IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 25, 2021, accepted July 8, 2021, date of publication July 12, 2021, date of current version July 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3096739

Multi-Layout Unstructured Invoice Documents Dataset: A Dataset for Template-Free Invoice Processing and Its Evaluation Using AI Approaches

DIPALI BAVISKAR[1], SWATI AHIRRAO[1], AND KETAN KOTECHA[2]

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India
[2]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India

Corresponding authors: Swati Ahirrao (swatia@sitpune.edu.in) and Ketan Kotecha (head@scaai.siu.edu.in)

This work was supported by the Research Support Fund of Symbiosis International (Deemed University).

ABSTRACT The daily transaction of an organization generates a vast amount of unstructured data such as invoices and purchase orders. Managing and analyzing unstructured data is a costly affair for the organization. Unstructured data has a wealth of hidden valuable information. Extracting such insights automatically from unstructured documents can significantly increase the productivity of an organization. Thus, there is a huge demand to develop a tool that can automate the extraction of key fields from unstructured documents. Researchers have used different approaches for extracting key fields, but the lack of annotated and high-quality datasets is the biggest challenge. Existing work in this area has used standard and custom datasets for extracting key fields from unstructured documents. Still, the existing datasets face some serious challenges, such as poor-quality images, domain-related datasets, and a lack of data validation approaches to evaluate data quality. This work highlights the detailed process flow for end-to-end key fields extraction from unstructured documents. This work presents a high-quality, multi-layout unstructured invoice documents dataset assessed with a statistical data validation technique. The proposed multi-layout unstructured invoice documents dataset is highly diverse in invoice layouts to generalize key field extraction tasks for unstructured documents. The proposed multi-layout unstructured invoice documents dataset is evaluated with various feature extraction techniques such as Glove, Word2Vec, FastText, and AI approaches such as BiLSTM and BiLSTM-CRF. We also present the comparative analysis of feature extraction techniques and AI approaches on the proposed multi-layout unstructured invoice document dataset. We attained the best results with BiLSTM-CRF model.

INDEX TERMS Artificial Intelligence (AI), information extraction, key field extraction, named entity recognition (NER), template-free invoice processing, unstructured data.

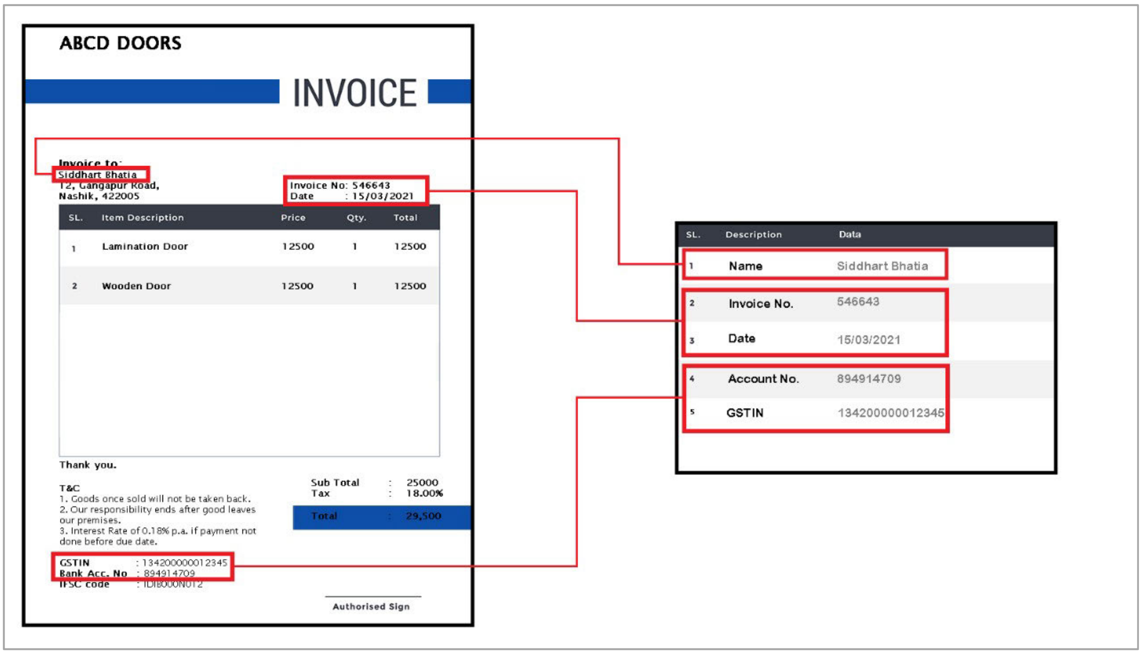
I. INTRODUCTION

Unstructured data is currently a buzzword in organizations. However, the efficient analysis of unstructured documents is challenging for almost 95% of business enterprises as unstructured data is not organized very well. It costs millions of dollars to analyze and integrate this data into the information management systems of the organization for a significant increase in productivity [1].

According to Forbes statistics [2], recently, unstructured data occupies 80% of generated data produced in email, official communication letters, scanned PDF, and many others. This unstructured data is growing exponentially with the frequent access to digital communication. Lots of insights or valuable information can be dwelled out of such data that would be utilized to make the right decisions on market situation or product.

The business process workflow of the organization consists of various template-based unstructured documents such as invoices, receipts, bills, and many others [3]. At present,

The associate editor coordinating the review of this manuscript and approving it for publication was Ting Wang.



ABCD DOORS

INVOICE

Invoice to:
Siddhart Bhatia
12, Gangapur Road,
Nashik, 422005

Invoice No: 546643
Date : 15/03/2021

SL.	Item Description	Price	Qty.	Total
1	Lamination Door	12500	1	12500
2	Wooden Door	12500	1	12500

Thank you.

T&C
1. Goods once sold will not be taken back.
2. Our responsibility ends after good leaves our premises.
3. Interest Rate of 0.18% p.a. if payment not done before due date.

Sub Total	:	25000
Tax	:	18.00%
Total	:	29,500

GSTIN : 134200000012345
Bank Acc. No : 894914709
IFSC code : IDIB000N012

Authorised Sign

SL.	Description	Data
1	Name	Siddhart Bhatia
2	Invoice No.	546643
3	Date	15/03/2021
4	Account No.	894914709
5	GSTIN	134200000012345

FIGURE 1. Sample Key Fields Extraction from Unstructured Documents such as Invoice Document.

handling these unstructured documents is generally a manual task. The existing frameworks for automatic key fields extraction tasks are expensive and error-prone. Consider the unstructured document type like invoices, which is of diverse types like invoices from the different suppliers or various department offices inside an organization, may have a unique invoice layout [4]. However, an invoice consists of few common structured data or information fields like the name of the supplier, supplier address, date of invoice, number of the invoice, total invoice amount, Goods & Service Tax number (GST number), and the item list. An automatic key fields extraction framework that can extract all this valuable data can significantly improve the productivity of the organization by reducing error-prone and laborious manual work.

Figure 1. depicts the example of key fields extraction from the unstructured documents such as the invoice document image.

The automatic extraction of the key fields from the unstructured documents can significantly relieve human efforts to process the piles of unstructured documents to get insightful data. Hence, developing a framework for efficient key field extraction from multi-layout unstructured documents is an urgent necessity [5]. However, key field extraction from multi-layout unstructured documents research faces few challenges discussed as follows:

A. CHALLENGE IN EXTRACTING INFORMATION FROM UNSTRUCTURED DOCUMENTS

The automatic and efficient key field extraction task is one of the challenging tasks as its solution is spanned across the use of Computer Vision (CV) and Natural Language Processing (NLP) [6]. The unstructured documents such as invoices, claim processing forms usually do not comprise "natural language" as other regular documents or paragraphs. Hence, they can not be treated as standard NLP tasks. The unstructured document has data presented inside the tables spanning multiple pages, with a variable number of sections and various layouts to organize the information. Text layout understanding is a fundamental solution to understand such varied unstructured documents. Considering the information extraction task as an image segmentation problem loses the text semantics, which can further complicate the processing of unstructured documents. The challenges in this research are stated as:

1. Availability of standard, high-quality, multi-layout unstructured documents datasets.
2. Inadequate data validation techniques for assessing the data quality.
3. Lack of a more precise and accurate automatic key field extraction framework for multi-layout unstructured documents dataset.
4. Lack of publicly available tools for automating the task of extracting key fields efficiently from the multi-layout unstructured documents.

Following are the main contributions of the proposed work:

1. Detailed architecture for the development of multi-layout unstructured invoice document dataset collected from the different suppliers of the organization.
2. Development of a high-quality and multi-layout unstructured invoice document dataset that includes complex and varied formats.
3. To employ the data quality assessment with a statistical data validation technique.

4. Development of a proposed framework for extracting key fields automatically from multi-layout unstructured invoice documents to create a structured output, using Artificial Intelligence (AI) approaches.
5. Exploring appropriate feature extraction techniques and AI approaches used to evaluate the proposed multi-layout unstructured invoice documents dataset.

B. PRACTICAL/INDUSTRY IMPLICATIONS OF PROPOSED WORK

The proposed work has numerous practical and industry significance for implementing key field extraction automation in the finance sector. Our research specifies that even though several other sectors such as healthcare and legal have started taking the benefits of automating key field extraction, further enhancements are essential to accomplish automatic key field extraction from multifaceted and diverse unstructured invoice documents obtained from various suppliers. Each invoice has unique layouts that can not be easily processed with existing automation tools which organizations currently have with them. The main reason behind them is their complexity and variety in which location or placing of a particular key field varies immensely as invoice format changes. In today's scenario, to the best of our knowledge, the dataset containing such varied, highly diverse, and high-quality invoice documents is scarce. If such a dataset gets available to the researchers, research in this area will advance. Various organizations can collaborate with researchers and can significantly become a pioneer by implementing end-to-end automation solutions. Hence, there are benefits of automation in key field extraction tasks in the finance sector or few other sectors, but organizations have some substantial challenges to deal with varied and multifaceted unstructured documents.

II. RELATED WORK

This section discusses the existing work in automatic key fields extraction tasks from unstructured documents. Early approaches used for recognizing and extracting the named entities (Named Entity Recognition) (NER) have used various domain-specific standard and custom datasets, which are elaborated further in this section.

A. STANDARD DATASETS

Existing studies can be categorized based on the publicly available/standard datasets for different domains and custom datasets. Few noteworthy publicly available/standard datasets found in the literature are discussed as follows:

Availability of public datasets for unstructured business documents such as invoices, legal contracts, and claim processing forms is scarce in this research area.

The study [7] presented a Form Understanding (FUNSD) dataset containing 199 completely annotated noisy scanned forms in JSON format. FUNSD consists of forms with diverse fields like Marketing, Advertisement, Science, and few others, which are used for text detection, Optical Character Recognition (OCR), and document layout understanding tasks. The number of samples in this dataset is very less, leading to poor generalizability of form understanding application.

The study [8] uses a dataset called RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) comprising scanned document images for 16 diverse categories and 25,000 document images per category, such as Letters, Memos, Resumes, and Invoices, and few others. It is a subcategory of IIT-CDIP Test Dataset Collection 1.0 containing very low resolution, low quality, and noisy document images. This dataset is used for document classification, layout understanding, and OCR with a bounding box in XML file format.

The ICDAR (International Conference on Document Analysis and Recognition) organizes various competitions for researchers in the document recognition and analysis domain. Various challenges for the competition include complex layout document recognition, text extraction, handwritten text recognition, and few others. The study [9] carried experiments on ICDAR provided SROIE (Scanned Receipts OCR and Information Extraction) dataset for extracting receipt fields. All 1000 scanned and annotated receipts in this dataset have a similar layout structure. However, several receipts in this dataset have a poor scan and low-resolution quality. The receipts also have folded and distorted corners and varied text sizes.

