iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.57926

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

Advancements in Document QnA: A Comprehensive Survey

Likhith V[1], Monish M[2], Abhishek A[3], Chandan VK[4], Usha CR[5]
Department of Artificial Intelligence and Machine Learning, K S Institute of Technology

Abstract: ***The increasing demand for effective document information extraction methods has underscored the necessity of addressing challenges related to semi-structured tables and diverse content formats. This survey extensively explores the intricate task of extracting information from documents with a particular emphasis on the challenges associated with precise Key Information Extraction (KIE) and their broader implications for enhancing document understanding efficiency. The survey delves into recent breakthroughs in this domain, with a special focus on notable approaches such as BROS, BloombergGPT, and the innovative Document Understanding Transformer (DonUT). Additionally, it provides a comprehensive analysis of various studies in Key Information Extraction (KIE) and Visual Document Understanding (VDU), elucidating the strengths and weaknesses of these endeavors. It also provides justification for highlighting DonUT lies in its unique OCR-free VDU model architecture based on Transformers, incorporating a pre-training objective that utilizes cross-entropy loss. The survey not only addresses current challenges but also illuminates promising avenues for advancing document text extraction techniques.***
Keywords: Document Information Extraction , Key Information Extraction (KIE) , Visual Document Understanding(VDU) , Transformers , Document Understanding Transformer(DonUT) , Optical Character Recognition(OCR) , Cross-entropy loss.

I. INTRODUCTION

Before commencing this exploration, it is imperative to elucidate key terminologies. Document text extraction denotes the process of retrieving semi-structured information from digital documents. Representation learning encompasses methodologies facilitating machines to discern meaningful patterns from data. Pre-trained language models are those initially trained on extensive datasets and subsequently fine-tuned for specific tasks. Spatial dependency parsing involves comprehending relationships among text tokens in documents.

In the dynamic landscape of document text extraction, where information serves as the lifeblood of decision-making, the demand for precise and efficient extraction techniques is paramount. This survey endeavors to navigate the intricacies of recent advancements in extracting valuable insights from form-like documents. From the subtleties of representation learning to groundbreaking pre-trained language models and spatial dependency parsing, the survey encapsulates a spectrum of methodologies designed to decipher the wealth of information concealed within semi-structured tables and diverse content formats.

The impetus for this survey arises from the escalating need for comprehensive solutions in document text extraction. As the volume and complexity of digital documents surge, extracting semi-structured information becomes a formidable task. Consider the challenges faced by businesses, researchers, or individuals dealing with copious amounts of information encapsulated in invoices, forms, and financial documents. Navigating through this complexity, accurately extracting key information, and comprehending contextual nuances are not merely technological challenges but fundamental prerequisites for informed decision-making.

The narrative of Natural Language Processing (NLP) unfolds as a captivating saga of technological evolution, marked by significant epochs reshaping interaction with machines. Originating with rule-based systems in the mid-20th century, NLP has undergone a metamorphosis, progressing through statistical methods and embracing machine learning. A pivotal moment emerged with the introduction of transformers in 2017, a revolutionary architecture redefining the language modeling landscape.

Transformers, with their attention mechanisms, facilitated a quantum leap in the capabilities of NLP models. Notable studies include BERT, RoBERTa, and Generative Pre-Trained Transformer (GPT), harnessing the power of large-scale pre-training to capture intricate language patterns and semantic nuances. The landscape of NLP underwent a paradigm shift as these models achieved state-of-the-art results in various language tasks.

Beyond traditional text processing, the integration of visual data ushered in a new era – Visual Document Understanding (VDU). This interdisciplinary field harmonizes linguistic intelligence with the ability to comprehend and interpret content from visual documents, transcending the limitations of Optical Character Recognition (OCR). In this context, a diverse array of models has been proposed to address challenges related to spatial relationships, contextual semantics, and varied document formats.

©IJRASET: All Rights are Reserved

Current VDU methods predominantly employ a two-stage approach, commencing with reading texts in the document image through OCR and progressing to holistic document understanding. The conventional pipeline involves separate modules for text detection, text recognition, and parsing. While these methods have yielded noteworthy results, they heavily rely on OCR, presenting inherent challenges.

