



# Multi-Layout Invoice Document Dataset (MIDD): A Dataset for Named Entity Recognition

Dipali Baviskar <sup>1</sup>, Swati Ahirrao <sup>1,\*</sup> and Ketan Kotecha <sup>2,\*</sup><sup>1</sup> Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India; dipali.baviskar.phd2019@sitpune.edu.in<sup>2</sup> Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune 412115, India

\* Correspondence: swatia@sitpune.edu.in (S.A.); head@scaai.siu.edu.in (K.K.)

**Abstract:** The day-to-day working of an organization produces a massive volume of unstructured data in the form of invoices, legal contracts, mortgage processing forms, and many more. Organizations can utilize the insights concealed in such unstructured documents for their operational benefit. However, analyzing and extracting insights from such numerous and complex unstructured documents is a tedious task. Hence, the research in this area is encouraging the development of novel frameworks and tools that can automate the key information extraction from unstructured documents. However, the availability of standard, best-quality, and annotated unstructured document datasets is a serious challenge for accomplishing the goal of extracting key information from unstructured documents. This work expedites the researcher's task by providing a high-quality, highly diverse, multi-layout, and annotated invoice documents dataset for extracting key information from unstructured documents. Researchers can use the proposed dataset for layout-independent unstructured invoice document processing and to develop an artificial intelligence (AI)-based tool to identify and extract named entities in the invoice documents. Our dataset includes 630 invoice document PDFs with four different layouts collected from diverse suppliers. As far as we know, our invoice dataset is the only openly available dataset comprising high-quality, highly diverse, multi-layout, and annotated invoice documents.



**Citation:** Baviskar, D.; Ahirrao, S.; Kotecha, K. Multi-Layout Invoice Document Dataset (MIDD): A Dataset for Named Entity Recognition. *Data* **2021**, *6*, 78. <https://doi.org/10.3390/data6070078>

Academic Editor: Joaquín Torres-Sospedra

Received: 6 July 2021

Accepted: 16 July 2021

Published: 20 July 2021

**DataSet:** <http://doi.org/10.5281/zenodo.5113009>

**DataSet License:** License under which the dataset is made available (CC-BY-4.0).

**Keywords:** Artificial Intelligence (AI); information extraction; Named Entity Recognition (NER); unstructured data

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Summary

Forbes statistics [1] state that the amount of data produced through daily transactions is very high, and 80% of this daily generated data is unstructured. Unstructured data include Portable Document Formats (PDF), emails, official letters, and many more.

Unstructured data are a valuable asset to the organization as they have a lot of information hidden in them. If an organization extracts these key insights and uses them for the decision-making process, it can significantly increase its operational efficiency [2]. However, manual processing and extracting key insights from such numerous and complex unstructured documents is naturally time-consuming and error-prone. Hence, developing an artificial intelligence (AI)-enabled tool for automatic key information extraction from unstructured data is a promising and upcoming research focus [3]. However, automatic extraction of key insights from unstructured document research faces certain key challenges [4,5]. One of the most fundamental and critical challenges is to obtain a high-quality,

standard, and annotated unstructured document dataset. The dataset is the foremost important entity for machine learning model training. Model robustness and accuracy depend on its learning from the training data. Therefore, it is always necessary to include variations in training data so that the model learns to recognize the unknown data. Standard datasets or publicly available datasets in this research face key challenges discussed as follows:

- Publicly available datasets consist of poor-quality, blurred, skewed, and low-resolution document images, leading to poor text extraction [6].
- Publicly available datasets are obsolete, consisting of the old or obsolete formats of documents [7].
- Publicly available datasets are domain-specific and task-specific. For example, the dataset proposed in [8,9] is used for the healthcare domain, and the dataset proposed in [10,11] is used for the legal contract analysis domain. In addition, they are used for specific tasks such as metadata extraction from scientific articles [12] or patient detail extraction from the clinical dataset [13].
- Publicly available datasets are unlabelled. Therefore, manual labeling or annotating the dataset is time-consuming and tedious [4,5].