The study [10] developed the annotated dataset named PubLayNet, consisting of more than 360 thousand document images for the document analysis tasks. This dataset is derived from PDF articles of the PubMed Central dataset. This dataset is also available in an XML representation. It is used for metadata extraction tasks like metadata on article author and title. However, poorly annotated PDF articles in this dataset contribute to noise in the dataset and result in poor performance.

The study [11] developed token-level annotations on a scientific articles dataset named DocBank, used for the article metadata extraction and document layout analysis tasks. It consists of 500 thousand document pages with the label annotations such as Abstract, Author, Caption, Figure, and few others.

The study [12] created a GROTOAP2 dataset of the scientific articles in PDF and Truewiz (XML) format. This dataset is derived from an open-access dataset called PubMed Central. It comprises of 13,210 ground truth, hierarchical structure scientific articles, signifying different zones of the publication contents such as full text, different sections, title, authors, and few other zones.

The study [13], [14] presented a dataset called PRImA (Pattern Recognition and Image Analysis) consisting of a variety of diverse documents to handle various challenges of document layout analysis. It includes a variety of magazine and technical/scientific journal document images. It is specifically used in the scientific literature domain.

The study [15], [16] uses datasets MIMIC (Medical Information Mart for Intensive Care) and i2b2 (Informatics for Integrating Biology and the Bedside) medical domain datasets for information extraction tasks from the clinical notes documents of the patient. It is a medical domain dataset.

From the above discussion, we conclude that the standard datasets face a few of the following challenges:

- Standard datasets are domain-specific or task-specific, and language-specific.
- Standard datasets are outdated, containing obsolete formats of the unstructured documents.
- Standard datasets are of poor quality, including noisy, low-resolution document images.
- There is a lack of data validation techniques for the data quality assessment of the standard datasets.
- There is a lack of a large, annotated corpus for the information extraction tasks.

These are few issues due to which the researchers have used custom datasets. But, custom datasets also face few challenges discussed as follows:

- Custom datasets have privacy issues. Unstructured documents such as invoices, office orders are private and confidential documents of the organization. Hence, public access to such documents is scarce.
- Manual annotations of unstructured documents are time-consuming, error-prone, and tedious tasks.
- Getting high-quality, large and enough documents for model training is another challenge in creating the custom dataset.

Table 1. summarizes the few other standard and custom datasets mentioned in the existing literature studies.

B. CHALLENGES WITH THE EXISTING DATASETS

The information extraction model must be trained with good quality data to obtain meaningful and valuable information extraction. Figure 2. illustrates the challenges with the existing datasets, which are discussed as follows:

1) PRIVACY ISSUES

- *Confidentiality [1]:* Most of the custom datasets that the researchers have used, contain confidential data about people, administration, or business enterprises. For instance, invoices are private documents for any company or organization, containing information about the Buyer Name, GST Number, and few other fields. These fields are considered as the confidential information for the organization. Hence, the availability of the datasets of such documents is a challenging issue.
- *Limited Publicly Available Datasets [24], [19]:* Various workshops/conferences provide the customized datasets to their participants. The organizers often provide training and test data to the participants before the competition. These datasets can be accessed publicly by registering separately for the same. For instance, ICDAR provides such types of data for various document understanding tasks. Several researchers have highlighted their research work in ICDAR.

2) POOR QUALITY

- *Dataset Quality [25], [26]:* Dataset containing quality images is an important aspect since Machine Learning (ML) model accuracy highly relies on the training data quality. Many existing datasets have distorted or skewed images, that is, tilted or off-centered images. Blur images and the images captured in the various lighting background is another challenge in few existing datasets. Noisy images are the images with some watermark or patterns in the background, handwritten remarks. Noise may also get added while scanning documents with a low-resolution camera, or images scanned with a low-quality scanner. Skewed and noisy images lead to low-performance of the OCR engine. Missing data values and few other errors lead to an insignificant extraction of data.
- *Outdated Datasets [27]:* Outdated or obsolete datasets are another challenge to the existing datasets as most of these datasets contain document structure with an old format/layout. The datasets like RVL-CDIP contain document images with blur, obsolete format, and missing data values, creating problems during the information extraction process. The problems like blur images and differences or variability in lighting conditions affect the model performance.

3) LACK OF DATA VALIDATION/QUALITY ASSESSMENT TECHNIQUES [28]–[30]

Data validation techniques assure to assess the quality of the dataset. Data validation helps to determine whether the available data is suitable for ML model training. The existing literature does not focus on various data quality assessment techniques for categorical/nominal data types. There are very few data validation methods available and discussed in the literature, such as K-fold Cross-Validation [28], Chi-Square [29], and Cohen's Kappa [30].

4) DOMAIN-SPECIFIC DATASETS

- *Task-Specific Datasets [28]:* The existing datasets address the need for a specific domain such as the Scientific Literature or the Medical domain. The existing datasets are task-specific related to metadata extraction tasks, Article Title or Bibliography Extraction Tasks, or Patient Details Extraction Tasks.
- *Handwritten and Printed Datasets [23]:* Few existing datasets contain images with handwritten characters, and there are various types of handwriting, which makes it difficult to locate and recognize the text, leading to inaccurate results. The recognition of various handwritings requires advanced OCR techniques.
- *Language-Specific Datasets [31], [32]:* The unstructured documents datasets are presented in various native languages, such as the French receipts dataset, Chinese

TABLE 1. Summary of Standard and Custom-Built Datasets in the Existing Literature.

Reference	Dataset	Dataset Type	Domain	Tasks	Language	Data Collection Period	Dataset Feature	Validation Methods
[17]	MNIST (Modified National Institute of Standards and Technology database)	Standard	Handwritten Recognition	Handwritten digit recognition with OCR, Computer Vision, Deep Learning	English	1999	Ten-digit classes from 0 to 9	No validation
[17]	MIDV 500 (Mobile Identity Document Video dataset)	Standard	Generic	Document analysis and recognition	English, Chinese	Not given	17 Identity cards, 14 Passport copies, 13 Driving Licenses, and 6 other types of identity documents.	No validation
[18]	DocRED	Standard	Generic	Relation extraction from documents	English	2019	Named entities and relationships among entities	Human annotated-Manual validation
[19]	ICDAR-2019	Standard	Finance	Receipt recognition and key information extraction	English	2019	Receipt fields such as good name and price	No Validation
[20]	2003 French news articles	Custom	News articles	Document sentiment analysis	French	2019	French article document consisting average 4000 words	No validation
[21]	Legal contract documents	Custom	Legal	Contract entity extraction	English	Not given	3500 English contract documents with 11 different elements	No validation
[19]	4, 484 scanned and annotated Spanish receipt dataset	Custom	Finance	Receipt recognition and key information extraction	Spanish	2019	9 different labels: Vendor Name, Vendor Taxi ID, Invoice Date, and a few other	No Validation
[22]	CloudScan Invoice dataset	Custom	Finance	Invoice field extraction	English	2017	8 different labels: Invoice Number, Invoice Date, Invoice Currency Type, and few others	No Validation
[23]	National Archives Forms (NAF) dataset	Custom	Generic	Form field extraction in the form of key-value pair	English	Not given	Instances of handwritten NAF form	No Validation

Passport, Identity Cards, and Patient Medical Receipt Datasets. Ambiguous language and incorrect morphological phrases pose a serious challenge to the information extraction approaches.

- *The Absence of Labeling/Annotations [24]:* Unlabeled corpus or domain-specific annotations is a serious challenge. The absence of a huge labelled corpus in the public datasets is another critical challenge.

C. NAMED ENTITY RECOGNITION (NER)

NER is the most significant and crucial task in information extraction. NER identifies the significant entities or elements from a given text, such as person or organization name, address, invoice number, and many others. Managing, sorting, and analyzing the unstructured text data becomes easier by extracting key elements from them. It also helps to deal with large datasets by discovering

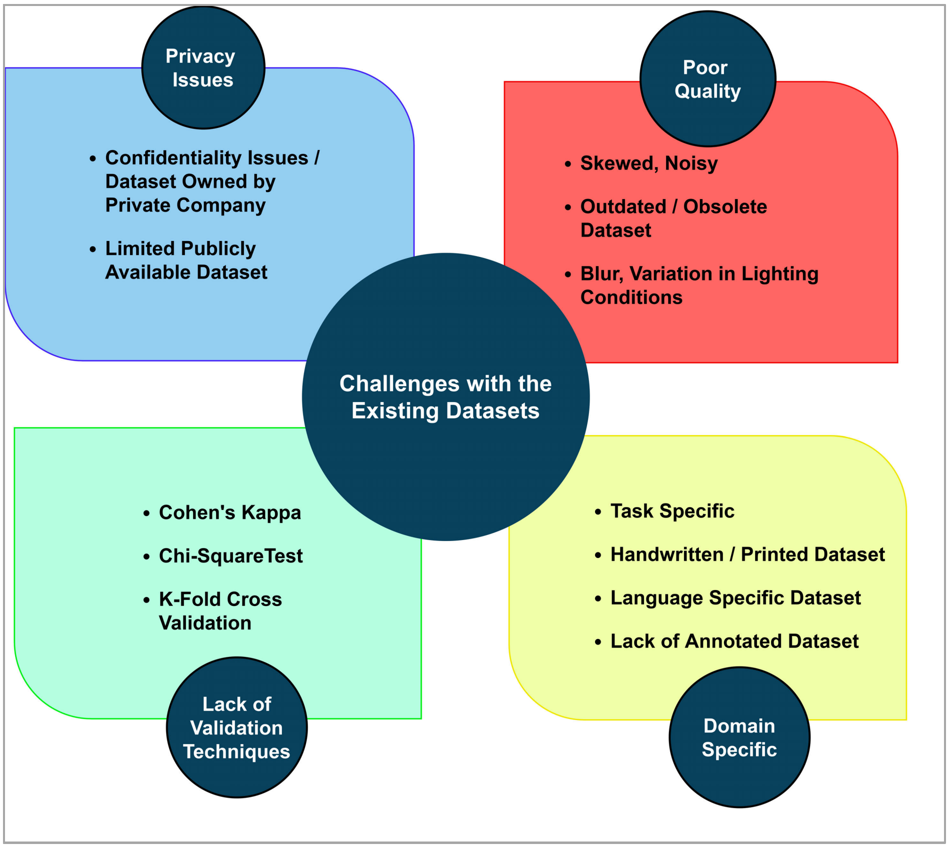

Privacy Issues
• Confidentiality Issues / Dataset Owned by Private Company
• Limited Publicly Available Dataset
Poor Quality
• Skewed, Noisy
• Outdated / Obsolete Dataset
• Blur, Variation in Lighting Conditions
Challenges with the Existing Datasets
• Cohen's Kappa
• Chi-SquareTest
• K-Fold Cross Validation
Lack of Validation Techniques
• Task Specific
• Handwritten / Printed Dataset
• Language Specific Dataset
• Lack of Annotated Dataset
Domain Specific

FIGURE 2. Challenges with the Existing Datasets.

crucial statistical information about the unstructured data.

Few studies for named entity identification and classification using various approaches found in the literature are discussed as follows:

The study [33] presented a conventional template-based approach for the identification and classification of invoice information. After a few pre-processing on invoice images like slant removal and rotation angle reduction, 20 invoice template images are used to match the regions of text blocks for extracting useful information from that block. However, template-matching methods have limitations in processing multiple invoice layouts. In a template-based method for pattern matching, several lines of code with regular expressions are needed, which is not suitable for highly diverse unstructured documents such as invoices and purchase orders.

