International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# Automated Information Extraction from Offer Letters Using Deep Learning: A Comparative Study

**Ramneet Singh Chadha[1], Chandrashekhar Aazad[2], Jasmehar Singh[3]**

[1]Scientist F, Embedded Systems, Centre for Development of Advanced Computing (C-DAC), Noida, Uttar Pradesh, India
[2]Project Engineer, Embedded Systems, Centre for Development of Advanced Computing (C-DAC), Noida, Uttar Pradesh, India
[3]Student, Shiv Nadar University,Noida, Uttar Pradesh, India

**Abstract:**
The fast move to digital for HR tasks has led to the creation of a lot of unstructured documents, especially offer letters, which provide important information about employees. Using usual approaches, this has not been processed well for extracting information. This study presents a comprehensive system for the automatic extraction of information from offer letters utilizing optical character recognition (OCR) and named entity recognition.
A comparison investigation has been conducted about the performance of the conventional Conditional Random Field (CRF) model and the advanced Bidirectional Encoder Representations from Transformers (BERT) model. This paper presents a system for generating synthetic data to address the challenges of data scarcity and privacy. This framework may create fake data for the process of getting information from offer letters, such as alternative layouts, semantic content, and noise. This paper evaluated the efficacy of the models on a demanding dataset of 500 documents. The findings indicate that, while the traditional Conditional Random Field model yielded satisfactory outcomes, the advanced Bidirectional Encoder Representations from Transformers model produced significantly superior results in comparison to the conventional model.
This experimental evaluation also proves the efficiency of the proposed system in extracting key entities such as candidate name, designation, salary, and date of joining accurately from the scanned offer letters. The comparative analysis proves that, although the proposed system using the CRF model works accurately in structured data, the BERT model has a better capacity for understanding the contextual relationship within the document text. Thus, the proposed transformer model proves to be effective in dealing with the variability in the document layouts, linguistic expressions, and document conditions.

**Keywords:** Document Information Extraction, Named Entity Recognition, OCR, CRF, BERT, Synthetic Data

## 1. INTRODUCTION

Offer letters are business documents that are usually semi-structured. These documents formalize the link between an applicant and an organization with regard to a given job. Though these documents are similar

in that they all serve a similar purpose for businesses, there is a lot of variation in structure, layout, design, and wording. It is possible to note variations in the design of the header and positioning of the entity data, wage breakdowns, legal limits, and styles used. This means that information such as applicant name, job title, wages, date of joining (DOJ), firm name, and location is available in unstructured data rather than structured data.

One of the most important parts of the existing Human Resource (HR) management systems is the automation of the information retrieval from the offer letters. More and more companies are trying to use technology for employee onboarding, payroll processing, background checks, and monitoring of compliance. It is a very tedious and labor-intensive job to manually extract information from the offer letters, and it is very easy to make errors, especially when dealing with a large number of papers. Parsing based on rules is not only unreliable but is also very difficult due to the differences in the format of the documents in different companies.

The standard methods for rule-based extraction use location heuristics and regex, which are manually crafted. These methods perform well for templates where the layout is fixed in stone, but they fail miserably if the layout of the document is altered in some way, such as if the wording is changed. Learning token-level dependencies, statistical labeling techniques such as Conditional Random Fields (CRF) [1], make for a more robust solution. However, these models are highly dependent on feature engineering and mostly use local context,making them less effective in dealing with semantic ambiguity and distractor entities.

The recent developments in deep learning, particularly in models that use transformers like **Bidirectional Encoder Representations from Transformers (BERT**)[2], have proven to be very effective in Natural Language Processing (NLP) tasks. This is due to the fact that they can use self-attention mechanisms and word representations to identify relationships in text. This is why they are particularly useful in data extraction. This is because there can be so many different things in one text. Moreover, while there has been a lot of development in research on how to interpret documents, the field of extracting offer letters for human resources has not been well studied. Not enough data is available in this field, and this is one of the biggest problems. In fact, real offer letters include private information about the individual and the company. It is not possible to obtain large datasets that are annotated due to privacy reasons and to follow the laws. This calls for the establishment of a training system that respects individual privacy as a requirement.