Due to these challenges, a few research studies proposed a custom dataset. However, the custom datasets are private and face confidentiality issues [4,14]. The custom dataset also includes documents with similar layouts or formats. Providing similar layout documents restricts the generalizability of key extraction tasks. Few recent studies highlight the significance of the template-free processing of unstructured documents [15,16]. Therefore, to encourage and advance automatic key information extraction research, we developed a dataset. Figure 1 shows an overview of key information extraction tasks from the invoice document. Our multi-layout invoice document dataset (MIDD) dataset contains 630 invoices with four different layouts of different suppliers. The dataset is of high-quality document images, which leads to high accuracy in text extraction. The dataset also helps to generalize the AI-enabled model as it comprises varied and complex layouts of documents. Table 1 summarizes our MIDD dataset specifications.

**INVOICE**

**Invoice to:**  
Santosh Naik,  
18, MG Road,  
Mumbai, 422654

**Invoice#** 4521484  
**Date** 21 / 12 / 2021

SL	Item Description	Price	Qty.	Total
1	144 Hz Screen	12500	1	12500
2	Speakers	12500	1	12500

**Payment Info:**  
GSTIN: 12345678910111  
BANK A/C No: 432785321  
PSC Code: 1800000000

**Sub Total:** 25000  
**Tax:** 18.00%  
**Total:** 29500

**Terms & Conditions**  
1. Goods once sold will not be taken back.  
2. Our Responsibility ends once the goods leave our premises.

Thank you for your business

Authorized Sign

SL	Description	Data
1.	Name	Santosh Naik
2.	Invoice Number	4521484
3.	Date	21/12/2021
4.	Account Number	432785321
5.	GSTIN	12345678910111

**Figure 1.** Key information extraction task from sample invoice document.

**Table 1.** Specification of multi-layout invoice document dataset (MIDD) dataset.

Subject Area	Natural Language Processing, Artificial Intelligence
More precise area	Named Entity Recognition
Type of files collected	Scanned invoice PDFs
Method of data acquisition	Scanner
Provided file format in our dataset	IOB and manually annotated
Source of data acquisition	Different supplier organizations
Number of files	630
Number of layouts	4

### 1.1. Research Purpose/Goal of Multi-Layout Invoice Document Dataset (MIDD)

- To provide the annotated and varied invoice layout documents in IOB format to identify and extract named entities (named entity recognition) from the invoice documents to the researchers working in this domain. Obtaining a high-quality and sufficient annotated corpus for automated information extraction from unstructured documents is the biggest challenge researchers face.
- To overcome the limitations of rule-based and template-based named entity extraction from unstructured documents traditionally used so far in information extraction approaches. Template-free processing is the only key to processing, and managing a huge pile of unstructured documents in the recent digitized era.
- To provide varied invoice layouts so that researchers can develop a generalized AI-based model that will train on various unstructured invoice layouts. Obtained structured output can later be utilized for integrating into information management application of the organization and used for the decision-making process.

### 1.2. Related Datasets

The research study [6,16–18] proposed key field extraction from a scanned receipts dataset named the ICDAR (International Conference on Document Analysis and Recognition) SROIE-2019 dataset. It has 1000 scanned receipt images with similar layouts, including 876 annotated receipts with labels such as the name of a company, address, date of receipt, and total amount.

The research study [19] used the RVL-CDIP dataset that includes scanned document images of different categories, including invoices as one of the categories. It has 25,000 images of every category. However, the dataset is obsolete and of poor-quality scanned documents.

A few research studies [20,21] built a custom invoice dataset for key field extraction tasks. However, these datasets are not publicly available to researchers due to privacy and confidentiality issues in invoice documents.