Few studies [34]–[36] reported the use of conventional Machine Learning (ML) algorithms such as Naïve Bayes and Support Vector Machine (SVM). SVM classifier maximizes the text categorization or classification accuracy in the study [34]. The study [35] uses SVM and Naïve Bayes, and OCR for various invoice field categorizations like Invoice Date, Invoice Number. The study [36] uses SVM and Naïve Bayes for identifying Biomedical named entities from medical document abstracts.

The study [37] classified and studied the NER approaches in three types: Rule-based/ Knowledge-based approach, Feature-inferring Neural Network approach, and Learning-based approach [37]. Rule-based/Knowledge-based approaches are highly dependent on human-defined rules and Lexical-resource or Vocabulary of a particular domain or language [32]. Learning-based approaches rely on learning from the data provided and include Supervised, Semi-supervised and Unsupervised learning algorithms. SVM [35], Hidden Markov Model (HMM) [38], and Naïve Bayes [36] are the few examples of the approaches under this category. Feature-inferring Neural Network approaches use automatic feature extraction and learning capabilities [39].

A model can be built with traditional ML algorithms such as SVM and Naive Bayes for key field recognition and classification tasks. However, traditional ML algorithms face certain challenges, as mentioned below:

- Traditional ML algorithms can not be used for end-to-end information extraction tasks. They are limited to key field recognition and classification tasks.
- A large unlabelled and unstructured text dataset poses a challenge to the traditional ML algorithms.
- Manual feature extraction and dimensionality reduction are other challenges in using traditional ML algorithms.

- Complex layout and arbitrarily placed elements or entities within a document pose a great challenge to the traditional ML algorithms.

Few of the challenges mentioned above, can be solved using Deep Learning (DL) approaches [6]. Automatic feature extraction and availability of pre-trained Neural Networks (NN) trained on huge unlabeled corpus are the main advantages of using DL approaches in information extraction tasks. Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and their variants are precisely and popularly employed DL approach for NER. Few DL approaches proposed in the existing literature are briefly discussed below:

RNN is a popular NN for sequence tagging or labeling tasks in NER. Each of the ''tokens'' in a given input sentence or sequence is assigned with a specific label or class in sequence labeling. RNN feeds back the previous network output and current state input, allowing the network to capture or store the sequential relationship between the entities. The study [39] proposed RNN for recognizing Biomedical Named Entities (BioNER) and detecting the interaction amongst ''proteins and drugs'' or ''genes and diseases'' by identifying relations between Biological Entities.

The study [40] proposed the Deep NN approach in two stages: Text detection and Text recognition, by combining CNN and RNN to create the NN named Convolutional Recurrent Neural Networks (CRNN). It is used to extract the textual contents from the document images of patient's medical laboratory reports. This extracted information helps the doctor to keep track of the patient's health conditions and diagnosis.

The study [19] proposed a Convolutional Universal Text Information Extractor (CUTIE) approach that uses CNN to extract key information from the ICDAR-2019 receipts dataset and other self-built datasets. It uses the semantic information and positional information of entities to train CNN and the word embedding layer.

The study [23] presented a template-free form field extraction method on NAF historical handwritten filled form dataset, with a varied layout and noisy form images using Fully Convolutional Network (FCN). FCN is used along with a Heuristic Detector function for detecting the relationship between label-value pairs.

Another study [41] combined CNN and RNN for Chinese medical invoice recognition tasks. CNN is used for extracting image features from medical invoices. RNN is used for identifying semantic information from the extracted features.

Short-term memory is the main disadvantage of simple RNNs. So, a more complex NN architecture that solves the short-term memory issue in RNN is proposed in few studies. Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) are extensively used architectures to fulfill this requirement. LSTM and GRU are the variants of RNN consisting of gating and cell mechanisms for performing memory-associated processing flow. The study [28] used the MIMIC and i2b2 clinical detail datasets to extract patient's symptoms and disease details with LSTM and GRU.

Bidirectional LSTM (Bi-LSTM) is another variant of RNN comprising of the forward and backward layers. The forward layer holds the text input sequence, and the backward layer process the input in inverse order. The hidden state combines each input ''token'' in sequence and information associated with each token in the form of intermediate representation. Following this method, in each iteration, the Bi-LSTM network has information on the complete document, and then it infers the correct label for each ''token.'' Few studies [4], [20], [21] proposes Bi-LSTM and Conditional Random Field (CRF) for text sequence modelling, text classification and extraction tasks. The study [20] used Bi-LSTM and CNN for analyzing document sentiments into positive, negative, and neutral classes. In this study, French articles from newspapers are used as a dataset for sentiment analysis tasks. The study [4] presented the Bi-LSTM-CRF model for key information extraction from the ICDAR-2019 receipt dataset. Spatial and text semantic features combinely used to train the Bi-LSTM-CRF model. The study [21] used Bi-LSTM combined with CRF as a top layer for extracting the legal contract entities such as the agreement details from the legal document dataset.

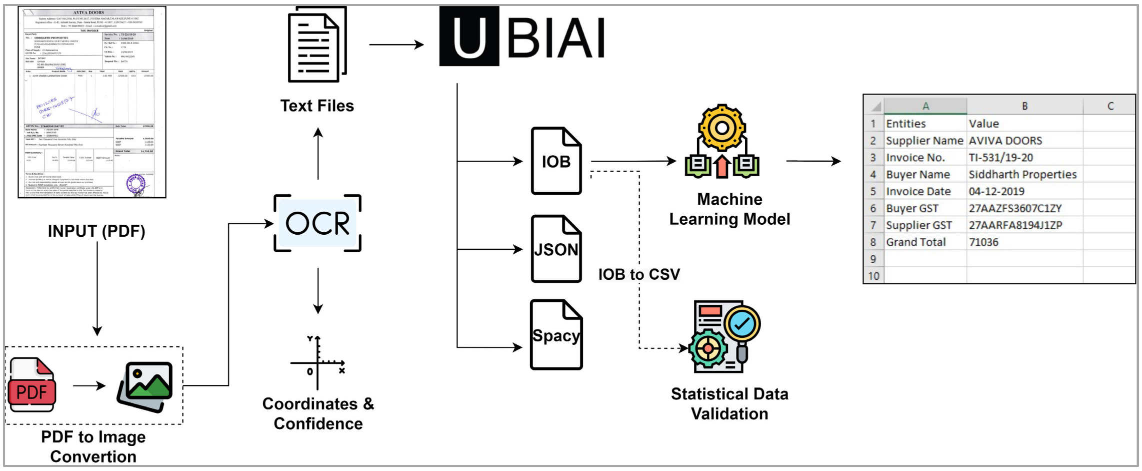
The study [22] proposes CloudScan-an invoice recognition and analysis system using Bi-LSTM for eight different fields of automatically generated invoice documents training dataset with the help of Universal Business Language (UBL) builder system. UBL is basically the XML format of the invoice.

Bidirectional Encoder Representations from Transformers (BERT) is another pre-trained DL unsupervised NLP model. After fine-tuning, BERT can efficiently perform many popular NLP tasks such as NER. BERT is significant, since it can process the data bidirectionally with a contextual understanding of data. BERT is a pre-trained network trained on a huge unlabeled corpus, so generally used for transfer-learning in NLP. BERT can easily deal with ambiguous languages. The study [18] proposed BERT for extracting relationships amongst entities at the document level from the DocRED dataset. Later, the BERT is fine-tuned with different parameters to improve the performance of the entity relation extraction task by 5% (F1 Score - 47%) than BERT without fine-tuning (F1 Score - 42 %). The study [42] proposed word and character level BERT models for NER tasks from various financial and biomedical documents. This study proposes transfer learning for 3000 invoice document datasets and biomedical document datasets (BioNLP09 and The BioCreative II Gene Mention (BC2GM)) for BERT training. BERT model is later fine-tuned for NER tasks on very few available labeled data.

To summarize, the existing literature focuses on key field extraction tasks on documents with similar layouts such as receipts. The literature lacks a dataset containing high-quality, multiple, or diverse layout documents. AI and NLP model accuracy mainly depend on sufficient and varied training data. Obtaining such varied and annotated data is problematic for many researchers and organizations due to

TABLE 2. Summary of Commonly used Deep Learning Techniques for NER Tasks in the Existing Literature.

Reference	Technique used	Hyperparameters	Feature Extraction method	Performance Metric	Remark
[39]	RNN, Bi-LSTM	Not given	Fast Text,Word2Vec	F-score: 67.2	Biomedical named entities (BioNER) recognition and extraction.
[40]	CNN, RNN	Learning rate: 0.0001 and Beta1rate: 0.5	CNN	Recall :99.5%, F1-score:99.1%, Average Precision: 90.9%	Extraction of texual information from images of patient's medical laboratory reports
[19]	CNN	Learning rate: 1e−3, Adam optimizer,Batch size:30 Embedding size: 128	Word embedding	Average Precision: 85 to 90.%	Gridded text with spatial and semantic knowledge of features.
[23]	FCN	Confidence threshold: 0.5, non-maximal suppression stage added	Convolution layer with RoIAlign, YOLOv2-to predict bounding box	Average Precision: 85 to 90.%	Template-free historical handwritten form data extraction.
[21]	Bi-LSTM-CRF	300-dimensional hidden states in Bi-LSTM,different dropout,learning rate and batch size for each entity extraction	Word embedding	Macro-average F1:0.80	Legal contract entity extraction
[42]	LSTM, BERT	LSTM layer size: 128, learning rate :0.0001,embedding size:25	BERT pre-trained network as feature extractor	F1:0.88	Recognizing named entities from financial and biomedical documents.
[18]	BERT	Batch size:4, Learning rate:10e-5,embedding size:128	BERT pre-trained network as feature extractor	F1:57.38	Extraction of relationship amongst the entities at document-level



	A	B	C
1	Entities	Value	
2	Supplier Name	AVIVA DOORS	
3	Invoice No.	TI-531/19-20	
4	Buyer Name	Siddharth Properties	
5	Invoice Date	04-12-2019	
6	Buyer GST	27AAZFS3607C1ZY	
7	Supplier GST	27AARFA8194J1ZP	
8	Grand Total	71036	
9			
10			

FIGURE 3. Detailed Process Flow for Key Fields Extraction from Multi-layout Unstructured Invoice Documents.

various issues, as discussed in Section II. Existing literature shows that sophisticated invoice reader systems currently in use are implemented on structured invoice datasets. Therefore, there is a scope to use AI on unstructured datasets. Many AI algorithms can not discern the key-value pairs when exposed to datasets with random structures. Such datasets can help improve the learning of the algorithms by making them learn more robust features. Our dataset will help researchers and organizations to evaluate their results on multi-layout, highly diverse, and varied invoices. This research also presents template-free processing of our multi-layout invoices collected from different suppliers using the BiLSTM and BiLSTM-CRF model.

Table 2. summarizes few Deep Learning approaches for Named Entity Recognition (NER) proposed in the existing studies.

