ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1130

AI-Based Purchase Order Automation Using Confidence-Aware Hybrid Extraction and ERP Integration

Gowtham S.[1]; Radhika M.[2]; Maduvanthi S.[3]; Thulasi P.[4]; Sanchith Shanmugha Sundaram R.[5]; Muthamizh Kavi E.[6]

[1]UG Scholar; Department of AI&ML IFET College of Engineering Villupuram, India
[2]Assistant Professor; Department of AI&ML IFET College of Engineering Villupuram, India
[3]Assistant Professor; Department of AI&ML IFET College of Engineering Villupuram, India
[4]Assistant Professor; Department of AI&ML IFET College of Engineering Villupuram, India
[5]UG Scholar Department of AI&ML IFET College of Engineering Villupuram, India
[6]UG Scholar Department of AI&ML IFET College of Engineering Villupuram, India

Publication Date: 2026/05/19

Abstract: However, some firms still opt to handle email purchase orders manually, leading to inefficiency, mistakes, and unwanted delays. In this regard, the emails are not only written in simple text form but are also scanned and/or provided as PDF files, thereby complicating the process of extracting data from such emails. This research suggests the use of a completely automated process that will manage the emails in order to structure the extracted data for use by the ERP system. For this reason, a confidence-aware hybrid approach was applied to extract data about the products, quantity, and even shipping details from the purchase orders based on the language model, named entities, and rule approaches. With the use of the confidence-aware approach, the system identifies reliable data, while the input quality controller handles text misrecognition and differences in email format. The whole process involves retrieving emails, doing OCR, validating the data, and importing into the ERP system.

Keywords: *Purchase Order Automation, Optical Character Recognition (OCR), Named Entity Recognition (NER), ERP Integration.*

How to Cite: Gowtham S.; Radhika M.; Maduvanthi S.; Thulasi P.; Sanchith Shanmugha Sundaram R.; Muthamizh Kavi E. (2026) AI-Based Purchase Order Automation Using Confidence-Aware Hybrid Extraction and ERP Integration.
International Journal of Innovative Science and Research Technology, 11(5), 583-588.
https://doi.org/10.38124/ijisrt/26may1130

I. INTRODUCTION

Most businesses still prefer using manual methods when it comes to processing purchase orders sent through emails. Generally, employees check the emails, extract the relevant information, and then key it into the ERP system, which includes the name of the item ordered, the quantity, and other details about shipment. This is very time-consuming and also increases the chances of making mistakes, especially when dealing with many emails. The problem gets even more complicated due to the fact that most emails contain unstructured text as well as attachments in the form of PDFs or JPEGs.

The technologies that are currently available for automation cannot be used here since they operate under the assumption that there will be some template to work with and certain patterns will exist. The truth is that emails can be quite different from one another, and people often write them using natural language and with missing information. Moreover, the scanned documents usually come out very poorly, making it hard to extract information using one technique alone.

This problem is tackled by the suggested solution by means of using a confidence-aware hybrid extraction system that employs different extraction techniques involving language model named entity recognition and rule-based parsing. The information relevant to purchase orders is extracted independently using different methods and a confidence-aware method for selecting the most accurate values for every attribute in a PO record. A document analysis is conducted on the input to take into account possible problems related to OCR noise and variations in the structure of emails. The entire process of handling emails includes several stages from IMAP document fetching up to

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1130

converting documents via OCR and validating and entering data into ERP systems.

II. RELATED WORKS

The area of extracting automated information from unstructured business texts has received ample attention via methods based on Optical Character Recognition (OCR) and Natural Language Processing (NLP). In a paper by Li et al. [1], they looked into the topic of entity extraction from business emails using NLP-based methods to show how structured information can be extracted from unstructured text. However, these methods are very sensitive to text quality, particularly with text obtained from scanning documents.

Modern studies have focused on merging machine learning algorithms, such as NER, with statistical and similarity approaches to boost extraction accuracy. Wiriyapistan and Sinthupinyo [2] developed a hybrid approach by merging conditional random fields with similarities to perform data extraction in unstructured texts. While these approaches refine the process of entity recognition, they still rely on one pipeline for data extraction, which limits their capability to handle emails with different structures and missing data.