A few research studies [8–10,12] proposed information extraction tasks on various domain-specific and task-specific datasets such as the I2b2 2010 (Informatics for Integrating Biology and the Bedside) clinical notes dataset, the MIMIC-III (Medical Information Mart for Intensive Care) dataset, the custom-built legal contract document dataset, and the GROTOAP2 dataset.

## 2. Data Description

To the best of our knowledge, our dataset is the first publicly available multi-layout invoice document dataset. The proposed MIDD dataset includes invoices of different layouts collected from different supplier organizations. Developing generalizability and model robustness are the main aims of collecting the highly diverse and complex invoice layouts. In addition, researchers may use training and testing samples from the dataset as per their requirements. Table 2 represents the details of varied layouts and the total number of invoice document PDFs collected for each layout. Table 2 also shows label naming conventions used while manually annotating the invoices using a UBIAI tool (Release 4.6.2021). There are 11 labels used in invoice annotations. Labels used are Supp\_N for Supplier name, Supp\_G for Supplier GST, BUY\_N for Buyers Name, BUY\_G for Buyers

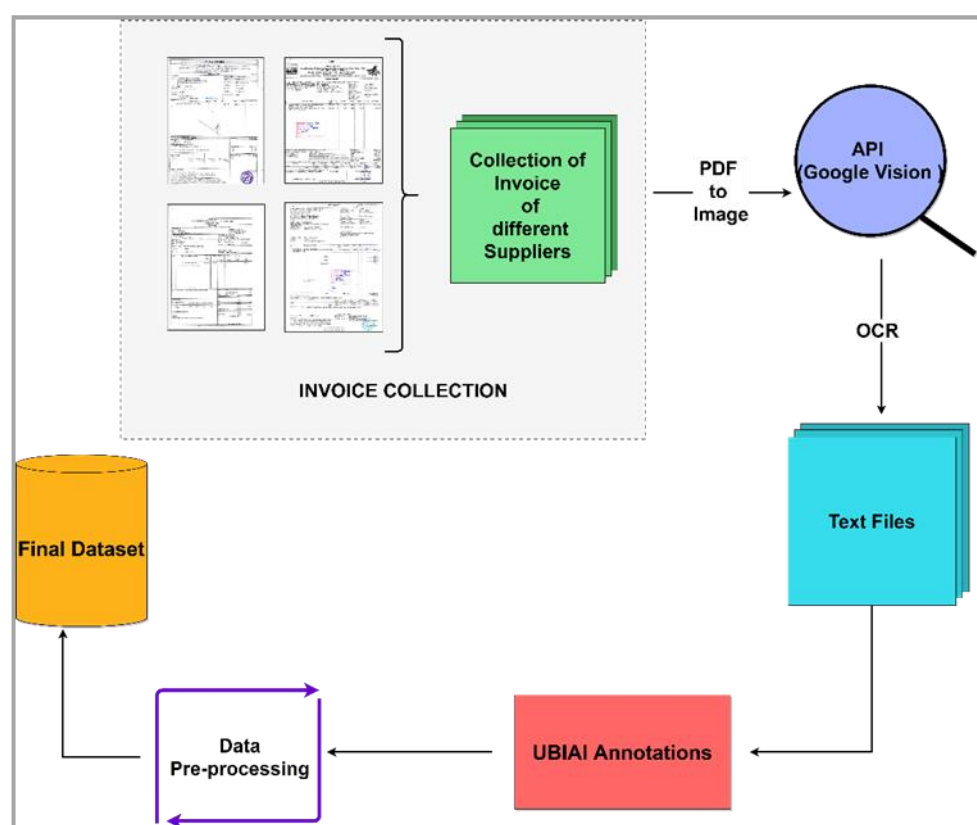
GST, GSTL for GST Label, INV\_NO for Invoice Number, INV\_L for Invoice Number Label, INV\_DL for Invoice Date Label, INV\_DT for Invoice Date, GT\_AMTL for Grand Total Amount Label and GT\_AMT for Grand Total Amount. Our dataset includes four different layout invoices from different supplier organizations. The number of scanned invoice PDFs for layout 1 is 196, layout 2 is 29, layout 3 is 14, and layout four is 391. Thus, our dataset has 630 total scanned invoice PDFs.