III. PROPOSED PROCESS FLOW

This section elaborates a detailed process flow for automatic and end-to-end key fields extraction from unstructured documents. Figure 3. depicts the process flow to understand the overall key fields extraction process used for extracting insights from multi-layout unstructured invoice documents. For this, we took the example of scanned invoice PDFs as unstructured documents, discussed in brief as follows:

Scanned invoice PDF is converted into an image and is processed for text recognition and extraction using OCR. OCR engine outputs the text file and also provides coordinates and confidence of the text. These text files are then passed on to the annotation tool, UBIAI, where every file is annotated based on the labels provided to UBIAI. The output of these annotation files can be downloaded in multiple formats, as shown above, such as JSON, IOB, and Spacy. Kruskal-Wallis

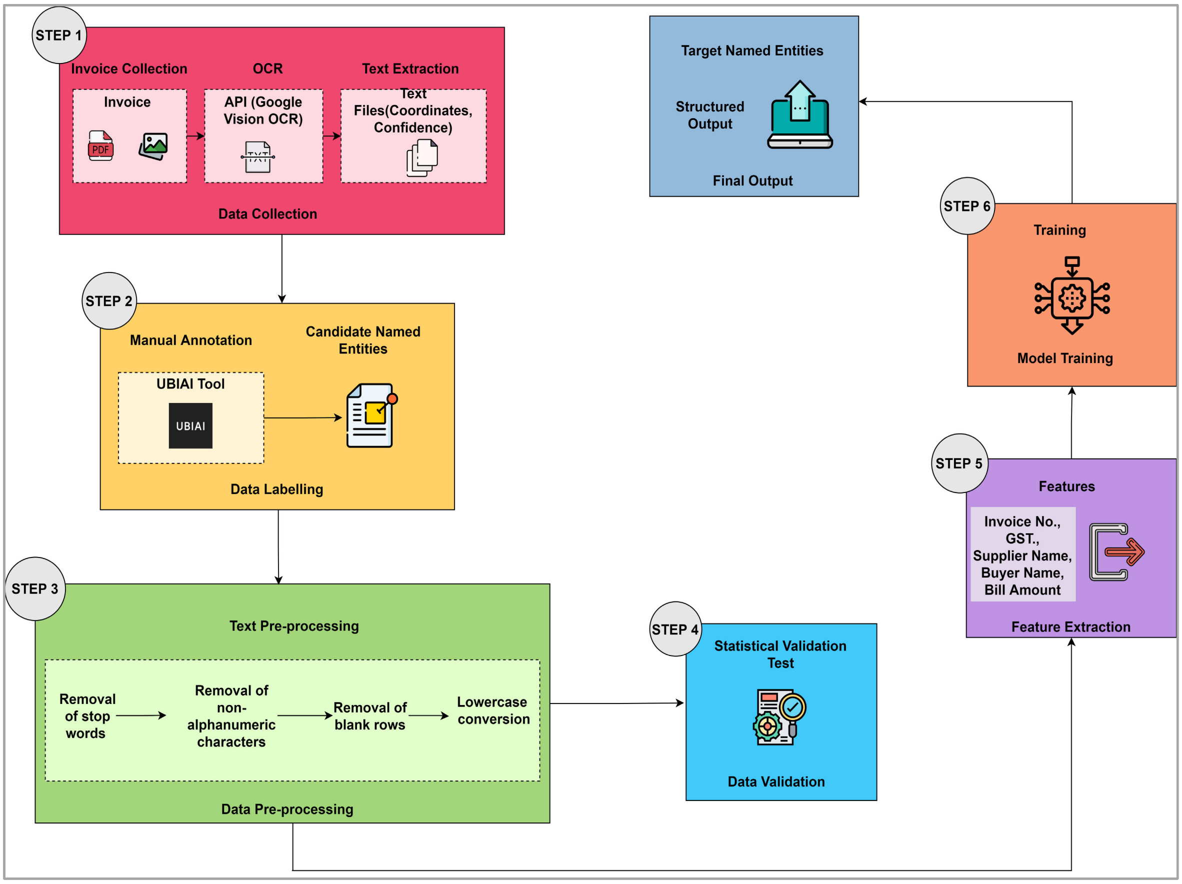


FIGURE 4. Proposed Framework for Key Fields Extraction from Multi-layout Unstructured Invoice Documents.

H test is used as a statistical test for assessing the data quality. IOB format is inputted for building a Machine Learning model. Structured output is extracted key fields from invoices that can be displayed in CSV, JSON, and PDF format.

A. PROPOSED FRAMEWORK

Figure 4. shows the proposed framework for key fields extraction from multi-layout unstructured invoice documents with the detailed steps discussed as follows:

1) DATA COLLECTION

Data collection includes gathering relevant unstructured documents such as scanned invoice PDF documents essential for daily financial transactions for every organization. These invoices can be in the form of images or scanned PDF documents. Publicly available standard datasets can be utilized for this task. Still, due to the issues discussed in Section II, the custom dataset is given preference over any other available datasets.

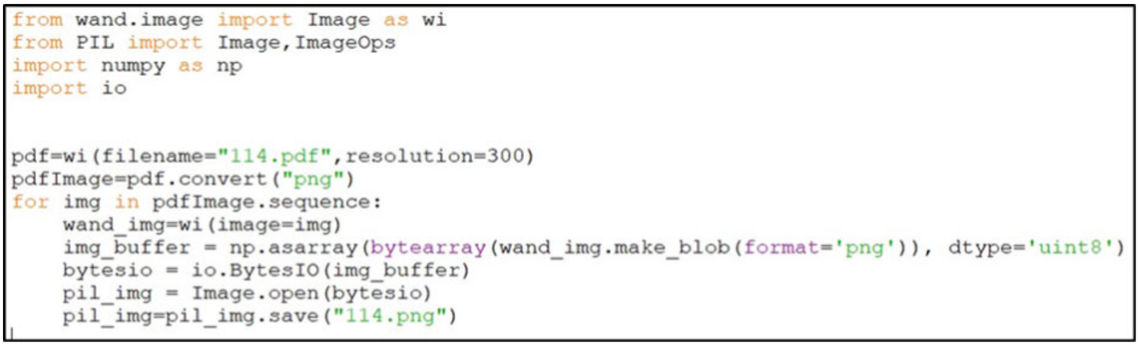
Table 3. shows the details of the data collected from the different suppliers of the organization, the number of documents collected per supplier and key fields, and their labels used. Our custom dataset includes supplier invoices of multiple and diverse layouts. Supplier1 with 196 invoices, Supplier2 with 6 invoices, Supplier3 with 29 invoices, Supplier4 with 14 invoices, Supplier5 with 391 invoices, Supplier6 with 406 invoices, and a few from other suppliers, which make a total of 1646 invoices. All these invoices are available in scanned PDF format, with each company having a different and complex layout of the invoice.

Collected invoice scanned PDFs are needed to convert into the image format for inputting to Optical Character Recognition (OCR) engine. Google Vision OCR is employed to detect, recognize and extract the textual contents from these document images. OCR extracts the text from a given image and outputs a text file for each image. Figure 5. shows the code block to convert the scanned PDF documents to the image format. Wand and PIL (Python Image Library) Python libraries convert PDF files to the image format. PIL is used for image processing operations and getting different image files like.png,.jpeg. All the images are then converted into their text file using Google Vision OCR. X-Y position coordinates (spatial distribution) of each extracted word and confidence score are also obtained as OCR output along with the extracted text.

- **Different layouts of invoices in our multi-layout unstructured invoice document dataset:** Figure 6. depicts the sample layouts of invoices contain in our multi-layout unstructured invoice documents dataset. It consists of a variety of supplier invoices having

TABLE 3. Multi-layout Unstructured Invoice Document Dataset Details.

Number of suppliers	Number of documents	Key fields with their labels assigned
Supplier 1	196	Supplier Name: Supp_N Supplier GST Number: Supp_G Buyer Name: Buy_N Buyer GST Number: Buy_G Invoice Date: INV_DT Invoice Number: INV_NO Grand Total Amount for Invoice: GT_AMT
Supplier 2	06	
Supplier 3	29	
Supplier 4	14	
Supplier 5	391	
Supplier 6	406	
Supplier 7	217	
Supplier 8	387	
Total	1646 invoice documents	



```
from wand.image import Image as wi
from PIL import Image,ImageOps
import numpy as np
import io


pdf=wi(filename="114.pdf",resolution=300)
pdfImage=pdf.convert("png")
for img in pdfImage.sequence:
    wand_img=wi(image=img)
    img_buffer = np.asarray(bytearray(wand_img.make_blob(format='png')), dtype='uint8')
    bytesio = io.BytesIO(img_buffer)
    pil_img = Image.open(bytesio)
    pil_img=pil_img.save("114.png")
```

FIGURE 5. PDF to Image Conversion Code.

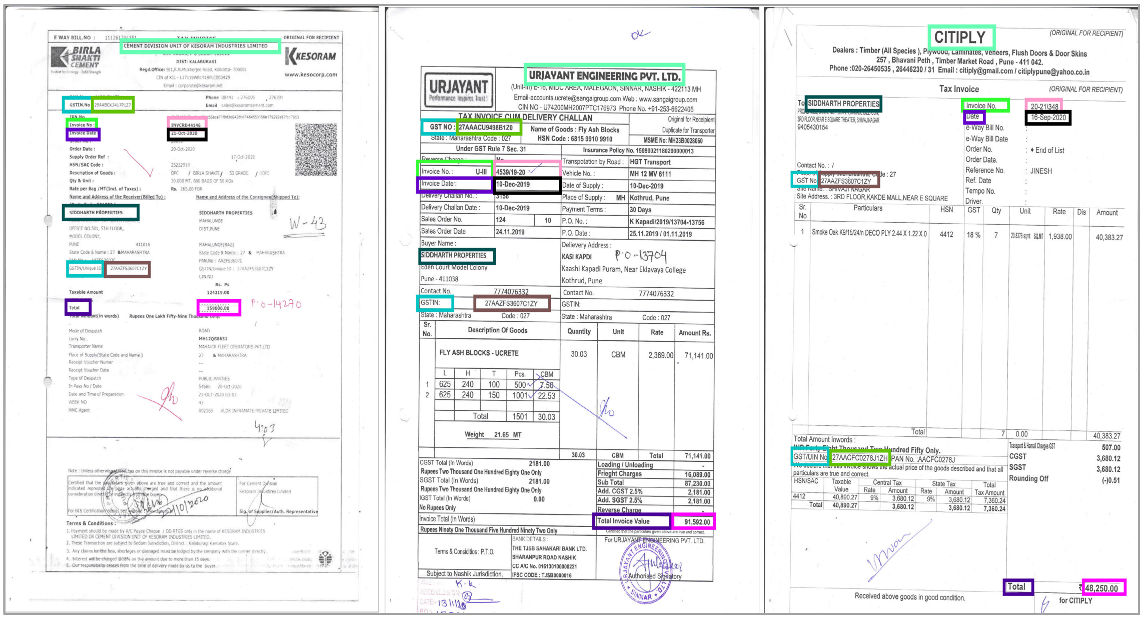
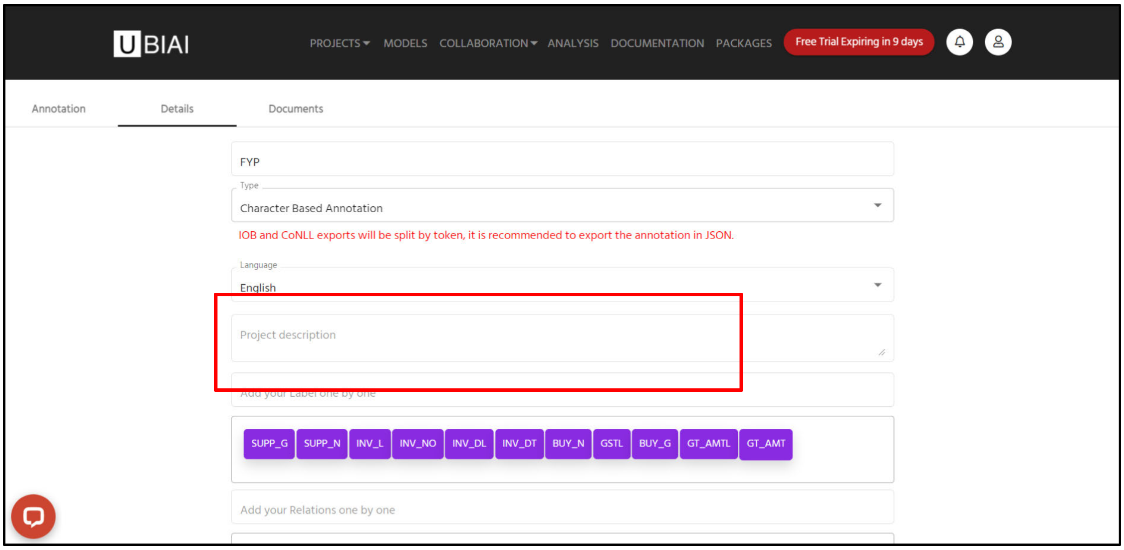

KESORAM
URJAYANT ENGINEERING PVT. LTD.
TAX INVOICE CUM DELIVERY CHALLAN
CITIPLY
Tax Invoice

FIGURE 6. Diverse and Multiple Invoice Layouts in Multi-layout Unstructured Invoice Documents Dataset.

complex and multiple layouts. Figure 6. shows that the location or placing of various data fields in the invoices varies, as the supplier invoice formats changes. Each supplier has its unique layout format for the invoices. For example, an invoice number label represented with a green bounding box is located at different places for different supplier invoice layouts. It is concluded from Figure 6 that our invoice dataset consists



FYP

Type
Character Based Annotation

IOB and CoNLL exports will be split by token, it is recommended to export the annotation in JSON.