The dependence on OCR introduces several challenges. Firstly, utilizing OCR as a pre-processing method is computationally expensive, even with off-the-shelf OCR engines. The cost escalates when aiming for high-quality OCR results. Additionally, off-the-shelf OCR methods lack flexibility in handling diverse languages or domain changes, compromising generalization ability. Training a dedicated OCR model demands extensive resources and large-scale datasets. Moreover, OCR errors propagate through the VDU system, particularly impacting subsequent processes. In languages with complex character sets like Korean or Chinese, OCR quality tends to be lower, exacerbating these challenges. Post-OCR correction modules are often adopted, but they inflate system size and maintenance costs, rendering them impractical in real-world application environments.

II. RELATED WORK

At the outset of the related work section, attention is directed towards endeavors within the domain of document key information extraction.

In the financial technology domain, [1] work presents BloombergGPT, a substantial 50 billion parameter language model trained on an extensive financial dataset. Uniquely constructed with a 363 billion token dataset from Bloomberg's sources, augmented with 345 billion tokens from general datasets, BloombergGPT outperforms existing models on financial tasks without compromising performance on general language model benchmarks. The paper elucidates modeling choices, training processes, and evaluation methodologies, offering a valuable contribution to NLP in finance.

In [2] , Address on the limitations of existing IOB tagging or graph-based formulations in structured information extraction (SIE), this paper proposes a novel formulation inspired by anchor-based object detectors in vision. It introduces DocTr, a Document Transformer, to detect and associate entity bounding boxes in visually rich documents. The approach, robust to text ordering, introduces a simple pre-training strategy for entity detection in the context of language, outperforming existing solutions on three SIE benchmarks.

Introducing DocFormerv2 [3] , a multi-modal transformer for Visual Document Understanding (VDU), the research addresses the challenges of understanding documents beyond OCR predictions. DocFormerv2, pre-trained with unsupervised tasks, exhibits state-of-the-art performance over strong baselines on nine datasets, showcasing its understanding of multiple modalities for VDU.

In [4] , the focus is on key information extraction (KIE) from document images, this paper introduces BROS, a pre-trained language model emphasizing the effective combination of text and layout. BROS, or BERT Relying On Spatiality, encodes relative positions of texts in 2D space, showcasing comparable or superior performance on four KIE benchmarks without relying on visual features. Real-world challenges in KIE tasks, such as minimizing errors from incorrect text ordering and efficient learning from fewer examples, are addressed, demonstrating the superiority of BROS over previous methods.

In [5] , methods to improve the serialization of forms, FormNet introduces a structure-aware sequence model leveraging rich attention and super-tokens. By explicitly recovering local syntactic information lost during serialization, FormNet outperforms existing methods on CORD, FUNSD, and Payment benchmarks with a more compact model size and less pre-training data.

In [6] , the realm of Visual Document Understanding (VDU), an OCR-free model for document understanding. Addressing challenges associated with OCR-based approaches, Donut employs a Transformer architecture with a cross-entropy loss pre-training objective. The model achieves state-of-the-art performances on various VDU tasks in terms of both speed and accuracy, overcoming computational costs, inflexibility, and error propagation associated with OCR engines. The introduction of a synthetic data generator enhances Donut's flexibility across languages and domains, marking a significant stride in OCR-free VDU research.

In [7] , Address the complexity and cost associated with multiple modules in traditional pipeline-based information extraction (IE) systems, the paper advocates for an end-to-end model. By formulating document IE as a sequence generation task, the paper demonstrates that a single end-to-end IE system can be built to achieve competent performance, simplifying development and maintenance in large-scale production.

The paper [8], delves on methods to overcome limitations in the traditional sequence tagging approach for information extraction (IE) from semi-structured document images, this work introduces SPADEs (SPAtial DEpendency parser). Formulated as a spatial dependency parsing problem, SPADEs models complex spatial relationships and an arbitrary number of information layers in documents in an end-to-end manner. Evaluation across various documents shows comparable or better performance compared to strong baselines.