Models like Chargrid [3] and LayoutLM [13] based on deep learning have shown significant advancements using both textual and visual information for document understanding tasks. Such models work well for structured documents such as invoices; however, they need a lot of data for training and can't adapt well to informal emails. Models like DocFormer [10] and LayoutLMv2 [9], which focus on multilayered understanding, are also complex to develop.

Further research has been done on incorporating extracted information into business applications. In their study, Krieger et al. [7] presented a machine learning approach to invoice processing, noting that the benefit of efficiency could be gained due to ERP incorporation. However, current systems have been developed with consideration of structured/semi-structured documents and cannot handle dynamic nature of emails or OCR errors.

From the reviewed literature, it can be observed that though each extraction method is good and efficient when tested in a lab environment, they become inflexible and unreliable when applied to actual cases using emails which might not be structured at all. There is no provision in most of the current models to measure the reliability of extracted data by comparing results from different methods. The proposed model employs a confidence-based hybrid extraction technique.

III. PROPOSED SYSTEM

The design of the system is such that it has been able to automate the process of extracting purchase orders from emails, and then feed these purchase order extracts into the ERP systems. Using its ability to analyze both the unstructured nature of the email content as well as the unstructured data provided by optical character recognition software (OCR), it becomes possible for the extraction process to give more precise outcomes. A confidence-aware hybrid extraction technique succeeds in achieving its objectives through the combination of different processes. The emails are processed in a sequential manner, starting with retrieving up to storing the data in structured form.

- *System Overview*

The developed framework is expected to process automatically the PO information from emails and extract it into the form that can be used in the enterprise systems. The first step is involve getting the incoming emails from the inbox based on the use of IMAP protocol. Then, the next step is related to selecting emails containing PO information. As the information may appear both in the body of the email and its attachments, the latter is need to be obtained and processed along with the emails themselves. If some documents are attached in PDF format or image, it is needed to scan the text from it, thus turning an attachment into the machine readable text. Once the text is ready, it should be preprocessed, which includes eliminating of any unnecessary information, formatting problems and other kinds of noise. As soon as the text becomes clear enough, it is transferred to the hybrid extraction block, where several techniques is work simultaneously, each trying to detect various fields in the PO, such as the product name and number, date and other relevant information. The results of each extraction technique is compared with the others by means of applying the confidence method, which is help choose the best output.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1130

