelligent Document Processing with AI is proven to be essential and necessary for the processing of unstructured documents.

However, there are several challenges and limitations that might affect the performance of which need to be addressed. There are major challenges regarding the use of Unstructured Data. Inaccurate text data is added by OCR text extraction in the form of noisy data sometimes. This is a key problem related to data. Unstructured data originating from multiple sources are not standard and have different formats of data. The lack of a sufficient supply of data and insufficient quality is also another challenge in data processing with unstructured data. Moreover, it is difficult to obtain information in a language that is complex such as Arabic, without the creation of dictionaries. It is difficult with information extraction methods to describe the semantic and contextual relationships among named entities in this ambiguous language.

Another challenge is posed by domain specific entities. For example, the biomedical datasets are different from any other dataset in terms of domain specific entities [2]. As the business process requires handling and processing different types of unstructured documents from the clients or suppliers such as invoices, ID cards and application forms, it poses a challenge as these documents differ in type, form and layout which require classification based on their type and nature. Therefore, the AI technologies used for document processing should be able to extract the relevant fields from the particular documents with the application of either template-based or template-free approach. Moreover, AI technologies should be able to process and improve the quality of scanned documents, which is often submitted with low quality scanners or mobile devices. The collection of relevant target data from documents is complicated when it comes to managing multi-pages of unstructured documents which consist of tables with data that span across different pages [2].

3. Method

A case study method will be used for this study due to various data and information gathering have been undertaken to study this topic. A case study is a research methodology that assists in phenomenon exploration in a certain context via various data sources. Besides, it embarks on the exploration of phenomena from different perspectives to uncover various aspects. In the case study, the study is conducted in real-time within its naturally occurring context, with the consideration that context will influence the outcome [15].

According to [16], case study research is a qualitative approach in which the investigator explores real-life, recent cases over time involving detailed data collection from numerous sources of information, and reports a case description. To present the complexity of the issue being examined, various methodological approaches will be used. A case study will include a case study database, a clear evidence chain, and multiple sources of data from research. The depth and richness of the case study description allow readers to have a better understanding of the case and to determine the applicability of findings beyond the setting. The evidence sources for case studies include questionnaires, interviews, documentation, direct observations, participant observations, archival records etc. Therefore, case studies involve trustworthy information and act as a reliable research method since case study research is an in-depth and intensive research methodology [15].

4. Result and discussion

A questionnaire is conducted to review and understand about this study and research objectives. The questionnaire results from the respondents are shown in the Fig. 1.

Fig. 1 Results from the questionnaire

RPA is the automation technology that strives to automate human tasks including business processes by utilizing software robots that interact with systems through their user interface. It can help to increase efficiency, decrease costs and efforts that humans spend on repetitive tasks [17]. One of the advantages of RPA is that it can automate numerous processes smoothly and decrease operational costs significantly due to it can handle repetitive tasks and time-saving as well as resource-saving. Besides that, prior programming knowledge is not a must to set up and implement RPA, leading to accessibility to employees who are without any programming knowledge and experience. Moreover, RPA supports most of the regular business processes with error-free automation and reduces the need for human intervention. Furthermore, software tools are scalable and do not get exhausted [2]. RPA can handle tedious and redundant tasks, thus employees can concentrate on more complicated and valuable tasks that contribute more value to the company. The employees

with more meaningful tasks can invest more time to develop new skills to be more qualified for their job [18].

On the other hand, RPA might be limited when it comes to reading unstructured documents with various layouts and formats. The documents need to be in the same format for software bots to read accurately. In addition, even minor changes in the automation application also require reconfiguration of the RPA software bot [2]. Moreover, it can be disruptive and less effective if without the appropriate tools although it seems like a good idea for a company to shift to a more advanced business system [19].

Even though it has been proven that a competitive advantage can be achieved with the use of advanced analytics and ML, enterprises nowadays are still facing significant challenges in the adoption of ML due to lack of clear strategy for AI, insufficient time and limited data available. Soft aspects which include leadership, decision making, and company culture are also some of the challenges which affect the adoption of ML [4].

It is possible to achieve adoption through the creation of a tangible business value. However, as part of product or process innovation, the main challenge is to find the right purpose and to deliver the technology in a compliant manner. In these circumstances, for the development of business value it is imperative to have deep experience in processes and a good understanding of commercial requirements. Efficiency can be achieved by automation and an increased pace of implementation of business processes, saving costs at the same time. Effectiveness refers to higher business process quality which consists of fewer errors and compliant process execution [4].