In [9], research introduces a groundbreaking approach utilizing representation learning to address the challenge of extracting structured information from form-like document images. The proposed system leverages knowledge about target field types to generate extraction candidates and employs a neural network architecture for learning dense representations based on neighboring words. These learned representations not only enhance extraction for unseen document templates but also offer interpretability, as demonstrated through loss cases.

The paper [10], emphasis on Innovating document image understanding, LayoutLM introduces a joint learning framework for text and layout information across scanned documents. Leveraging image features to incorporate visual information, LayoutLM achieves state-of-the-art results in form understanding, receipt understanding, and document image classification. The pioneering approach marks the first time text and layout are jointly learned in a single framework for document-level pre-training.

The work [11] proposes a graph-based approach for detecting tables in document images, particularly in unconstrained formats. Utilizing Graph Neural Networks (GNNs), the model describes local repetitive structural information of tables based on location, context, and content type, independent of language quality. Experimentally validated on two invoice datasets, the proposed model achieved encouraging results, and the novel dataset contributed to the community enhancing benchmarking opportunities.

The paper [12], mentions replication study evaluates the impact of key hyperparameters and training data size in BERT pretraining. The findings reveal that BERT was significantly undertrained and can match or exceed the performance of subsequent models. Achieving state-of-the-art results on GLUE, RACE, and SQuAD, this research underscores the importance of overlooked design choices, questioning the sources of reported improvements.

TABLE I
COMPARISON WITH SIMILAR WORKS

Sl. No.	Year	Title	Description
1	2023 [1]	BloombergGPT: A Large Language Model for Finance	BloombergGPT, tailored for financial applications, exhibits remarkable performance in this domain. However, its applicability outside finance is constrained due to a lack of domain specificity, potentially limiting its versatility. Furthermore, the model's immense size, with 50 billion parameters, contributes to significant computational costs and extensive training times, posing practical challenges for broader deployment.
2	2023 [2]	DocTr: Document Transformer for Structured Information Extraction in Documents	DocTr presents an innovative solution for structured information extraction. Despite its novel approach, the model's domain-agnostic nature may result in suboptimal performance in specialized fields. The reliance on a pre-training strategy adds a layer of complexity, especially in scenarios where annotated data for fine-tuning is limited.
3	2023 [3]	DocFormerv2: Local Features for Document Understanding	DocFormerv2 showcases state-of-the-art performance in Visual Document Understanding (VDU). However, its efficacy is contingent upon the quality and diversity of pre-training tasks, potentially limiting its adaptability to languages or domains not well-covered in pre-training. This poses challenges for understanding languages beyond the model's training scope.
4	2022 [4]	BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents	BROS takes a unique approach to key information extraction, encoding 2D spatial information. While it effectively addresses information extraction challenges, the model may encounter difficulties with intricate layouts or non-standard document structures. Additionally, its efficiency in handling diverse languages and domains requires further exploration.
5	2022 [5]	FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction	FormNet excels in correcting token serialization issues, enhancing information extraction from forms. However, its performance might be compromised when dealing with forms that significantly deviate from the structures seen in its training data. This limitation raises concerns about the model's adaptability to unconventional layouts.