**Table 2.** Proposed multi-layout invoice document dataset features.

Layouts	Number of PDFs	Size of Invoices (in MB)	Labels in Dataset
Layout 1	196	164	Invoice Number: INV_NO
Layout 2	29	25.8	Invoice Date: INV_DT
Layout 3	14	23.6	Buyer Name: BUY_N
Layout 4	391	353	Supplier Name: SUPP_N
			Buyer GST Number: BUY_G
			Supplier GST Number: SUPP_G
Total	630	566.4	Grand Total Amount: GT_AMT

### 3. Methods

Figure 2 shows the detailed process of our multi-layout invoice document dataset creation.



**Figure 2.** Process of multi-layout invoice document dataset creation.

#### 3.1. Data Acquisition

Invoices from different supplier organizations are acquired to obtain variations in layouts of invoices. All the supplier invoices are scanned using HP Laserjet M1005 MFP Printer and Scanner (HP: Pune, India) in PDF Format. Each supplier has its unique and own layout or format of their invoices. The varied and multiple layouts of invoices later will help researchers in the template-free processing of various unstructured documents. As shown in Table 2, four different layout invoices from different supplier organizations

are collected. The number of scanned invoice PDFs for layout 1 is 196, layout 2 is 29, layout 3 is 14, and layout four is 391. Thus, our dataset has 630 total scanned invoice PDFs.

Figure 2 illustrates that, as scanned PDFs of invoices are collected, each of these scanned PDFs is converted into an image as the initial step. Conversion from PDF to image is required as any optical character recognition (OCR) engine takes the image as its input for text detection and extraction. Later, all the corresponding text files obtained after OCR are input for the data annotation tool. The Google Vision OCR (Version 2.3.2) engine is used for text extraction, and the UBIAI annotation tool is used for manual annotations. Google Vision OCR output is validated manually by cross verifying the extracted text and original text contents. Roughly 30% of invoices from each invoice layout were manually verified and validated against original scanned PDF contents for evaluating OCR accuracy. As a result, it was observed that Google Vision OCR gives 90% accuracy in text extraction for each supplier invoice layout. The MIDD dataset was also evaluated using AI approaches such as BiLSTM and BiLSTM-CRF for key fields extraction tasks [22].

Figure 3 shows the sample invoices of all four different layouts from our dataset. Different key fields are colored with different colors to understand the meaning of “different layouts of invoices.” For example, the invoice number key field shown with the green rectangular surround box is positioned at different locations in all four layouts of invoices collected from different suppliers. Likewise, all other invoice key fields take a different position as the supplier organization changes. Thus, our dataset is useful for dealing with a real-world situation where organizations have their own unique format or layout of unstructured documents.

Table 3 summarizes more statistical information on the proposed dataset. Invoices collected for MIDD belong to the construction firm having different suppliers for different material purchases.

**Table 3.** Statistical summary of MIDD.

Period of invoice collection	From March 2015 to October 2020 for all supplier organizations
Type of supplier products/items in invoices	Building construction material such as doors, cement, glass
Number of words per invoice page	350 words per page on average. (One .csv file for one invoice page)
Number of named entities labeled for each invoice	11 labels in each invoice, including entity heading name and actual value of that entity. (For example, INV_DL for Invoice Date Label, INV_DT for actual Invoice Date value)
Size of one invoice PDF of any layout	500 KB minimum 1.5 MB average 3 MB maximum
Image quality of invoice PDF	300 dpi
Software used during MIDD construction	Python-3 Jupiter Notebook, Google colab, wand library (version 0.6.6) for converting invoice PDF to image Google vision OCR (version 2.3.2) for text extraction from image UBIAI Framework for NER annotations
Hardware used during MIDD construction	HP Laserjet M1005 MFP Printer and Scanner and HP Pavilion Laptop AMD RYZEN NVIDIA GEFORCE GTX card



Figure 3. Sample invoices of all four layouts from our dataset.