In this research, the above challenges are addressed by introducing a synthetic data generation system capable of producing realistic offer letters with diverse layouts, linguistic variations, and injected visual noise. The generated dataset is integrated into a complete processing pipeline consisting of image preprocessing, Optical Character Recognition (OCR), and Named Entity Recognition (NER). Furthermore, a comparative evaluation is conducted between a conventional Conditional Random Field (CRF) model and a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model in order to analyze their effectiveness under controlled yet challenging document conditions.

The primary contributions of this paper are summarized as follows:

- A privacy-preserving synthetic data generation engine is designed to simulate real-world offer letter variability, including template diversity, semantic distractors, and visual noise.
- A complete end-to-end information extraction pipeline integrating OCR and NER is developed for processing scanned offer letter documents.
- A detailed comparative evaluation between CRF and BERT models is performed, highlighting trade-

offs in performance, robustness, and interpretability.

The remainder of this paper is organized as follows. Section 2 reviews related work in document information extraction and sequence labeling. Section 3 formalizes the problem definition. Section 4 describes the proposed methodology, including synthetic data generation and model design. Section 5 presents the experimental setup and evaluation metrics. Section 6 discusses results and comparative analysis. Finally, Sections 7-9 outline limitations, future work, and conclusions.

## 2. RELATED WORK

In the domains of natural language processing and document understanding, one of the most important areas of attention has been the extraction of information from documents. The majority of the initial systems relied on rules that were established manually, regular expressions, and template matching approaches. In the case of highly ordered papers with predetermined patterns, such as standardized forms or bills, these techniques proved to be effective. However, rule based systems have a limited capacity for generalization, and therefore require continuous manual revisions whenever the formats of documents undergo changes.

The limitations of rule-based systems prompted the development of probabilistic sequence labeling techniques, such as Conditional Random Fields (CRF)[1], which were developed to solve these shortcomings. The conditional probability of label sequences is represented by CRFs, which are based on the tokens that have been observed. CRFs also allow for dependency between neighboring labels. Because of its ability to incorporate a wide variety of lexical and contextual information, this framework has garnered a lot of support for solving Named Entity Recognition (NER) difficulties. CRF models are substantially dependent on manually built feature engineering and primarily collect local contextual data, despite the fact that they are more resilient than rule-based techniques.

As the neural networks improved, deep learning architectures started to emerge in the place of the statistical models that had been trained using features. Improved performance was achieved through the application of Recurrent Neural Networks (RNN), more especially Bidirectional Long Short-Term Memory (BiLSTM)[2] networks coupled with Conditional Random Field (CRF) layers, denoted as BiLSTM-CRF. These improvements were achieved through the autonomous learning of distributed word representations. However, despite the improvements, the sequential processing of RNN-based models faces some limitations in terms of parallel processing and the challenges that may be encountered in the processing of long-range interactions.

The introduction of Transformer architectures marked a significant and revolutionary advancement in the field of natural language processing. In order to comprehend the global contextual connections that exist between tokens, transformer models, particularly BERT, make use of self-attention mechanisms. By applying rigorous pretraining and fine-tuning approaches, BERT was able to achieve better performance across a variety of natural language processing benchmarks, including natural language understanding (NER). Because of its ability to imitate bidirectional context, it is particularly well-suited for information extraction tasks at the document level that are characterized by semantic ambiguity.

Researchers in the field of document interpretation have improved text-only models by incorporating spatial layout information into their models. Textual and positional embeddings are incorporated into models such as LayoutLM, LayoutLMv2, and LayoutLMv3 in order to efficiently interpret visually complex documents such as contracts, receipts, and forms[4,5,6]. The performance of these models is significantly improved in circumstances when spatial alignment and document structure are of critical

importance. On the other hand, they require bounding box annotations and higher processing resources, both of which may not always be feasible in situations where resources are restricted or when privacy is a concern.