Language
English

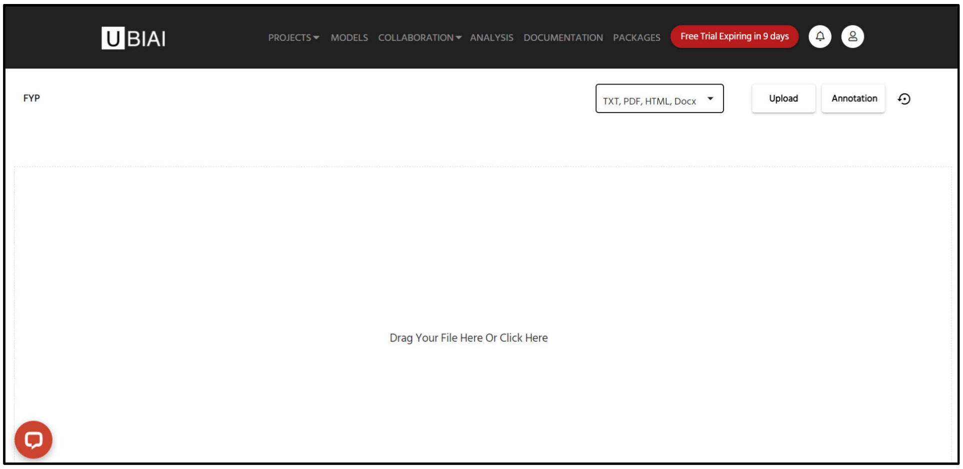
Project description

Add your Label one by one

SUPP_G SUPP_N INV_L INV_NO INV_DL INV_DT BUY_N GSTL BUY_G GT_AMTL GT_AMT

Add your Relations one by one

FIGURE 7. **Process of Adding the Labels using the UBIAI Tool.**



FYP

TXT, PDF, HTML, Docx

Drag Your File Here Or Click Here

FIGURE 8. **Process of Uploading the Text Files to the UBIAI Tool.**

of heterogeneous and complex layout invoices which are close to a real-world scenario, which will help the researchers in this domain to understand the experiments performed on such varied and complex unstructured documents dataset.

2) DATA LABELING/ANNOTATION

Once the text files are ready, each extracted entity is labeled as per the requirement. So, the labels used for annotations are Supp_N for Supplier name, Supp_G for Supplier GST number, BUY_N for Buyer Name, BUY_G for Buyer GST number, GSTL for GST Label, INV_NO for Invoice Number, INV_L for Invoice Number Label, INV_DL for Invoice Date Label, INV_DT for Invoice Date, GT_AMTL for Grand Total Amount Label and GT_AMT for Grand Total Amount. The entities are labeled using these 11 labels in total. UBIAI data annotation tool is used for labeling the entities. The desired labels are provided to the UBIAI tool. Figure 7. shows the UBIAI interface to add labels.

After the labels are added, all text files are uploaded either by drag and drop or by browsing. Figure 8. shows the UBIAI interface for uploading the text files.

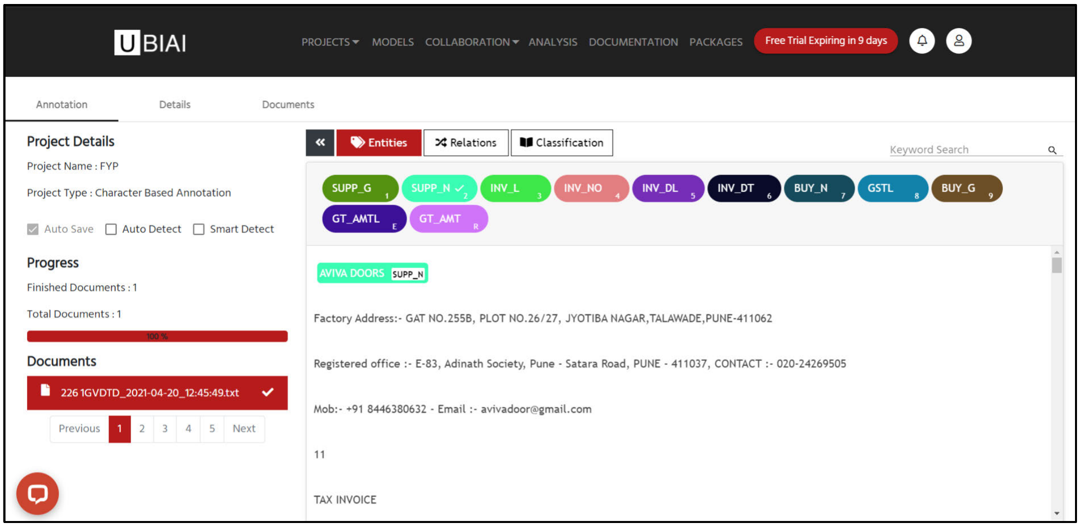
Text files can be annotated by selecting labels, one by one and highlighting the respective text. Figure 9. shows the UBIAI interface for the annotations. The output of UBIAI can be downloaded in multiple formats such as JSON, Spacy, and IOB, as shown in Figure 10. Later, IOB output format files are imported from the UBIAI tool and transformed into CSV.

3) DATA PRE-PROCESSING

Figure 11. shows the pipeline used to perform data pre-processing, which is discussed as follows:

a: REMOVAL OF STOP WORDS

Stop words, for instance, pronouns (we, his), conjunctions (and, or), articles (a, an, the), do not contribute to understand the meaning of a sentence or document. Thus, it is necessary to remove such stop words to reduce the memory size as well



Project Details

Project Name : FYP

Project Type : Character Based Annotation

☑ Auto Save ☐ Auto Detect ☐ Smart Detect

Progress

Finished Documents : 1

Total Documents : 1

100 %

Documents

226 1GVDTD_2021-04-20_12:45:49.txt ✓

AVIVA DOORS SUPP_N

Factory Address:- GAT NO.255B, PLOT NO.26/27, JYOTIBA NAGAR,TALAWADE,PUNE-411062

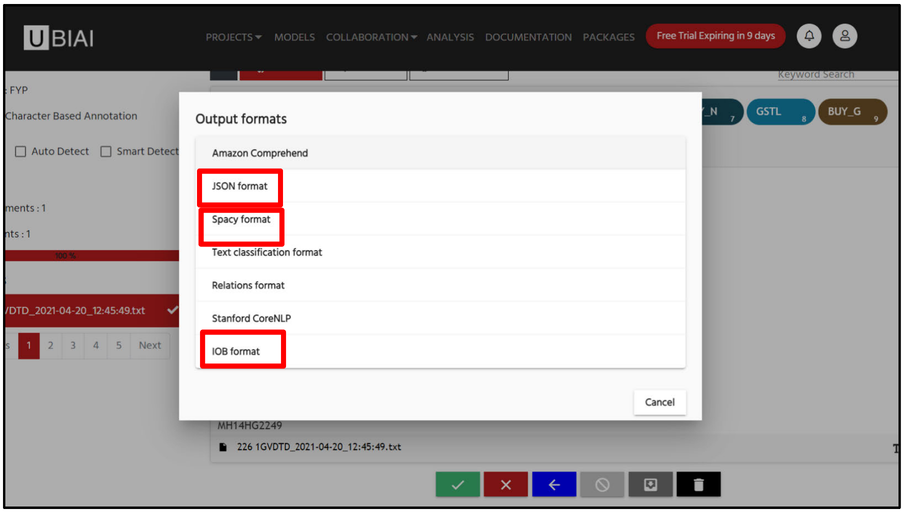
Registered office :- E-83, Adinath Society, Pune - Satara Road, PUNE - 411037, CONTACT :- 020-24269505

Mob:- +91 8446380632 - Email :- avivadoor@gmail.com

11

TAX INVOICE

FIGURE 9. Process of Annotating the Text Files to the UBIAI Tool.



: FYP

Character Based Annotation

☐ Auto Detect ☐ Smart Detect

ments : 1

nts : 1

100 %

/DTD_2021-04-20_12:45:49.txt ✓

Output formats

Amazon Comprehend

JSON format

Spacy format

Text classification format

Relations format

Stanford CoreNLP

IOB format

MH14HG2249

226 1GVDTD_2021-04-20_12:45:49.txt

FIGURE 10. UBIAI Tool Output Formats.

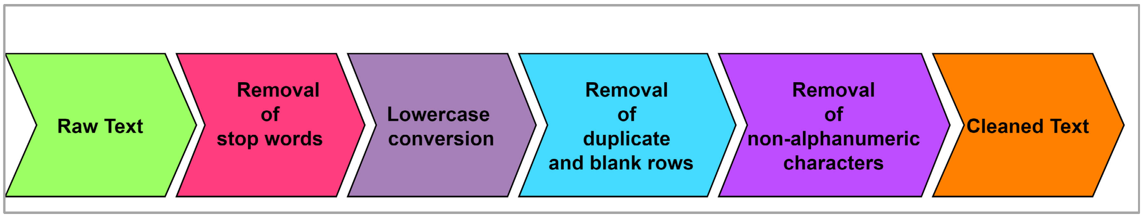

Raw Text
Removal of stop words
Lowercase conversion
Removal of duplicate and blank rows
Removal of non-alphanumeric characters
Cleaned Text

FIGURE 11. Data Pre-processing Pipeline.

as processing overhead. Articles, pronouns, and conjunctions are removed as a first and important pre-processing stage from the obtained.csv files.

b: LOWERCASE CONVERSION

All the tokens obtained after Stop-word removal are converted to lowercase characters. It is an essential step to maintain the consistent format of the output.

c: REMOVAL OF DUPLICATE AND BLANK ROWS

This is an important data pre-processing step to remove the duplicate entries and blank rows added due to OCR errors.

d: REMOVAL OF NON-ALPHANUMERIC CHARACTERS

Non-alphanumeric characters such as ":" or "-" are removed using regular expressions. The invoice date field may contain

TABLE 4. Results from the Kruskal Wallis H Test.

Statistics	Word2Vec	Glove	FastText	Inference
(H-value)	386.056	414.604	425.981	The invoice documents collected have a similar distribution in a different group of layouts.
P-value	0.994	0.932	0.863	The Null Hypothesis is accepted, and invoice documents collected have a similar distribution in a different group of layouts.

the "/" character, which is important to be kept for understanding the date entity. Hence, a regular expression is not applied to each entity.

4) DATA VALIDATION

Data validation is employed to assess the data quality, and it ensures that the data is complete, unique and there are no blank or null values. Validating the accuracy, clarity, and details of the data is necessary to mitigate further errors and meaningless information extraction. Without validating data, there can be the risk of decisions based on data with erroneous values.