### 3.2. Data (Named Entities) Annotation

Converted text files are manually annotated. UBAI is a convenient and simple-to-use text labeling/annotation tool used for most Natural Language Processing (NLP) tasks such as named entity recognition (NER). By selecting the “free version” of the UBAI package under the “Package” option, all decided labels for invoice entities (for example, INV\_L for invoice label and INV\_DT for actual Invoice Date value) to annotate the invoice document are supplied through its interface manually. After providing labels, the converted text files are inputted manually either by the “drag and drop” facility or the “browse” facility of the UBAI interface. After all the text files are inputted, they can be annotated by choosing labels individually and highlighting the respective text within a complete text file, as shown in Figure 4. This also provides an annotation file export facility in multiple formats such as spacy, IOB, and JSON. As shown in Figure 2, the text files obtained after OCR are provided to UBAI with the labels to be annotated. Table 2 shows the name of the labels used to annotate the text files. Finally, the annotated text file of each invoice in IOB format is exported. IOB labels are like part-of-speech (POS) labels, but they signify the inside, outside and beginning of a word. In NER, every word in the text file, also called “token,” is labeled with an IOB label, and then adjacent tokens are joined together depending on their labels. Later IOB files are transformed into a .csv file. Figure 5 shows the sample IOB file of

the invoice. The first column represents the “token,” and the second column represents the IOB tags of the respective tokens.

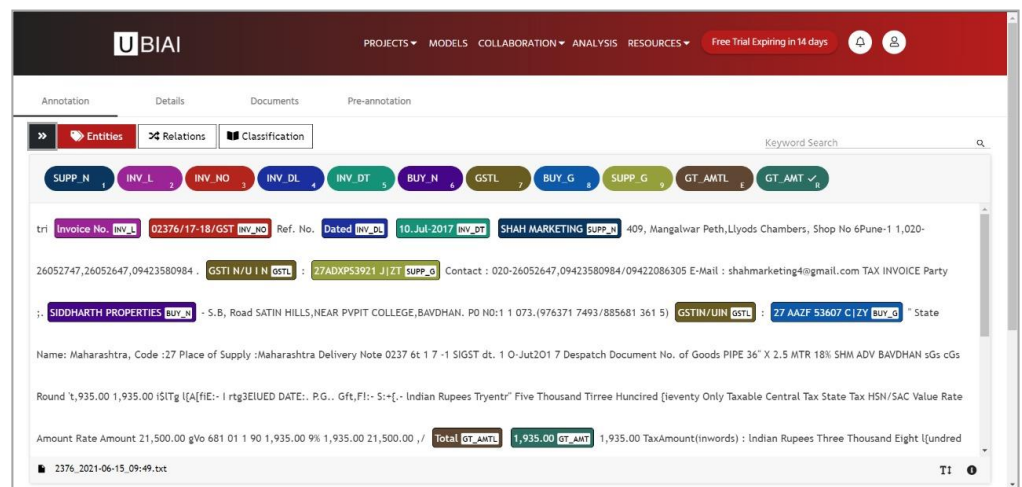


Figure 4. UBI AI interface for manual labeling each converted invoice text file.