In the field of document artificial intelligence, the fabrication of synthetic data has emerged as an essential technique for protecting users' privacy during machine learning. It is possible to facilitate prolonged training while protecting sensitive information through the fabrication of real artificial papers. Recent studies have shown that synthetic datasets that have been rigorously built can closely mimic distributions that occur in the real world if suitable variability and noise injection techniques are incorporated into the construction of the datasets.

The extraction of information from documents has received a significant amount of research; however, there is less attention paid to documents that are HR-centric, such as offer letters. These papers present a number of specific challenges, such as semantic diversions (for example, a variety of dates and financial data) and templates for businesses that are not consistent with one another. The outcomes of our research contribute to the advancement of this discipline by combining the generation of synthetic data with a methodical evaluation of statistical and transformer-based neural network models under regulated yet difficult conditions.

## 3. PROBLEM DEFINITION

Given a scanned offer letter image I, the objective is to extract a structured set of entities E = {e1, e2, ..., ek}, where each entity corresponds to predefined fields such as Name, Designation, Salary, Date of Joining, Company, and Location. The document image I is first processed using an OCR engine, producing a token sequence:

$$T = \{t_1, t_2, \ldots, t_n\} \quad (1)$$

where:

- **T** represents the token sequence extracted from the document
- $\mathbf{t_i}$ represents the i-th token in the sequence
- **n** denotes the total number of tokens in the document

Each token must be assigned a label indicating whether it belongs to a particular entity class. The labeling scheme used in this work follows the BIO tagging format, which is widely used in Named Entity Recognition tasks.

$$Y = \{B-X, I-X, O\} \quad (2)$$

where:

- B-X indicates the beginning of an entity of type X
- I-X indicates that the token is inside an entity of type X
- O indicates that the token does not belong to any entity

The goal is to learn a function:

$$f: T \rightarrow Y \quad (3)$$

This function predicts the appropriate entity label for each token in the sequence. that maximizes the conditional probability

$$P(Y|T) \quad (4)$$

P(Y | T) represents the probability of predicting the correct label sequence Y given the input token sequence T.

For CRF, this is modeled using conditional probability distributions over label sequences [1]. For BERT, contextual embeddings are learned via transformer encoders [2].

## 4.PROPOSED METHODOLOGY

A. System Architecture

As shown in Fig. 1, the overall process involves image preprocessing, OCR, NER, and post-processing.

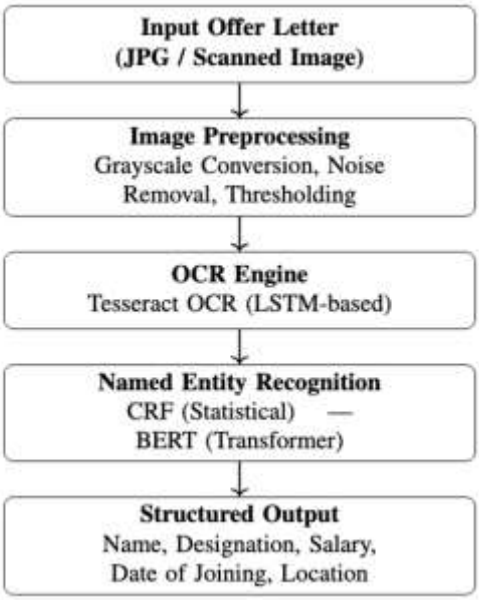


**Fig. 1. End-to-end architecture for automated offer letter information extraction**

### B. Synthetic Data Generation

Due to privacy constraints, real offer letters cannot be used at scale. A Python-based synthetic data engine was developed using the Faker library. Five distinct templates were created, reflecting real-world corporate layouts. Linguistic variation, visual noise (Gaussian blur, contrast reduction, skew), and semantic distractors were injected to simulate realistic scanning conditions.