6	2022 [6]	OCR-free Document Understanding Transformer	Donut's OCR-free approach overcomes challenges associated with OCR, yet its effectiveness relies on the quality of the synthetic data generator. The Transformer architecture, while advantageous, may face difficulties with highly complex document structures. These limitations suggest considerations for improving synthetic data quality and handling intricate layouts.
7	2021 [7]	Cost-effective End-to-end Information Extraction for Semi-structured Document Images	The proposed end-to-end IE model simplifies development but may encounter performance degradation with highly specialized or complex document types. Its reliance on a sequence generation task might limit efficacy in scenarios with varied document structures, prompting further exploration of tailored solutions for diverse document types.
8	2021 [8]	Spatial Dependency Parsing for Semi-Structured Document Information Extraction	SPADEs efficiently tackle spatial complexities in information extraction. However, it may face challenges with diverse document types, particularly those with unconventional spatial relationships. The model's performance could be impacted when applied to documents with irregular layouts, suggesting a need for robustness in handling varied spatial structures.
9	2020 [9]	Representation Learning for Information Extraction from Form-like Documents	The system excels in extracting structured information but may lack domain specificity, affecting precision for specialized document types. Additionally, the computational cost associated with representation learning poses limitations, emphasizing the need for efficient solutions to handle large-scale datasets.
10	2020 [10]	LayoutLM: Pre-training of Text and Layout for Document Image Understanding	LayoutLM innovates document-level pre-training, incorporating layout and style information. However, challenges arise in scenarios where such information plays a less crucial role. The model's reliance on visual details may lead to computational costs, especially when dealing with large-scale document datasets.
11	2019 [11]	Table Detection in Invoice Documents by Graph Neural Networks	The graph-based approach excels in table detection, leveraging structure perception. However, its effectiveness may vary with irregular or non-tabular document structures. The model's reliance on structure perception may limit its performance on documents with unconventional layouts, necessitating improvements for enhanced adaptability.
12	2019 [12]	RoBERTa: A Robustly Optimized BERT Pretraining Approach	While shedding light on BERT's undertraining, the study's limitations include a lack of openness, hindering community collaboration. Additionally, BERT's general language understanding might not be universally strong, impacting performance across diverse linguistic contexts. Improved transparency and language adaptability could enhance the study's contributions.

III.CONCLUSIONS

This survey navigated the dynamic landscape of document text extraction, unraveling the evolution of key information extraction (KIE) methodologies and the symbiotic relationship between NLP, transformers, and Visual Document Understanding (VDU). The challenges posed by semi-structured tables and diverse content formats were dissected, emphasizing the demand for precise extraction techniques. As the survey traversed recent breakthroughs, BROS emerged as a beacon, showcasing the prowess of a pre-trained language model optimizing text and layout information for superior KIE across multiple benchmarks. The financial domain found a champion in BloombergGPT, a robust 50 billion parameter language model tailored for diverse tasks. However, the star of the show was undoubtedly DonUT, pioneering an OCR-free VDU model.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

DONUT's Transformer architecture, coupled with a cross-entropy loss pre-training objective, heralds a new era, eliminating the computational costs, inflexibility, and error propagation associated with OCR engines. Its synthetic data generator amplifies flexibility, proving instrumental in various languages and domains.

The survey has also addressed forecasted promising avenues for the future of document text extraction. The amalgamation of NLP, transformers, and VDU stands as a testament to the relentless pursuit of efficiency and accuracy in deciphering the wealth of information embedded in form-like documents. The torch is passed to an exciting future, where continuous innovation will shape the next chapter in document understanding and extraction.

REFERENCES

[1] Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg and Gideon Mann. "BloombergGPT: A Large Language Model for Finance." *ArXiv* abs/2303.17564 (2023).

[2] Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R. Manmatha, Vijay Mahadevan. DocTr: Document Transformer for Structured Information Extraction in Documents. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19584-19594 (2023).

[3] Appalaraju, Srikar & Tang, Peng & Dong, Qi & Sankaran, Nishant & Zhou, Yichu & Manmatha, R.. DocFormerv2: Local Features for Document Understanding.(2023).

[4] Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., Park, S.: Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. Proceedings of the AAAI Conference on Artificial Intelligence 36(10), 10767–10775 (Jun 2022).

[5] Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, Tomas Pfister. FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction (2022).

[6] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-Free Document Understanding Transformer. In Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII. Springer-Verlag, Berlin, Heidelberg, 498–517.

[7] Hwang, W., Lee, H., Yim, J., Kim, G., Seo, M.: Cost-effective end-to-end information extraction for semi-structured document images. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3375–3383. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021).