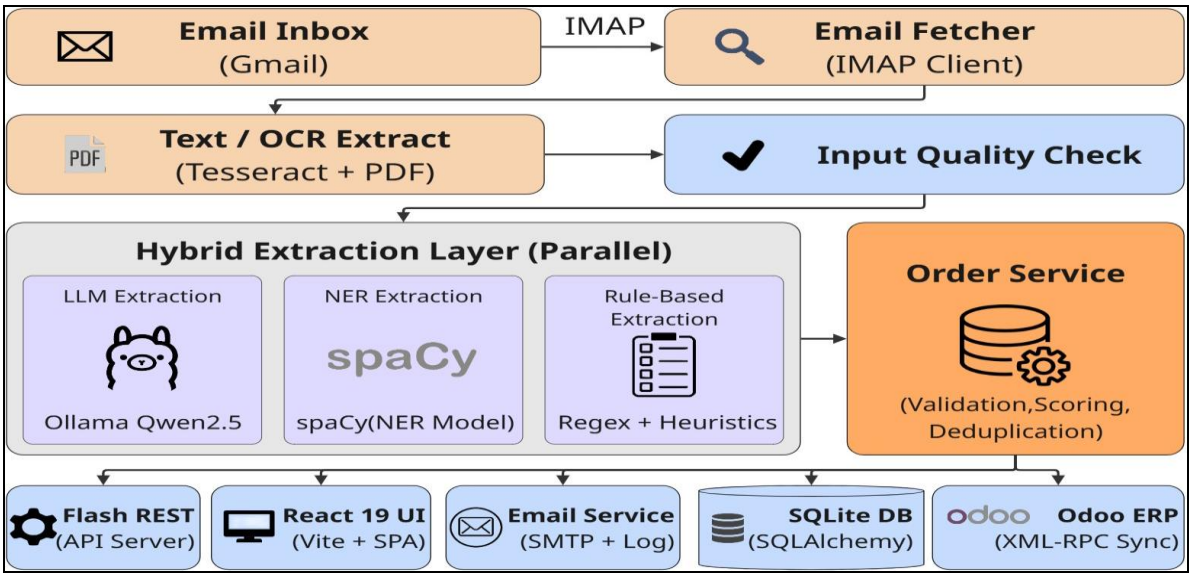

Fig 1 System Architecture

➢ *Input Quality Assessment*

The assessment of the quality of the input text takes place in order to consider mistakes in the text owing to OCR and variability of emails. There are several factors that are taken into account in the process of assessing the quality of text, for example, completeness, key phrases, noise, and other features. Based on analysis, a quality grade can be defined.

➢ *Hybrid Extraction Approach*

A combination of multiple extractor models is used simultaneously to improve the efficiency of information extraction process for emails that contain non-structured data. Language modeling aids in understanding the text context and produces structured output, whereas the named entity recognition model recognizes entities like the product name, quantity, and date. Extraction rules are applied to extract structured data from semi-structured documents.

The methods run concurrently using similar input, which means that several possible values are provided for each field value. Such an approach considers different styles of writing, missing values, and possible mistakes made when using OCR technologies. Combining contextual analysis with recognizing patterns guarantees the extraction of relevant information from the source.

This system uses a confidence-weighted approach in order to evaluate the results generated by different methods of extraction and to decide upon the final structured outcome. Every extraction method is generate a certain value for particular features, and these results is evaluated using a scoring process that is take into account such things as the reliability of the methods and the completeness of the data generated. This scoring process takes place using the following steps.

$$S_{i,j}=C_{i,j}\times Q\times W_j \quad (1)$$

$S_{i,j}$ is the score associated with technique i in field j, $C_{i,j}$ being the confidence level of the extraction technique, Q is the input quality score, and W_j refers to the weight of importance of the field. The selected values are then determined based on these scores using the rule specified below:

$$V_j=\text{argmax}(S_{i,j}) \quad (2)$$

Here (V_j) refers to the ultimate choice made for field (j). The method chosen at this level means that the process does not depend on one specific algorithm for extracting information, but uses several algorithms instead, giving more flexibility to the process. As a result, this allows the process to deal with poor quality data better and reduce the number of errors.

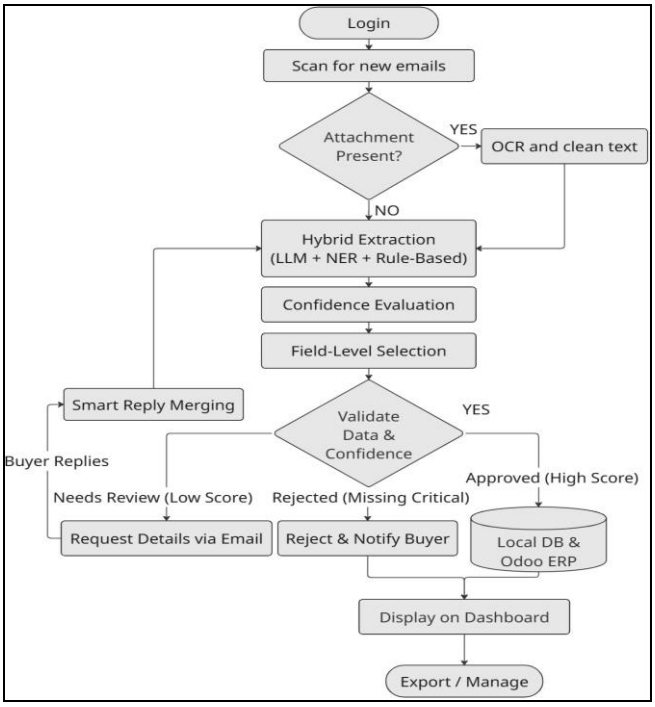

Fig 2 Workflow of the Proposed Method

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1130

➢ *Data Validation and ERP Integration*

The process involves an additional stage of verification after the fusion process based on confidence which determines the final set of values used, thus making sure that the information gathered fulfills the necessary requirements in terms of precision and consistency before being put into further use. In terms of verification, it guarantees that all of the fields necessary such as the name of the product, quantity, and other information related to delivery are present, as well as making sure that the numerical information falls into acceptable levels. Moreover, this step ensures that the data is consistently formatted throughout, thus minimizing the chance for errors during processing caused by lack or inconsistency of information.