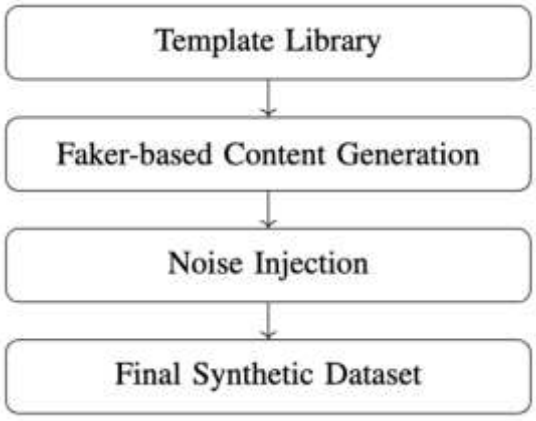


**Fig. 2. Synthetic data generation workflow**

### C. Preprocessing and OCR

All the images were converted to grayscale and the OCR was performed using Tesseract 4.0 LSTM-based OCR [7,8]. Preprocessing improved character recognition accuracy, particularly under noisy conditions.

### D. Named Entity Recognition

The following two models were tested:

- CRF: Implemented with the help of the sklearn-crfsuite library with various hand-crafted lexical as well as contextual features.
- BERT: The bert-base-cased model was fine-tuned with the AdamW optimizer.

## 5. EXPERIMENTAL SETUP

The BERT model has been fine-tuned with the AdamW optimizer, with a learning rate set to $5*10^{\wedge-5}$, for 3 epochs. The batch size is set to 16, with a maximum sequence length set to 128 tokens. The training is performed on a CPU-based environment.

### A. Dataset

TABLE I
DATASET STATISTICS

| Property | Value |
|---|---|
| Total Documents | 500 |
| Training Set | 400 |
| Test Set | 100 |
| Target Entities | 6 |

### B. Evaluation Metrics

Precision, Recall, and F1-score were computed using strict exact-match criteria at the token level.

## 6. RESULTS

The comparison chart of CRF and BERT models is given in Table~II, and a visual representation is given in Fig.~3. The results show that both models perform with high accuracy on the information extraction problem in the offer letter, but the transformer-based BERT model performs better on all evaluation criteria compared to the traditional CRF model.

The CRF model has a precision of 0.9850, recall of 0.9603, and F1-score of 0.9725, as given in Table~II. The results show that both models perform with high accuracy on extracting information, but the recall value shows that sometimes the model is not able to recognize entities, especially when entities are used in unconventional contexts.

TABLE II
MODEL PERFORMANCE COMPARISON

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| CRF | 0.9850 | 0.9603 | 0.9725 |
| BERT | 0.9916 | 0.9916 | 0.9916 |

The BERT model has a precision, recall, and F1-score of 0.9916, showing that this model is performing better on all criteria. The improvement in recall shows that this model is performing better on extracting entities even when entities are used in unconventional contexts.

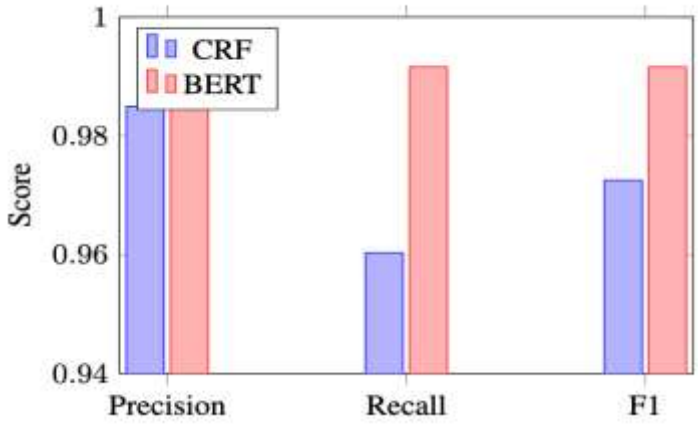


**Fig. 3. Comparison of Precision, Recall, and F1-score for CRF and BERT models.**

Overall, these findings confirm the superiority of deep context models in terms of robustness for document-level information extraction tasks. While the CRF model is still computationally lightweight and interpretable, the BERT-based approach offers advantages in terms of accuracy and generalization when dealing with semi-structured documents such as HR offer letters.