ML is expected to address the clear business needs and integrate them into the business process so that it can deliver efficiency, preferably in the form of an end-to-end automation, simplifying the access of business users to technology. In this respect, 'everyday AI' which is capable of functioning without long data preprocessing and training activities is expected. In order to achieve this simplification, it is possible to integrate the technology into the process where the data is stored and improve the user experience [4].

In addition, access to data and its quality are two of the key elements that contribute to the success of ML. Firstly, it is necessary to consider the aspects of data gathering, access and transfer, anonymisation, annotation, curation, secure storage and deletion. Nevertheless, strict observance is required in the areas of legal frameworks and general regulations. When more data is created or collected, AI must be trained and evolved over time. The algorithms are expected to learn directly from user behavior, which is then used to internalize their knowledge. ML is currently still considered as an assistance system which coexists with humans, and processes the data provided by humans. This is expected to happen automatically, in a way that human correction is used to improve algorithms and that this knowledge is not forgotten by the machine. This can be described as continuous learning. As a result, users will be able to anticipate increasing accuracy over time, build trust and depend on the results.

Furthermore, culture also plays an essential role in the introduction of machine learning and therefore needs to be adequately climate governed by innovation principles, resources, time, etc. The leadership's mindset and employees' willingness to accept change are the main reasons for this. It will have to be derived from and integrated with the proper strategy. It seems that enterprises with pragmatic and hands on attitudes tend to be more agile. They are aware of business needs, have well defined use cases, and provide the necessary environment for implementation such as fewer organizational obstacles, empowerment of staff, no fear of losing control or jobs. In fact, there seems to be a reluctance and much slower adoption by customers who are deeply entrenched in bureaucracy and hierarchies of managers without an explicit innovation culture [4].

5. Conclusion

The study focused on the potential and challenges of Information Document Processing with Artificial Intelligence techniques to manage unstructured data or documents. The prior study conducted by the researchers on IDP with AI techniques indicated that there are numerous benefits and challenges for AI-based techniques, but it is still a very popular and trending technique that most companies are adapting and utilising. According to the findings of this study, AI techniques such as RPA and ML have various benefits to companies, such as increased efficiency, reduced costs, and support for error-free automation of business processes. In addition, they are also facing some challenges, such as a lack of AI strategy, insufficient data, and culture, in the adoption process for the IDP with AI techniques. However, it can be overcome through value creation, integration, data considerations, and fostering an innovation-driven culture. IDP with AI techniques has been growing and increasingly important in each industry. Therefore, it is vital for companies to emphasise, adopt and utilise the current technologies to improve their business performance.

References

1. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.