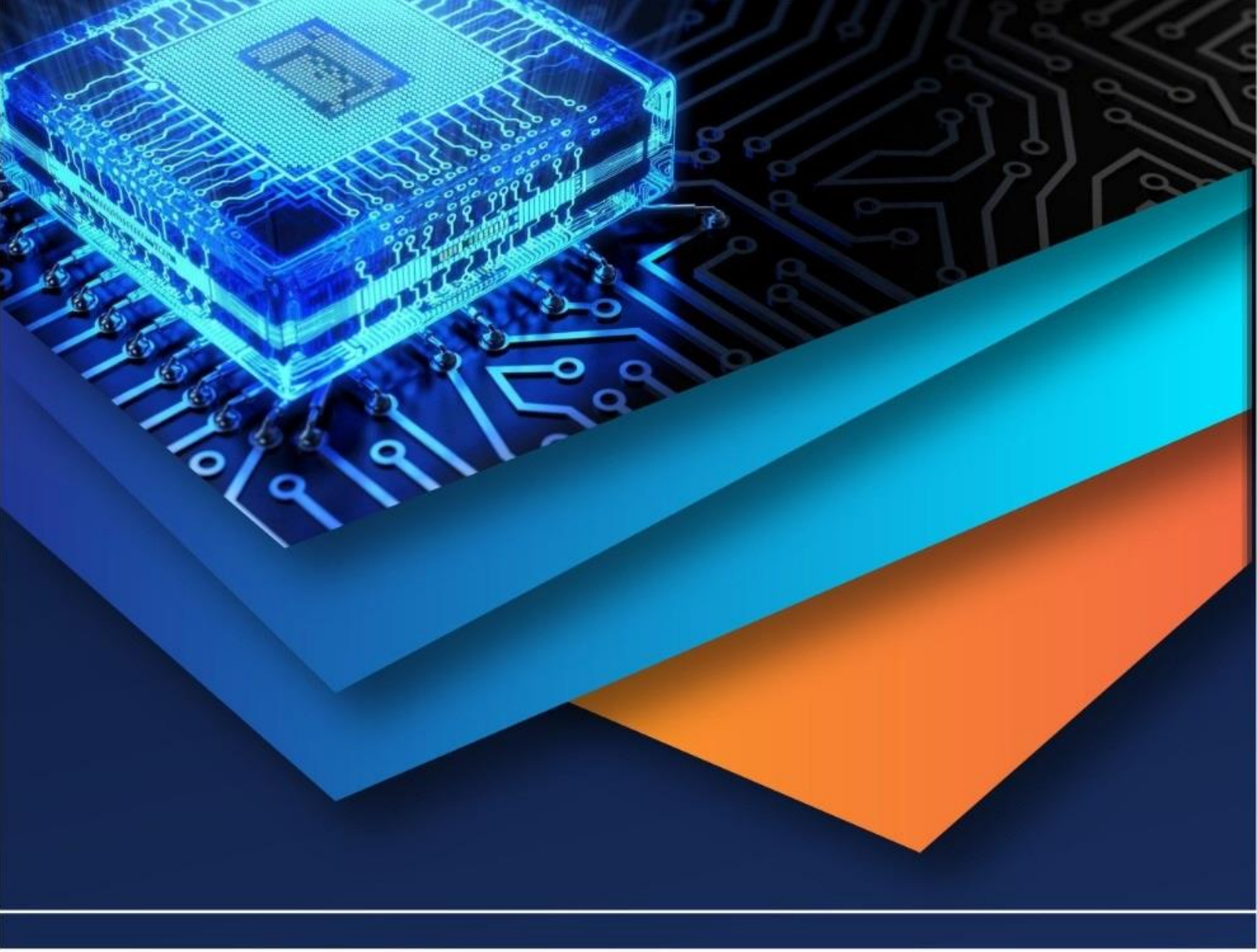
[8] Hwang, W., Yim, J., Park, S., Yang, S., Seo, M.: Spatial dependency parsing for semi-structured document information extraction. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 330–343. Association for Computational Linguistics, Online (Aug 2021).

[9] Majumder, B.P., Potti, N., Tata, S., Wendt, J.B., Zhao, Q., Najork, M.: Representation learning for information extraction from form-like documents. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6495–6504. Association for Computational Linguistics, Online (Jul 2020).

[10] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 1192–1200. KDD '20, Association for Computing Machinery, New York, NY, USA (2020).

[11] Riba, P., Dutta, A., Goldmann, L., Forn´es, A., Ramos, O., Llad´os, J.: Table detection in invoice documents by graph neural networks. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 122–127 (2019).

[12] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized Bert pretraining approach." 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)