## 7. ERROR ANALYSIS

The CRF errors were predominantly observed in cases with high distractor density and non-traditional entity placement. BERT reported impressive robustness with minor degradation in extreme cases of OCR noise.

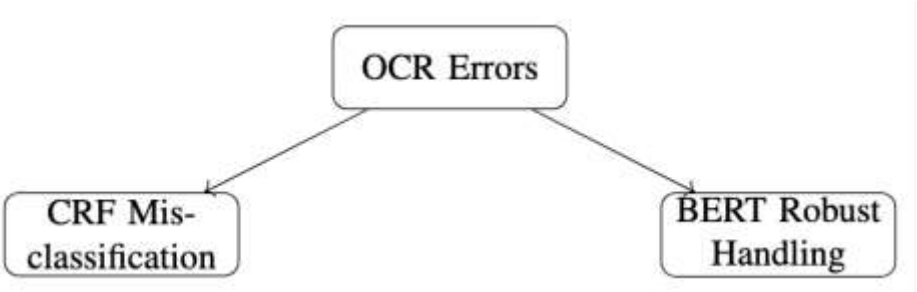


**Fig. 4. Primary sources of extraction errors**

## 8. DISCUSSION

The experimental results indicate an important trade-off between model interpretability, efficiency, and performance. The Conditional Random Field (CRF) model displays robust efficacy in cases where entities follow predictable language patterns and appear in anticipated situations. CRF is highly interpretable and transparent since it is based on handcrafted features such as capitalization, prefixes, suffixes, and adjacent token features. This feature makes CRF an attractive model in situations where resources are limited.

The CRF model has been seen to face constraints when dealing with complex document structure, semantic ambiguities, and a high concentration of distractors. The CRF model has been found to face difficulties in identifying the specific entity (i.e., Date of Joining), as seen in various offer letters containing different financial values or dates, because of irrelevant but syntactically similar distractors. The reason behind this is that the CRF model is not able to understand long-range semantic links.

The performance of the proposed model based on BERT has been consistently better than CRF for all objects. The reason for this is that BERT leverages bidirectional [9] self-attention for obtaining contextual token representations that capture both local and global knowledge. This allows for accurate disambiguation of objects based on sentence-level semantics rather than surface properties. The proposed model has shown better robustness against heterogeneity of layouts, paraphrase, and semantic noise.
Despite the increased computational cost, the BERT model is nevertheless feasible in practice because of the relatively small sizes of offer letter documents and the possibility of CPU-based inference. The results indicate that the transformer model is more suitable for use in an enterprise-level document comprehension system that emphasizes accuracy and robustness over interpretability.

## 9. LIMITATIONS

While the proposed technique promises the achievement of high accuracy in the process of extraction, certain constraints must be pointed out. In the first place, the data set that is used in this study is artificially created. In spite of the significant attempts that were made in replicating realistic variations using multiple templates, language variations, and visual noise, it is possible that the artificially created data may not entirely represent all the boundary cases that are possible in actual business offer letters. For example, cases involving handwritten text, low-quality images captured using smartphone cameras, and heavily stylized company branding were not directly addressed.
Secondly, the proposed technique is heavily dependent on the quality of the output provided by the OCR engine. This is a significant limitation. Errors that may be introduced during the OCR engine may have an adverse effect on the NER process. In spite of the significant attempts that were made in addressing this issue during the preprocessing stage, the accuracy of the OCR engine is a significant limitation.
The existing methodology considers the documents as linear text sequences, and the two-dimensional layout information is not explicitly incorporated. As a result, the depiction of the items that are placed in the tables or multicolumn formats may not be fully achieved. This constraint particularly affects the wage delineations that are depicted in the tabular form. The assessment of the offer letters is only concerned with the documents in the English language. The effectiveness of the system in dealing with the code-mixed or bilingual texts has not been investigated, which forms a constraint in the implementation in internationally dispersed organizations.