The system is store the validated information in a structured form on the database, and then it is used for processing by means of the ERP system. This allows companies to use the purchase order information extracted to perform functions such as managing inventory, tracking orders, and controlling procurement without having to input information manually. Thus, the procedure allows transforming unstructured email into structured enterprise data, improving data management. Performance is enhanced due to validation and automatic ERP integration, minimizing risks of any errors and maintaining credible data for business purposes.

IV. RESULT AND DISCUSSION

This part of the paper focuses on testing the efficiency of the developed system in extracting purchase order information from email. The dataset used for the evaluation purpose consists of email samples, including text, PDF attachment files, and images scanned into emails. The samples have been collected from different environments (both simulated and real life). The main emphasis has been put on the following areas in the testing process: information extraction precision, behavior of the system after validation stage, and efficiency in reducing manual handling.

➢ *Extraction Performance Analysis*

The extraction process was evaluated on the basis of how well the individual methods used in the extraction process performed compared to the hybrid technique that was created. While the individual methods such as the language model, named entity recognition, and rules-based parsing perform effectively under controlled conditions, they become inconsistent when faced with varied email styles and unreliable optical character recognition outputs. The hybrid method uses all these approaches and utilizes a confidence-based approach to select the best value for every field.

The performance evaluation results using quantitative analysis are presented in Table I and show the superiority of the hybrid approach in terms of accuracy compared to individual methods. This can also be illustrated in Fig. 3, where it can be observed that the proposed method has superior performance compared to individual methods.

Table 1 Performance Analysis of Extraction Methods

Metric	Value
OCR Text Extraction Accuracy	82–88%
LLM Extraction Accuracy	88%
NER Extraction Accuracy	82%
Rule-Based Accuracy	75%
Proposed Method Accuracy	93%
Avg. Processing Time	8–12 s/email
Manual Effort Reduction	75–85%