The features extracted from our dataset are of a nominal or categorical type of data, for example, Buyer Name, Invoice Date, and Supplier Name. For Nominal or Categorical data, non-parametric statistical tests are more suitable. We have used the Kruskal Wallis H (KWH) test to calculate the medians of each invoice layout subgroups. For example, we validate similar distribution in all invoice documents layouts such as Supplier1 layout, Supplier2 layout, and other supplier layouts. KWH is a popular rank-based non-parametric statistical test performed to decide the significant statistical variances amongst two or more data sample groups. Invoice documents with different Supplier layouts are data samples in our case.

a: KRUSKAL-WALLIS H TEST STEPS

- The first step is to input the numeric vectors generated with word embedding to the Kruskal-Wallis test. We used Word2Vec, Glove, and FastText word embeddings to perform KWH-test.
- KWH provides the H-test statistics for the sample distributions. Based on these H-test statistics and degree of freedom, a p-value is calculated. Together H and p-value determine the validity of a null hypothesis. Calculate the H-statistics using the following formula:

$$H\ statistics = \left[\frac{12}{s\,(s+1)}\sum_{j=1}^{c}\frac{T_j^2}{s_j}\right] - 3\,(s+1) \qquad (1)$$

whereas,

s = summation of all sample sizes
T_j = summation of ranks in j^{th} sample
c = total sample count
s_j = size of j^{th} sample

- After H statistics, the second step is to evaluate any significant differences amongst the data samples by calculating "the p-value." A probability is calculated by observing two data samples and is compared to the significance level to evaluate the Null Hypothesis. Significance level, usually represented as α or alpha of 0.05 ($\alpha = 0.05$), is normally considered. The Null Hypothesis (H0) defines that the medians of all the data samples (invoice documents layouts in our case) are all equal, and no variance exists amongst the data sample distributions.

b: INTERPRETATION OF THE KWH TEST

The final value of a statistic is used to conclude whether to accept or discard the Null Hypothesis. Provided the Null Hypothesis (basic assumption) that the two data samples are taken out from the same population with the same distribution, the p-value is the probability of observing those two data samples of the population. The p-value is compared to the Significance level. If the p-value is less than or equal to the Significance level, that is, $\alpha <= 0.05$, we reject the Null Hypothesis (H0). Hence, it can be determined that the given data samples have a different distribution. Suppose the p-value is greater than the Significance level, that is, $\alpha >= 0.05$. In that case, it is concluded that there is insufficient evidence for rejecting the Null Hypothesis and data samples possess the same distribution.

H0: The median of different invoice sample layouts is equal.

H1: The median of different invoice sample layouts is not equal.

The p-value for the data samples came out to be 0.994 using Word2vec, 0.932 using GloVe, and 0.863 using FastText. The obtained value is statistically significant, and there is no indication for rejecting the Null Hypothesis. Hence, it can be observed that all the given data samples take up equal distribution. Table 4. displays the results obtained by the KWH test.

5) FEATURE EXTRACTION

Feature extraction is performed to identify the required features from a large dataset. It reduces the dimensions of a large dataset. Various techniques are used for feature extraction. Word embedding is a widely used feature extraction method

TABLE 5. Multi-layout Unstructured Invoice Documents Dataset Size Details.

Number of suppliers	Number of documents	Size (in MB)	Labels Used
Supplier 1	196	164 MB	• Supplier Name • Supplier GST Number • Buyer Name • Buyer GST Number • Invoice Date • Invoice Number • Grand Total Amount for Invoice
Supplier 2	06	4.39 MB	
Supplier 3	29	25.8 MB	
Supplier 4	14	23.6 MB	
Supplier 5	391	353 MB	
Supplier 6	406	203 MB	
Supplier 7	217	185 MB	
Supplier 8	387	378 MB	
Total	1646	1336.79 MB	
Number of training samples	**1153**	**70% of total Invoices**	
Number of testing samples	**493**	**30% of total Invoices**	

from unstructured documents. It is a numeric vector input that represents a word. It works on the idea that unstructured textual data is transformed into numeric vectors, which can be understandable by any Machine Learning model. Popular word embedding techniques include - GloVe (Global Vectors), Word2vec, and FastText.

6) MODEL TRAINING

In ML model training, based on the input data, the model is supposed to recognize, learn and extract expected and relevant feature values. The model performance is evaluated using different metrics like Precision, Recall, and F1-score.

IV. DATASET EVALUATION

A. EXPERIMENTAL SETUP

Experiments were executed on HP Pavilion Gaming Laptop (AMD Ryzen 5) with a Nvidia Graphics card with 4 GB memory (GeForce GTX 1650 Ti graphics card). The Deep Learning models were trained using Nvidia DGX-Server with 4 Nvidia Tesla V-100 GPUs with 32 GB memory. Due to the limited capability of the HP Pavilion laptop, Google Colab with CPU runtime was used.

B. SIZE OF MULTI-LAYOUT UNSTRUCTURED INVOICE DOCUMENT DATASET

Table 5. shows the details of the number of invoice documents collected from each supplier, the document size, and the number of training and testing sample sizes, used to train the model. Target key fields to be extracted in the form of structured output from the whole invoice document are shown as "Labels used" in Table 5.

C. FEATURE EXTRACTION TECHNIQUES USED

Machine Learning model processes and understands any text input only in the form of vector or numeric representation. The model performance is eventually dependent on selecting appropriate word embeddings, so the model gets sufficient information on the input. Hence, text representation as a numerical vector or embedding is a promising research focus. The potential effect of various feature extraction techniques is investigated for extracting key fields from unstructured documents. It is discussed as follows:

1) Word2Vec

It is the most well-known and popularly used word-level vector representation. Word2Vec uses Neural Networks to provide the word context of the center word based on adjacent words (CBOW) or Vice Versa (skip-gram). It is a probability-based prediction method that predicts the word context given the adjacent words [43].

2) GloVe

It is the most commonly used pre-trained word-level embedding method to get the global co-occurrence and local context of a word. It is a statistical frequency-based method to obtain the word-to-word co-occurrence information [43].

3) FastText

Obtaining the vector representation or embeddings of rarely used words in the corpus is sometimes difficult with other word embeddings such as Word2Vec and GloVe. This problem is solved using FastText word embedding representation. A subword representation based on n-gram is a popular characteristic of FastText. In other words, it also processes out of vocabulary words. Hence, in most of the text analysis and extraction tasks, it performs better than other word embeddings. FastText allows subword meaning to be captured by a model, even though the complete word is not present in the corpus [44]. For example, features extracted using this technique are – Buyer and Supplier Name and their respective GST Number, Invoice Number, Invoice Date, Grand Total, and few others.

V. RESULTS AND DISCUSSION

Various AI approaches are used for recognizing and extracting the named entities from multi-layout unstructured invoice documents. The presented multi-layout unstructured invoice documents dataset is also evaluated using these AI approaches discussed as follows:

A. BiLSTM (BIDIRECTIONAL LONG SHORT-TERM MEMORY)

BiLSTM is a Recurrent Neural Network (RNN) that processes the text sequence with two LSTM, a forward LSTM, and a backward LSTM. It is used when the prediction tasks

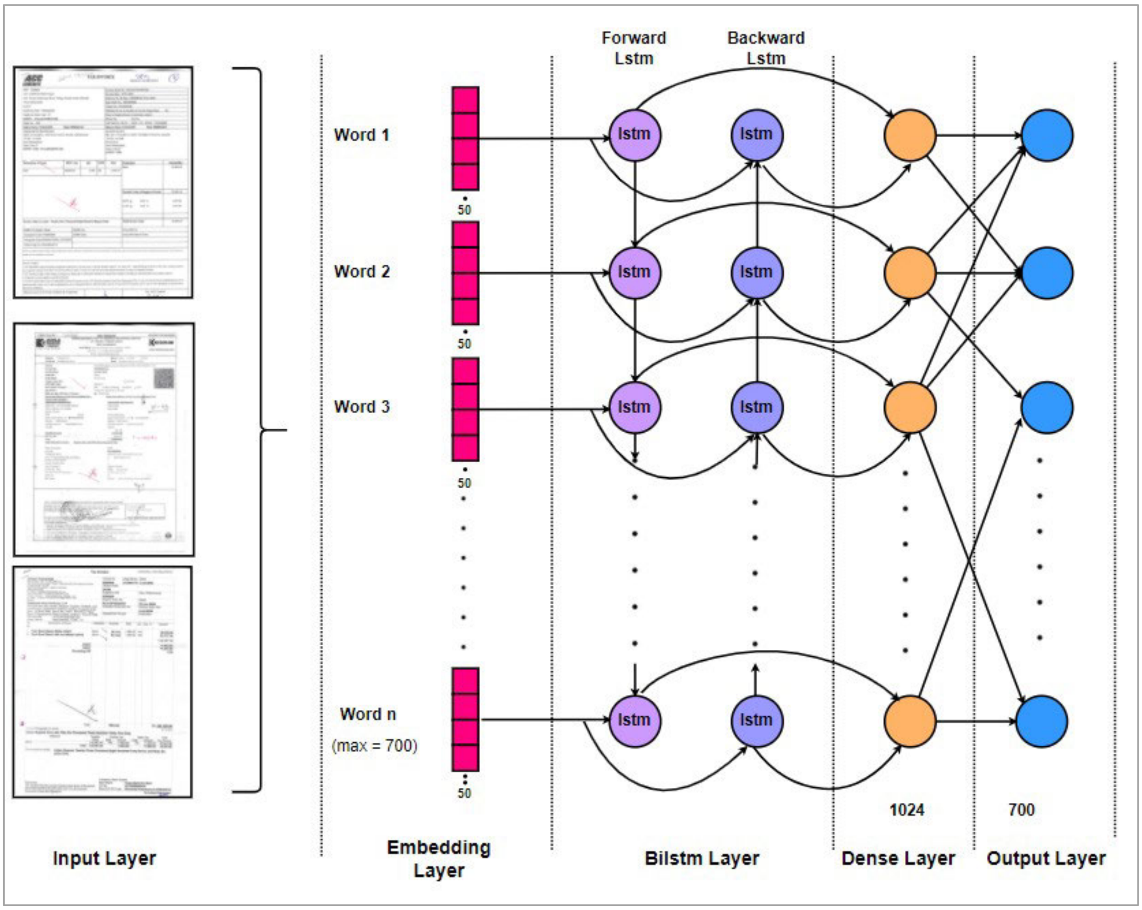


FIGURE 12. BiLSTM Model Architecture.

TABLE 6. Dataset Evaluation Results using BiLSTM with Various Feature Extraction Methods.

Sr. No	Model	Features	Entities Extracted	Precision	Recall	F1-score
1.	BiLSTM	Word2Vec	Invoice Number	0.00	0.00	0.00
			Invoice Date	0.00	0.00	0.00
			Buyer Name	0.00	0.02	0.00
			Buyer GST Number	0.00	0.00	0.00
			Supplier Name	0.29	0.40	0.32
			Supplier GST Number	0.00	0.00	0.00
			Grand Total Amount	0.00	0.00	0.00
	BiLSTM	GloVe	Invoice Number	0.10	0.08	0.09
			Invoice Date	0.01	0.27	0.02
			Buyer Name	0.17	0.17	0.16
			Buyer GST Number	0.02	0.01	0.01
			Supplier Name	0.02	0.01	0.01
			Supplier GST Number	0.01	0.06	0.02
			Grand Total Amount	0.00	0.00	0.00
	BiLSTM	FastText	Invoice Number	0.04	0.03	0.03
			Invoice Date	0.00	0.01	0.00
			Buyer Name	0.11	0.53	0.13
			Buyer GST Number	0.00	0.00	0.00
			Supplier Name	0.33	0.61	0.40
			Supplier GST Number	0.00	0.00	0.00
			Grand Total Amount	0.00	0.00	0.00

depend on the previous as well as the future input sequence. Table 6. demonstrates the potential effect of word embeddings like Word2Vec, GloVe and FastText used along with BiLSTM to extract key fields from multi-layout unstructured invoice documents dataset. Figure 12. shows the details of the BiLSTM model architecture used.