	A	B	C
1	Text	Tag	Column1
2	GST	O	
3	INVOICE	O	
4	(	O	
5	ORIGINAL	O	
6	FOR	O	
7	RECIPIENT	O	
8	)	O	
9	SHAH	B-SUPP_N	
10	MARKETING	I-SUPP_N	
11	MANGALWAR	O	
12	PETH	O	
13	Shop	O	
14	No.6	O	
15	Llyods	O	
16	Chamber	O	
17	Barne	O	
18	Road	O	
19	409	O	
20	Mangalwar	O	
21	Peth	O	
22	Pune	O	
23	GSTIN	B-GSTL	
24	/	I-GSTL	
25	UIN	I-GSTL	
26	:	O	
27	27ADXP53921J1ZT	B-SUPP_G	
28	State	O	
29	Name	O	
30	Maharashtra	O	

Figure 5. Sample IOB (.csv) file of the invoice.

### 3.3. Data Pre-Processing

Figure 6 shows the data pre-processing steps carried out to obtain the pre-processed data.

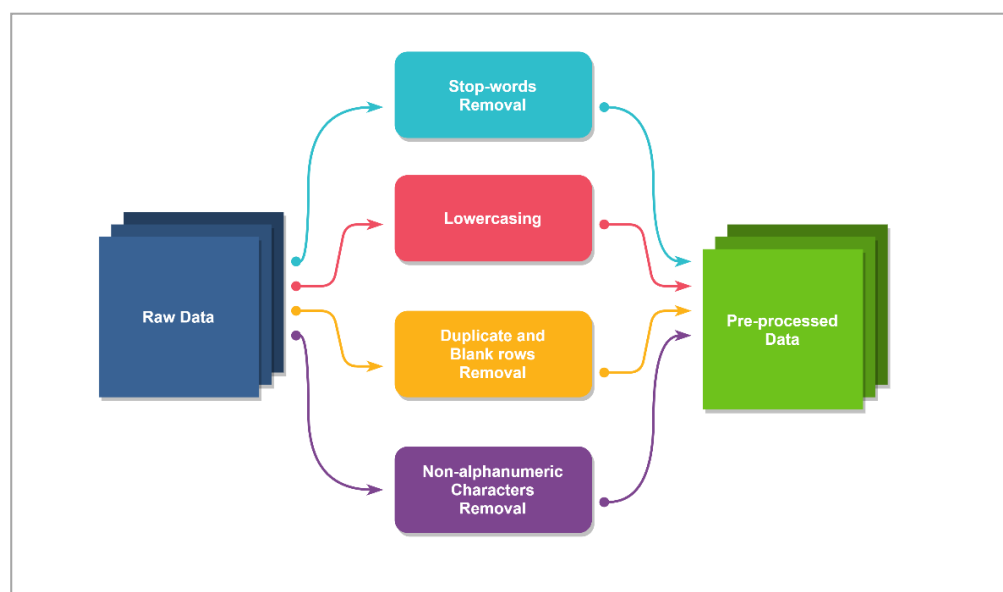


Figure 6. Data pre-processing steps.

*Stop-word removal:* stop-words like English articles (the, an, a) and conjunction (or, and) are removed from the data. Stop-words take up memory size and require processing time. In addition, stop words have no contribution to the meaning of a sentence, so these stop-words are removed from the data.

*Lowercasing:* all the tokens or words are converted into lowercase to get the consistent output.

*Duplicate and blank-row removal:* all the duplicate rows and blank rows are manually removed from the .csv files using the “Remove rows” option in Microsoft Excel.

*Non-alphanumeric characters’ removal:* few non-alphanumeric characters like “(” are removed from the data files. Few non-alphanumeric characters like “:” are kept in the data files as they are part of some fields like Invoice Date.

### 3.4. Practical Applications/Use-Cases of MIDD

Currently, available literature mainly focuses on information extraction from receipts with a similar format or layout. However, in a practical scenario, organizations receive invoices from various suppliers having their unique structure or layout. Publicly available datasets lack an invoice document dataset which has varied invoice layouts.