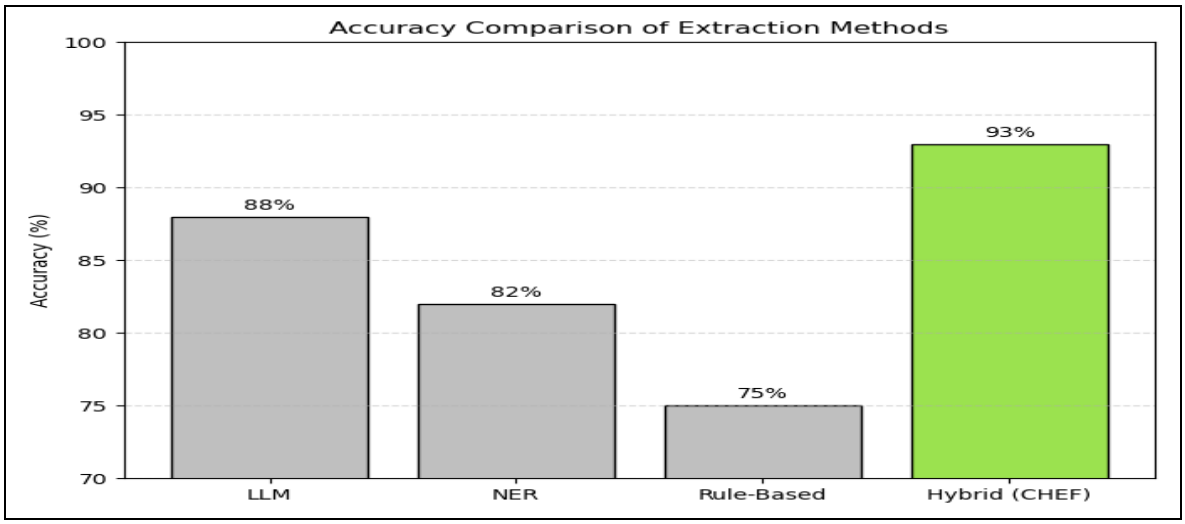

Fig 3 Extraction Accuracy Comparison

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1130

➢ *Validation Outcome Analysis*

After the data extraction and fusion process, there is a validation process to verify whether the data can be used immediately or needs further verification. Validation results is divided into three categories, namely approved, needs review, and rejected. The approved category is the largest group among all the processed data; this proves that the system is capable of producing accurate results from the input data. On the other hand, a small amount of processed data falls under the needs review category because of low confidence levels, while only a few of the data are placed in the rejected category due to insufficient data.

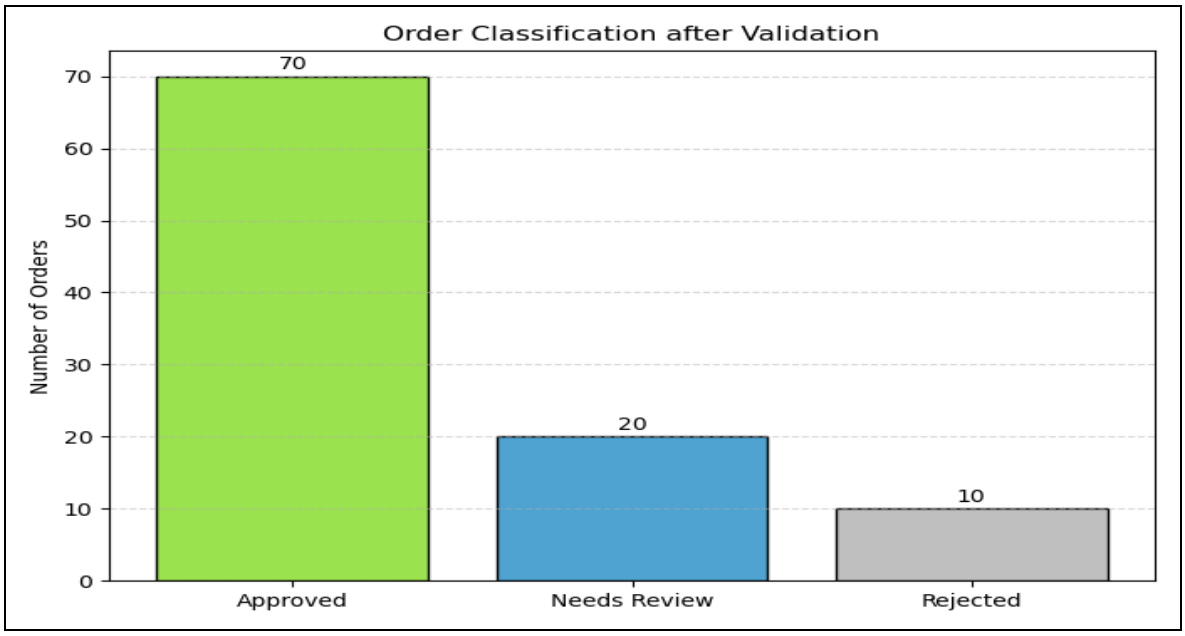

Fig 4 Order Classification Based on Validation Results

➢ *Discussion*

The findings of the experiment show that the developed system is an effective solution for automatic purchase order processing via emails. The use of the hybrid extraction approach leads to improved accuracy through the application of several methods of data extraction, whereas the confidence-based approach allows selection of the most trustworthy values. Input quality evaluation makes the system robust to different types of input data, such as OCR errors or different formats of documents. The system is highly valuable, as it successfully processes various types of texts; nevertheless, it may face challenges with processing very poor inputs. The integration with ERP systems makes it possible to use data in the users' business processes.