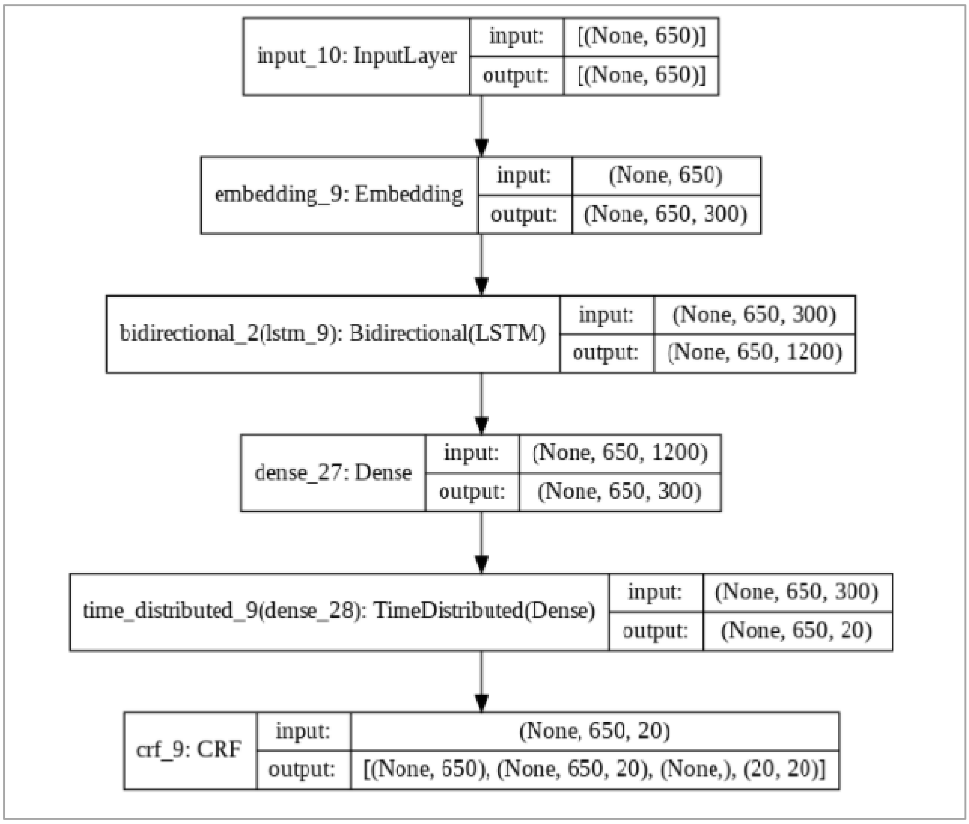


FIGURE 13. **BiLSTM-CRF Model Architecture.**

1) INPUT LAYER

It takes the word embeddings of the words present in the IOB file.

2) EMBEDDING LAYER

It converts the embeddings passed by the input layer to dense vectors of the fixed size.

3) BiLSTM LAYER

It captures the sequences present in the data.

4) DENSE LAYER

It introduces non-linearity in the output using the 'ReLU' activation function.

5) OUTPUT DENSE LAYER

The final output layer with the same dimensions as the input layer. It extracts the desired named entities.

Table 6. shows that Word2Vec and GloVe word embedding methods are usually used with language models by including the hidden layers to obtain the contextual embeddings. However, training GloVe and Word2Vec on unstructured documents like invoices is not a good idea, since formats of Invoice Numbers, Total Invoice Amounts, and Invoice Dates are very much unique [44]. The embeddings obtained for such unique words will be of low quality and will not be useful for key fields extraction tasks. FastText performs better in most of the key field extraction tasks, unlike Word2Vec and GloVe [43].

B. BiLSTM-CRF (BiLSTM-CONDITIONAL RANDOM FIELD)

It is another sequence processing model with an understanding of Bidirectional LSTM and data labeling logic by CRF. BiLSTM with CRF model learns the weights related to data samples. It achieves improved results over BiLSTM in natural language processing or analysis tasks [45]. Table 7. shows the results obtained using the BiLSTM-CRF model. Figure 13. shows the details of the BiLSTM-CRF model architecture used. CRF layer is added to the BiLSTM to understand the internal data labeling structure in IOB formatted tags.

Figure 13. Also shows that the BiLSTM-CRF model includes four hidden layers comprising two LSTM layers, one dense layer and one time- distributed dense layer. We have used MSE (Mean Squared Error) loss function as the objective function. MSE takes the loss between the predicted and actual key field values and backpropagates it for learning.

Table 7. shows that BiLSTM-CRF performs better than BiLSTM with word embeddings used, such as Word2Vec, GloVe, and FastText. BiLSTM-CRF has a more complicated architecture for better word representation with the multiple dense layers, and the CRF layer works well in the text classification task. CRF also understands well the internal labeling structure of entity tags.

TABLE 7. Dataset Evaluation Results using BiLSTM-CRF.

Sr. No	Model	Features	Entities Extracted	Precision	Recall	F1-score
	BiLSTM-CRF	Word embedding layer	Invoice Number	0.98	0.88	0.93
			Invoice Date	0.76	0.87	0.81
			Buyer Name	0.95	0.97	0.96
			Buyer GST Number	0.60	0.96	0.74
			Supplier Name	0.96	0.72	0.83
			Supplier GST Number	0.97	0.73	0.83
			Grand Total Amount	0.90	0.75	0.82

VI. LIMITATIONS

A. SIZE OF THE DATASET

Data used for training serves as a pillar of any AI model. Proposed model learning is heavily dependent on the variations in layouts of the unstructured documents and the number of sample invoices of each layout in the traning data. Limited availability of such multi-layout varied invoices is a challenge in this research area.

B. MANUAL DATA ANNOTATION

Invoice documents are manually annotated, which is error-prone and dependent on the knowledge of the annotator on the data labeling task.

C. RESULTS DEPEND ON SELECTING FEATURE EXTRACTION METHODS

Most popular word embedding methods are Word2Vec, Glove, and FastText. Word embedding methods aim to find the relationship or context amongst the words or subwords present in the document and represent it in a numerical vector. The unstructured documents like invoices contain uncommon words as their key fields such as Invoice Number, Total Invoice Amount, and Invoice Date. Hence, key-field extraction research is dependent on the type of word embeddings used.

D. LIMITED DATA VALIDATION METHODS FOR EVALUATING DATA

The data quality is assessed with limited statistical tests. In this research work, we employed the Kruskal Wallis H test for checking the quality of data. Few more tests can be used to access the data quality.

E. LIMITED DL APPROACHES

The proposed multi-layout unstructured invoice documents dataset is evaluated with BiLSTM and BiLSTM-CRF. Pre-trained DL approaches such as BERT and its variants like RoBERTa, can be employed for key field extraction from unstructured documents.

VII. FUTURE WORK

A. INCREASING THE DATASET SIZE

AI model robustness and accuracy rely on the size of the data used for model training. Collecting more invoice documents is challenging since invoices are confidential documents to business enterprises. Hence, data augmentation is an interesting research focus in this area for increasing the data size.

B. AUTOMATING THE DATA ANNOTATION PROCESS

Manually annotating the invoices or any other unstructured documents is tedious, laborious, error-prone, and time-consuming. Automating the annotations will simplify and save the manual work so that researchers can focus on advancements in the model, rather than putting efforts into data labeling tasks. Hence, automating data annotation is another research advancement in this area.

C. USE OF OBJECT DETECTION AND CLASSIFICATION (YOLOv5) APPROACH

YOLOv5(You Only Look Once), a Convolutional Neural Network (CNN) for detecting an object in real-time and key fields extraction from the unstructured documents such as Driving License and Identity Cards, is another fascinating and upcoming research trend for many researchers.

D. DEVELOPING THE SYSTEM FOR HANDWRITTEN TEXTUAL CONTENTS EXTRACTION

Handwritten unstructured documents such as shop-keeper receipts and clinical notes serve as important data for key-field extraction and analysis tasks. Another future direction is to automate the recognition and key-field extraction of handwritten characters in such unstructured documents. The researchers may advance this research by developing a framework or model for key field extraction from printed and handwritten unstructured documents combinedly.

E. DEVELOPING AN AI-BASED TOOL FOR KEY FIELD EXTRACTION FROM UNSTRUCTURED DOCUMENTS

Developing an AI-based tool for key field extraction from unstructured documents is another prospective research direction that helps the business organizations extract key fields and reduce the manual efforts of data entry and verification of the important data insights from the unstructured documents.

F. TRANSFER LEARNING FOR KEY FIELD EXTRACTION

Transfer learning approaches such as BERT and the variants of BERT like RoBERTa can be employed to evaluate the performance of our multi-layout unstructured invoice documents dataset.

VIII. CONCLUSION

This work contributes to the research area by developing a multi-layout, high-quality, and highly diverse unstructured invoice documents dataset. The quality of data samples from the developed multi-layout unstructured document dataset is assessed using a statistical data validation method known as the Kruskal Wallis H test. The results obtained from this statistical test suggest that all the data samples are appropriate to train the model. This work also presents the detailed process flow for end-to-end key fields extraction tasks from complex and varied unstructured documents. This work also presents a comparative analysis of the results of different feature extraction techniques employed and investigates the suitability of feature extraction techniques for invoice documents datasets. Our observation suggests that the dynamic word embeddings with the embedding layer of the AI-based model perform better than word embeddings such as Word2Vec, GloVe, and FastText. Word embeddings like Word2Vec, GloVe, and FastText do not provide the proper vector representation of unique words present in unstructured documents like invoice number or the total invoice amount. This work also demonstrates the comparative analysis of various AI approaches for template-free processing and key field extraction, such as BiLSTM and BiLSTM-CRF. We evaluated our dataset successfully and efficiently using BiLSTM-CRF. We achieved the best results despite complex, multi-layout, and varied unstructured invoice documents included in the dataset. This work is able to achieve the generalization in key fields extraction task with the help of presented multi-layout, high-quality, and highly diverse unstructured invoice documents dataset. It will help fellow researchers to have a deep dive into this research area and consider our work and dataset as a baseline work. This work also presents few promising and interesting future research directions for key field extraction tasks.

REFERENCES

[1] A. C. Eberendu, "Unstructured data: An overview of the data of big data," *Int. J. Comput. Trends Technol.*, vol. 38, no. 1, pp. 46–50, Aug. 2016, doi: 10.14445/22312803/ijctt-v38p109.

[2] *30 Eye-Opening Big Data Statistics for 2020: Patterns are Everywhere*. Accessed: Dec. 5, 2020. [Online]. Available: https://kommandotech.com/statistics/big-data-statistics/

[3] *Unstructured Data Process Automation*, Everest Group, Dallas, TX, USA, 2019.

[4] S. Patel and D. Bhatt, "Abstractive information extraction from scanned invoices (AIESI) using end-to-end sequential approach," 2020, *arXiv:2009.05728*. [Online]. Available: http://arxiv.org/abs/2009.05728

[5] D. Baviskar, S. Ahirrao, and K. Kotecha, "A bibliometric survey on cognitive document processing," Library Philosophy Pract. (e-journal), Univ. Nebraska-Lincoln, Lincoln, Nebraska, Tech. Rep. 4557, 2020. [Online]. Available: https://digitalcommons.unl.edu/libphilprac/4557

[6] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha, "Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions," *IEEE Access*, vol. 9, pp. 72894–72936, 2021, doi: 10.1109/access.2021.3072900.