- The proposed MIDD dataset has many practical implications for extracting named entities as a structured output from the huge pile of unstructured invoice documents. In addition, end-to-end automation of invoice information extraction workflow helps the accounting department in every organization for quick invoice processing and to verify accounts payable and receivable.
- Automated key field extraction from financial documents such as invoices impacts the performance of the business by customer onboarding and verification processes. It can reduce significantly the cost employed for manual data entry and verification of thousands of daily received invoices.

## 4. Conclusions

The proposed work developed a multi-layout invoice document dataset consisting of 630 invoices with four different layouts from different supplier organizations. It contributes to this research by providing a highly diverse, high-quality, and annotated dataset, which is



very useful in Natural Language Processing tasks such as named entity recognition (NER). The dataset size and contents are sufficient to build a generalized AI-based model used for template-free invoice processing for key information extraction tasks. The proposed work presents the statistical summary of the proposed MIDD dataset. It also highlighted the detailed process flow of our dataset creation, which fellow researchers can utilize as the guideline for creating a custom dataset. To the best of our knowledge, our dataset is the first publicly available multi-layout invoice document dataset. The proposed MIDD dataset will be useful for researchers and practitioners working for end-to-end automation across various sectors. This research work unlocks opportunities for researchers working in this area.

## 5. Future Work

1. Increase the data size in MIDD. We aim to increase the number of supplier invoice layouts to achieve more data diversity.
2. Automatic data annotation. Automatic data annotations make the researcher's task simpler and quicker. Therefore, we aim to find a way to annotate invoice documents automatically.
3. Use of pre-trained neural networks. Pre-trained Neural Networks such as BERT and their variants can be utilized on MIDD to evaluate its performance.

**Supplementary Materials:** The following are available online at <https://www.mdpi.com/article/10.3390/data6070078/s1>, MIDD Dataset.

**Author Contributions:** Conceptualization, S.A. and K.K.; methodology, D.B.; investigation, D.B.; resources, S.A., K.K., and D.B.; data curation, D.B.; writing—original draft preparation, D.B.; writing—review and editing, S.A. and K.K.; funding acquisition, S.A. and K.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by “Research Support Fund of Symbiosis International (Deemed University).”

**Data Availability Statement:** Data is contained within the article or Supplementary Material. The data presented in this study are available in Supplementary Material. Authors can use this data for research purposes only by citing our research article.

**Data Citation:** [MIDD] Dipali Baviskar, Swati Ahirrao and Ketan Kotecha. 2021. Multi-layout Invoice Document Dataset (MIDD): A Dataset for Named Entity Recognition.

**Acknowledgments:** We would like to thank Symbiosis International (Deemed University) for providing research facilities. We also want to thank Saarrthi Group Constructions Pvt. Ltd., Pune, for providing invoice documents from different suppliers for our research work.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. 30 Eye-Opening Big Data Statistics for 2020: Patterns Are Everywhere. Available online: <https://kommandotech.com/statistics/big-data-statistics/> (accessed on 5 December 2020).
2. Philosophy, L.; Ahirrao, S.; Baviskar, D. A Bibliometric Survey on Cognitive Document Processing. *Libr. Philos. Pract.* **2020**, 1–31.
3. Baviskar, D.; Ahirrao, S.; Potdar, V.; Kotecha, K. Efficient Automated Processing of the Unstructured Documents using Artificial Intelligence: A Systematic Literature Review and Future Directions. *IEEE Access* **2021**, *9*, 72894–72936. [CrossRef]
4. Adnan, K.; Akbar, R. Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *Int. J. Eng. Bus. Manag.* **2019**, *11*, 1–23. [CrossRef]
5. Adnan, K.; Akbar, R. An analytical study of information extraction from unstructured and multidimensional big data. *J. Big Data* **2019**, *6*, 1–38. [CrossRef]
6. Palm, R.B.; Laws, F.; Winther, O. Attend, copy, parse end-to-end information extraction from documents. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 329–336.
7. Reul, C.; Christ, D.; Hartelt, A.; Balbach, N.; Wehner, M.; Springmann, U.; Wick, C.; Grundig, C.; Büttner, A.; Puppe, F. OCR4all—An open-source tool providing a (semi-)automatic OCR workflow for historical printings. *Appl. Sci.* **2019**, *9*, 4853. [CrossRef]