https://link.springer.com/article/10.1007/s42979-021-00592-x
2. Baviskar, D., Ahirrao, S., Potdar, V., & Kotecha, K. (2021). Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. IEEE Access, 9, 72894-72936. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9402739
3. Sukrutha, K. S., Harini, S., & Kusuma, M. V. (2023). A Study on Intelligent Document Processing Using AWS. International Journal for Multidisciplinary Research (IJFMR), 5(4), 1-5. https://www.ijfmr.com/papers/2023/4/4308.pdf
4. Janasz, T., Mortensen, P., Reisswig, C., Weller, T., Herrmann, M., Crnoja, I., & Höhne, J. (2021). Advancements in ML-enabled intelligent document processing and how to overcome adoption challenges in enterprises. Die Unternehmung, 75(3), 340-358. https://www.nomos-elibrary.de/10.5771/0042-059X-2021-3-340/advancements-in-ml- enabled-intelligent-document-processing-and-how-to-overcome-adoption-challenges-in- enterprises-jahrgang-75-2021-heft-3?page=1
5. Cutting, G. A., & Cutting-Decelle, A. F. (2021). Intelligent Document Processing--Methods and Tools in the real world. arXiv preprint arXiv:2112.14070. https://arxiv.org/ftp/arxiv/papers/2112/2112.14070.pdf
6. Sambetbayeva, M., Kuspanova, I., Yerimbetova, A., Serikbayeva, S., & Bauyrzhanova, S. (2022). Development of Intelligent Electronic Document Management System Model Based on Machine Learning Methods. Eastern-European Journal of Enterprise Technologies, 1(2), 115.https://www.researchgate.net/publication/359049293_Development_of_intelligent_electronic_document_management_system_model_based_on_machine_learning_methods
7. Adnan, K., & Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. International Journal of Engineering Business Management, 11, 1847979019890771.https://journals.sagepub.com/doi/epub/10.1177/1847979019890771
8. Fernando, L. (2023). Intelligent Document Processing: A Guide for Building RPA Solutions (Edition 2023.1.). Notion Press.
9. Ribeiro, J., Lima, R., Eckhardt, T., & Paiva, S. (2021). Robotic process automation and artificial intelligence in industry 4.0–a literature review. Procedia Computer Science, 181, 51-58. https://www.sciencedirect.com/science/article/pii/S1877050921001393
10. Ling, X., Gao, M., & Wang, D. (2020). Intelligent document processing based on RPA and machine learning. In 2020 Chinese Automation Congress (CAC) (pp. 1349-1353).
11. Anandika, A., & Mishra, S. P. (2019). A study on machine learning approaches for named entity recognition. In 2019 International Conference on Applied Machine Learning (ICAML) (pp. 153-159). IEEE. https://ieeexplore.ieee.org/document/8989189
12. Macri, C. Z., Teoh, S. C., Bacchi, S., Tan, I., Casson, R., Sun, M. T., ... & Chan, W. (2023). A case study in applying artificial intelligence-based named entity recognition to develop an automated ophthalmic disease registry. Graefe's Archive for Clinical and Experimental Ophthalmology, 1-10. https://link.springer.com/article/10.1007/s00417-023-06190-2
13. Paneru, S., & Jeelani, I. (2021). Computer vision applications in construction: Current state, opportunities & challenges. Automation in Construction, 132, 103940. https://www.sciencedirect.com/science/article/pii/S0926580521003915
14. Bose, P., Srinivasan, S., Sleeman IV, W. C., Palta, J., Kapoor, R., & Ghosh, P. (2021). A survey on recent named entity recognition and relationship extraction techniques on clinical texts. Applied Sciences, 11(18), 8319. https://www.mdpi.com/2076-3417/11/18/8319
15. Rashid, Y., Rashid, A., Warraich, M. A., Sabir, S. S., & Waseem, A. (2019). Case study method: A step-by-step guide for business researchers. International journal of qualitative methods, 18, 1609406919862424. https://journals.sagepub.com/doi/full/10.1177/1609406919862424
16. Alpi, K. M., & Evans, J. J. (2019). Distinguishing case study as a research method from case reports as a publication type. Journal of the Medical Library Association: JMLA, 107(1),1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6300237/
17. Santos, F., Pereira, R. and Vasconcelos, J.B. (2020), Toward robotic process automation implementation: an end-to-end perspective, Business Process Management Journal, 26 (2), 405-420. https://doi.org/10.1108/BPMJ-12-2018-0380
18. Costa, D., São Mamede, H., & Silva, M. M. D. (2022). Robotic Process Automation (RPA) adoption: a systematic literature review. Engineering Management in Production and Services, 14(2), 1-12. https://repositorioaberto.uab.pt/handle/10400.2/14058
19. Ansari, W. A., Diya, P., Patil, S., & Patil, S. (2019). A review on robotic process automation-the future of business organizations. In 2nd International conference on advances in science & technology (ICAST). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3372171

Authors Introduction

Dr. Raenu Kolandaisamy

He received his PhD from the Faculty of Computer Science & Information Technology, University Malaya in 2020. He is currently an Assistant Professor in UCSI University, Malaysia. His research interest areas are Wireless Networking, Security, VANET and IoT.

Dr. Heshalini Rajagopal

She received her PhD and Master's degree from the Department of Electrical Engineering, University of Malaya, Malaysia in 2021 and 2016, respectively. She received the B.E (Electrical) in 2013 Currently, she is an Assistant Professor in UCSI University, Kuala Lumpur, Malaysia. Her research interest includes image processing, artificial intelligence and machine learning.

Dr. Indraah Kolandaisamy

She is a senior lecturer in School of Business Management, College of Business, Universiti Utara Malaysia. Indraah A/P Kolandaisamy holds a doctorate in Management from Universiti Kebangsaan Malaysia in 2015. Her D.B.A work is on organizational citizenship behavior among public sector in Malaysia. She obtained her MSc (Management) and Bachelor in International Business Management (Hons) from Universiti Utara Malaysia respectively in 2007 and 2005.

Dr. Glaret Shirley Sinnappan

She holds the position of Assistant Professor in the Department of Information and Communication Technology at Tunku Abdul Rahman University College.