V. CONCLUSION AND FUTURE WORK

The research highlights the automation process used in generating structured data through email purchase orders for use in an ERP. The research aims to solve the problems that arise from manual data entry through its implementation of a hybrid extraction solution which employs multiple extraction techniques that include language models and named entity recognition and rule-based methods. The system operates through multiple techniques which work together to produce results and uses a confidence-based decision mechanism to determine the most trustworthy data for each field. The method increases data extraction accuracy through unstructured email content and OCR text processing.

The output from the system shows that when many extraction tools are combined, more accuracy can be achieved than when only one extraction method is used. This system enables the reduction of manual labor involved in processing the purchase orders, while at the same time ensuring that the processing time does not exceed acceptable limits. The system enhances process reliability by incorporating quality checks on the inputs before being loaded into the ERP system.

Future improvements will concentrate on developing solutions for situations which require processing extremely degraded input data that includes documents with severe distortions and text that is difficult to read. The system can be extended through the implementation of learning systems which use feedback to enhance extraction efficiency as time progresses. The system will gain broader use in various locations through its ability to support multiple languages. The solution will achieve better deployment results in extensive operational environments through system performance enhancements and improved connections to enterprise platforms.

REFERENCES

[1]. J. Li, S. Sen, and N. Zaman, "Entity extraction from business emails," *International Journal of Information Technology and Computer Science*, vol. 7, no. 9, pp. 15–22, Aug. 2015.

[2]. S. Wiriyapistan and S. Sinthupinyo, "Extracting structured data from unstructured text using conditional random field and Jaccard similarity," in

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1130

Proc. 11th Int. Conf. Information Technology (ICIT), 2019, pp. 103–106.

[3]. A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Höhne, and J. Faddoul, "Chargrid: Towards understanding 2D documents," in *Proc. EMNLP*, 2018, pp. 4459–4469.

[4]. C. Sage, R. Aussem, and H. Elghazel, "End-to-end extraction of structured information from business documents with pointer-generator networks," in *Proc. Workshop on Structured Prediction for NLP*, 2020, pp. 43–52.

[5]. X. Holt and A. Chisholm, "Extracting structured data from invoices," in *Proc. Australasian Language Technology Association Workshop*, 2018, pp. 53–59.

[6]. X. Liu, F. Gao, Q. Zhang, and H. Zhao, "Graph convolution for multimodal information extraction from visually rich documents," in *Proc. NAACL-HLT*, 2019, pp. 32–39.

[7]. F. Krieger, P. Drews, and B. Funk, "Automated invoice processing: Machine learning-based information extraction for long tail suppliers," *Intelligent Systems with Applications*, vol. 20, 2023.

[8]. V. Perot, M. Rusinol, and D. Karatzas, "LMDX: Language model-based document information extraction and localization," in *Findings of ACL*, 2024, pp. 15140–15168.

[9]. Y. Xu, Y. Lv, M. Cui, et al., "LayoutLMv2: Multi-modal pre-training for visually-rich document understanding," in *Proc. ACL*, 2021, pp. 2579–2591.

[10]. S. Appalaraju, B. D. Trainor, M. Jain, et al., "DocFormer: End-to-end transformer for document understanding," in *Proc. IEEE/CVF ICCV*, 2021.

[11]. Z. Huang, Y. Chen, J. Li, and J. Zhou, "ICDAR2019 competition on scanned receipt OCR and information extraction," in *Proc. ICDAR*, 2019, pp. 1516–1520.

[12]. T. A. N. Dang and D. N. Thanh, "End-to-end information extraction by character-level embedding and multi-stage attentional U-Net," in *Proc. British Machine Vision Conference (BMVC)*, 2019.

[13]. Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of text and layout for document image understanding," in *Proc. AAAI*, 2020, pp. 11993–12000.

[14]. V. P. d'Andecy, E. Hartmann, and M. Rusinol, "Field extraction by hybrid incremental and a-priori structural templates," in *Proc. Int. Workshop on Document Analysis Systems (DAS)*, 2018, pp. 251–256.

[15]. T. I. Denk and C. Reisswig, "BERTgrid: Contextualized embedding for 2D document representation and understanding," in *Proc. NeurIPS Workshop on Document Intelligence*, 2019.