[7] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "FUNSD: A dataset for form understanding in noisy scanned documents," 2019, *arXiv:1905.13538*. [Online]. Available: http://arxiv.org/abs/1905.13538, doi: 10.1109/icdarw.2019.10029.

[8] M. Kerroumi, O. Sayem, and A. Shabou, "VisualWordGrid: Information extraction from scanned documents using a multimodal approach," 2020, *arXiv:2010.02358*. [Online]. Available: http://arxiv.org/abs/2010.02358

[9] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar, "ICDAR2019 competition on scanned receipt OCR and information extraction," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1516–1520, doi: 10.1109/ICDAR.2019.00244.

[10] X. Zhong, J. Tang, and A. J. Yepes, "PubLayNet: Largest dataset ever for document layout analysis," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1015–1022, doi: 10.1109/ICDAR.2019.00166.

[11] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou, "DocBank: A benchmark dataset for document layout analysis," 2020, *arXiv:2006.01038*. [Online]. Available: http://arxiv.org/abs/2006.01038, doi: 10.18653/v1/2020.coling-main.82.

[12] D. Tkaczyk, P. Szostek, and Ł. Bolikowski, "GROTOAP2—The methodology of creating a large ground truth dataset of scientific articles," *D-Lib Mag.*, vol. 20, nos. 11–12, Nov. 2014, doi: 10.1045/november14-tkaczyk.

[13] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher, "A realistic dataset for performance evaluation of document layout analysis," in *Proc. 10th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2009, pp. 296–300, doi: 10.1109/ICDAR.2009.271.

[14] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "Efficient and effective OCR engine training," *Int. J. Document Anal. Recognit.*, vol. 23, no. 1, pp. 73–88, Mar. 2020, doi: 10.1007/s10032-019-00347-8.

[15] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *J. Biomed. Informat.*, vol. 77, pp. 34–49, Jan. 2018, doi: 10.1016/j.jbi.2017.11.011.

[16] A. Abbas, M. Afzal, J. Hussain, and S. Lee, "Meaningful information extraction from unstructured clinical documents," *Asia–Pacific Adv. Netw.*, vol. 48, pp. 42–47, Oct. 2019. Accessed: Sep. 17, 2020. [Online]. Available: https://www.researchgate.net/publication/336797539_Meaningful_Information_Extraction_from_Unstructured_Clinical_Documents

[17] Y. S. Chernyshova, A. V. Sheshkus, and V. V. Arlazarov, "Two-step CNN framework for text line recognition in camera-captured images," *IEEE Access*, vol. 8, pp. 32587–32600, 2020, doi: 10.1109/ACCESS.2020.2974051.

[18] X. Han and L. Wang, "A novel document-level relation extraction method based on BERT and entity information," *IEEE Access*, vol. 8, pp. 96912–96919, 2020, doi: 10.1109/ACCESS.2020.2996642.

[19] X. Zhao, E. Niu, Z. Wu, and X. Wang, "CUTIE: Learning to understand documents with convolutional universal text information extractor," 2019, *arXiv:1903.12363*. [Online]. Available: http://arxiv.org/abs/1903.12363

[20] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM model for document-level sentiment analysis," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 3, pp. 832–847, 2019, doi: 10.3390/make1030048.

[21] I. Chalkidis and I. Androutsopoulos, *A Deep Learning Approach to Contract Element Extraction* (Frontiers in Artificial Intelligence and Applications), vol. 302. Amsterdam, The Netherlands: IOS Press, 2017, pp. 155–164, doi: 10.3233/978-1-61499-838-9-155.

[22] R. B. Palm, O. Winther, and F. Laws, "CloudScan—A configuration-free invoice analysis system using recurrent neural networks," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 406–413, doi: 10.1109/ICDAR.2017.74.

[23] B. Davis, B. Morse, S. Cohen, B. Price, and C. Tensmeyer, "Deep visual template-free form parsing," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 134–141, doi: 10.1109/ICDAR.2019.00030.

[24] K. Adnan and R. Akbar, "Limitations of information extraction methods and techniques for heterogeneous unstructured big data," *Int. J. Eng. Bus. Manag.*, vol. 11, pp. 1–23, Dec. 2019, doi: 10.1177/1847979019890771.

[25] R. B. Palm, F. Laws, and O. Winther, "Attend, copy, parse end-to-end information extraction from documents," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 329–336, doi: 10.1109/ICDAR.2019.00060.

[26] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *J. Big Data*, vol. 6, no. 1, p. 91, 2019.

[27] C. Reul, D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, and F. Puppe, "OCR4all—An open-source tool providing a (semi-)automatic OCR workflow for historical printings," *Appl. Sci.*, vol. 9, no. 22, p. 4853, Nov. 2019, doi: 10.3390/app9224853.

[28] J. M. Steinkamp, W. Bala, A. Sharma, and J. J. Kantrowitz, "Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes," *J. Biomed. Informat.*, vol. 102, Feb. 2020, Art. no. 103354, doi: 10.1016/j.jbi.2019.103354.

[29] M. Binkhonain and L. Zhao, *A Review of Machine Learning Algorithms for Identification and Classification of Non-Functional Requirements*, vol. 1. Amsterdam, The Netherlands: Elsevier, 2019.

[30] S. Gehrmann, F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Foote, E. T. Moseley, D. W. Grant, P. D. Tyler, and L. A. Celi, "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives," *PLoS ONE*, vol. 13, no. 2, Feb. 2018, Art. no. e0192360, doi: 10.1371/journal.pone.0192360.

[31] Y. Ye, S. Zhu, J. Wang, Q. Du, Y. Yang, D. Tu, L. Wang, and J. Luo, "A unified scheme of text localization and structured data extraction for joint OCR and data mining," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 2373–2382, doi: 10.1109/BigData.2018.8622129.

[32] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay named entity recognition based on rule-based approach," *Int. J. Mach. Learn. Comput.*, vol. 4, no. 3, pp. 300–306, 2014, doi: 10.7763/ijmlc.2014.v4.428.

[33] Y. Sun, X. Mao, S. Hong, W. Xu, and G. Gui, "Template matching-based method for intelligent invoice information identification," *IEEE Access*, vol. 7, pp. 28392–28401, 2019, doi: 10.1109/ACCESS.2019.2901943.

[34] P. Sahare and S. B. Dhok, "Multilingual character segmentation and recognition schemes for Indian document images," *IEEE Access*, vol. 6, pp. 10603–10617, 2018, doi: 10.1109/ACCESS.2018.2795104.

[35] W. Liu, Y. Zhang, and B. Wan, "Unstructured document recognition on business invoice," Mach. Learn., Stanford iTunes Univ., Stanford, CA, USA, Tech. Rep., 2016. [Online]. Available: http://cs229.stanford.edu/proj2016/report/LiuWanZhang-UnstructuredDocumentRecognitionOnBusinessInvoice-report.pdf

[36] N. Kanya and T. Ravi, "Named entity recognition from biomedical text—An information extraction task," *ICTACT J. Soft Comput.*, vol. 6, no. 4, pp. 1303–1307, Jul. 2016, doi: 10.21917/ijsc.2016.0179.

[37] T. Al-Moslmi, M. G. Ocana, A. L. Opdahl, and C. Veres, "Named entity extraction for knowledge graphs: A literature overview," *IEEE Access*, vol. 8, pp. 32862–32881, 2020, doi: 10.1109/ACCESS.2020.2973928.

[38] S. Joshi, P. Shah, and A. K. Pandey, "Location identification, extraction and disambiguation using machine learning in legal contracts," in *Proc. 4th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Dec. 2018, pp. 1–5, doi: 10.1109/CCAA.2018.8777631.

[39] N. Perera, M. Dehmer, and F. Emmert-Streib, "Named entity recognition and relation detection for biomedical information extraction," *Frontiers Cell Develop. Biol.*, vol. 8, p. 673, Aug. 2020, doi: 10.3389/fcell.2020.00673.

[40] W. Xue, Q. Li, and Q. Xue, "Text detection and recognition for images of medical laboratory reports with a deep learning approach," *IEEE Access*, vol. 8, pp. 407–416, 2020, doi: 10.1109/ACCESS.2019.2961964.

[41] F. Yi, Y.-F. Zhao, G.-Q. Sheng, K. Xie, C. Wen, X.-G. Tang, and X. Qi, "Dual model medical invoices recognition," *Sensors*, vol. 19, no. 20, p. 4370, Oct. 2019, doi: 10.3390/s19204370.

[42] S. Francis, J. V. Landeghem, and M.-F. Moens, "Transfer learning for named entity recognition in financial and biomedical documents," *Information*, vol. 10, no. 8, p. 248, Jul. 2019, doi: 10.3390/info10080248.

[43] B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C.-J. Kuo, "Evaluating word embedding models: Methods and experimental results," *APSIPA Trans. Signal Inf. Process.*, vol. 8, pp. 1–13, Jul. 2019, doi: 10.1017/ATSIP.2019.12.

[44] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 11, pp. 1297–1304, Nov. 2019, doi: 10.1093/jamia/ocz096.

[45] Y. Li, T. Liu, D. Li, Q. Li, J. Shi, and Y. Wang, "Character-based BiLSTM-CRF incorporating POS and dictionaries for Chinese opinion target extraction," *Proc. Mach. Learn. Res.*, vol. 95, pp. 518–533, Nov. 2018. [Online]. Available: https://github.com/kdsec/chinese-opinion-target-extraction

DIPALI BAVISKAR received the master's degree in computer science and engineering from the MGM College of Engineering, Nanded. She is currently pursuing the Ph.D. degree with the Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune. She is working as an Assistant Professor with the School of Computer Engineering and Technology, MIT-WPU, Pune. Her research interests include machine learning, deep learning, and natural language processing.

SWATI AHIRRAO received the Ph.D. degree from Symbiosis International (Deemed University), Pune, Maharashtra, India. She is currently working as an Associate Professor with the Symbiosis Institute of Technology, Pune. According to Google Scholar, her articles have 71 citations, with an H-index of 3 and an i10-index of 2. She has published over 31 research papers in international journals and conferences. Her research interests include big data analytics, machine learning, deep learning, natural language processing, and reinforcement learning.

KETAN KOTECHA received the M.Tech. and Ph.D. degrees from IIT Bombay. He is currently the Head of the Symbiosis Centre for Applied Artificial Intelligence (SCAAI), the Director of the Symbiosis Institute of Technology, and the Dean of the Faculty of Engineering, Symbiosis International (Deemed University). He has expertise and experience in cutting-edge research and projects in AI and deep learning for the last 25+ years. He has published more than 100 widely in a number of excellent peer-reviewed journals on various topics ranging from cutting-edge AI, education policies, teaching-learning practices, and AI for all. He has published three patents and delivered keynote speeches at various national and international forums, including at Machine Intelligence Laboratory, USA; IIT Bombay under the World Bank Project; and the International Indian Science Festival organized by the Department of Science Technology, Government of India. He was a recipient of the two SPARC projects worth INR 166 lakhs from MHRD Government of India in AI in collaboration with Arizona State University, USA, and The University of Queensland Australia. He was also a recipient of numerous prestigious awards like Erasmus+ faculty mobility grant to Poland, DUO-India Professors Fellowship for research in responsible AI in collaboration with Brunel University, U.K.; LEAP grant at Cambridge University, U.K.; UKIERI grant with Aston University U.K.; and a grant from the Royal Academy of Engineering, U.K., under Newton Bhabha Fund. He is an Associate Editor of IEEE Access.