## 10.FUTURE WORK

This analysis reveals multiple intriguing avenues for research. An inherent progression of the proposed system is the incorporation of layout-aware transformer models like LayoutLM or DocFormer, which concurrently analyse textual content and the spatial arrangement of documents. The integration of bounding box information can substantially enhance extraction efficacy for visually intricate documents, especially those featuring tables and complicated formatting.
A major focus is placed on multilingual and cross-lingual information extraction. Future studies will focus on the fine-tuning of the multilingual transformer model, allowing it to handle information written in languages other than English, as well as documents that contain information in multiple languages. In addition, active learning strategies may be leveraged, allowing the system to improve itself by learning from documents with minimal annotation effort, while at the same time ensuring that privacy laws are adhered to by requesting input from the human annotator when the system is not confident about the prediction.

Subsequently, confidence estimate and uncertainty modeling for the extracted items will be investigated in the forthcoming research. Adding confidence information for the extracted variables can contribute to the stability of the system and smooth integration into the companies' processes.

## 11. CONCLUSION

In this research, an comprehensive framework for automatic information extraction from offer letters using OCR and Named Entity Recognition methodologies was proposed and implemented[11]. The proposed approach effectively deals with the challenges of document diversity, data scarcity, and privacy preservation through an efficient synthetic data generation strategy.

The extensive investigation showed that, although Conditional Random Fields provide a lightweight and interpretable baseline, transformer-based models such as BERT show better performance and resistance to semantic distractors and layout variations[12]. The almost state-of-the-art results achieved by the BERT model validate the effectiveness of transfer learning for document-level information extraction in sensitive sectors of enterprises.

The findings of the study add to the existing body of research in the field of document interpretation, and the proposed approach offers valuable insights that may be useful in the implementation of intelligent HR automation systems. The proposed approach offers an accurate, scalable, and privacy-sensitive solution that may be easily generalized to other business documents beyond offer letters.

## REFERENCES

1. J. Lafferty, A. McCallum, and F. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” in Proceedings of the 18th International Conference on Machine Learning (ICML), 2001, pp. 282–289.
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
3. A. Vaswani et al., “Attention Is All You Need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
4. X. Liu, Y. Xu, S. Xu, and J. Gao, “LayoutLM: Pre-training of Text and Layout for Document Image Understanding,” in Proceedings of KDD, 2019, pp. 1192–1200.
5. Y. Xu et al., “LayoutLMv2: Multi-modal Pre-training for Visually- Rich Document Understanding,” in Proceedings of ACL, 2021, pp. 2579–2591.
6. Y. Huang et al., “LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking,” in Proceedings of ACM MM, 2022, pp. 4083–4091.
7. R. Smith, “An Overview of the Tesseract OCR Engine,” in Proceedings of ICDAR, 2007, pp. 629–633.
8. T. Breuel et al., “High Performance OCR for Printed English and Fraktur Using LSTM Networks,” in Proceedings of ICDAR, 2013, pp. 683–687.
9. M. Peters et al., “Deep Contextualized Word Representations,” in Proceedings of NAACL, 2018, pp. 2227–2237.
10. J. Howard and S. Ruder, “Universal Language Model Fine-tuning for Text Classification,” in Proceedings of ACL, 2018, pp. 328–339.
11. M. Appalaraju et al., “DocFormer: End-to-End Transformer for Document Understanding,” in Proceedings of ICCV, 2021, pp. 993–1003.

12. D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson, 2023.