8. Abbas, A.; Afzal, M.; Hussain, J.; Lee, S. Meaningful Information Extraction from Unstructured Clinical Documents. Available online: [https://www.researchgate.net/publication/336797539\\_Meaningful\\_Information\\_Extraction\\_from\\_Unstructured\\_Clinical\\_Documents](https://www.researchgate.net/publication/336797539_Meaningful_Information_Extraction_from_Unstructured_Clinical_Documents) (accessed on 17 September 2020).
9. Steinkamp, J.M.; Bala, W.; Sharma, A.; Kantrowitz, J.J. Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes. *J. Biomed. Inform.* **2020**, *102*, 103354. [[CrossRef](#)] [[PubMed](#)]
10. Joshi, S.; Shah, P.; Pandey, A.K. Location identification, extraction and disambiguation using machine learning in legal contracts. In Proceedings of the 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 14–15 December 2018; pp. 1–5. [[CrossRef](#)]
11. Shah, P.; Joshi, S.; Pandey, A.K. Legal clause extraction from contract using machine learning with heuristics improvement. In Proceedings of the 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 14–15 December 2018; pp. 1–3. [[CrossRef](#)]
12. Tkaczyk, D.; Szostek, P.; Bolikowski, L. GROTOAP2—The methodology of creating a large ground truth dataset of scientific articles. *D-Lib Mag.* **2014**, *20*, 11–12. [[CrossRef](#)]
13. Yang, J.; Liu, Y.; Qian, M.; Guan, C.; Yuan, X. Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding. *Appl. Sci.* **2019**, *9*, 3658. [[CrossRef](#)]
14. Eberendu, A.C. Unstructured Data: An overview of the data of Big Data. *Int. J. Comput. Trends Technol.* **2016**, *38*, 46–50. [[CrossRef](#)]
15. Davis, B.; Morse, B.; Cohen, S.; Price, B.; Tensmeyer, C. Deep visual template-free form parsing. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; pp. 134–141. [[CrossRef](#)]
16. Zhao, X.; Niu, E.; Wu, Z.; Wang, X. Cutie: Learning to understand documents with convolutional universal text information extractor. *arXiv* **2019**, arXiv:1903.12363.
17. Patel, S.; Bhatt, D. Abstractive information extraction from scanned invoices (AIESI) using end-to-end sequential approach. *arXiv* **2020**, arXiv:2009.05728.
18. Huang, Z.; Chen, K.; He, J.; Bai, X.; Karatzas, D.; Lu, S.; Jawahar, C.V.V. ICDAR2019 competition on scanned receipt OCR and information extraction. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 1516–1520.
19. Kerroumi, M.; Sayem, O.; Shabou, A. VisualWordGrid: Information extraction from scanned documents using a multimodal approach. *arXiv* **2020**, arXiv:2010.02358.
20. Palm, R.B.; Winther, O.; Laws, F. CloudScan—A Configuration—Free Invoice Analysis System Using Recurrent Neural Networks. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR); Kyoto, Japan, 9–15 November 2017, Volume 1, pp. 406–413.
21. Liu, W.; Zhang, Y.; Wan, B. Unstructured Document Recognition on Business Invoice. 2016. Available online: <http://cs229.stanford.edu/proj2016/report/LiuWanZhang-UnstructuredDocumentRecognitionOnBusinessInvoice-report.pdf> (accessed on 18 November 2020).
22. Baviskar, D.; Ahirrao, S.; Kotecha, K. Multi-layout Unstructured Invoice Documents Dataset: A dataset for Template-free Invoice Processing and its Evaluation using AI Approaches. *IEEE Access* **2021**, *1*. [[CrossRef](#)]