Unlocking the Potential of Unstructured Data in Finance Through Document Intelligence

Sriranjani Ramakrishnan
sriranjani.ramakrishnan@aexp.com
American Express, AI Labs
India

Himanshu Sharad Bhatt
himanshu.s.bhatt@aexp.com
American Express, AI Labs
India

Sachin Raja
sachin.raja@research.iiit.ac.in
IIIT-Hyderabad
India

C. V. Jawahar
Jawahar@iiit.ac.in
IIIT-Hyderabad
India

ABSTRACT

With the recent advancements, organizations have brought data to the forefront of their digital transformation journeys. Financial services industry is also moving towards adopting data-driven strategies for improved and faster decision making and providing enhanced customer experience. While advances generally in Artificial Intelligence (AI) and specifically in Machine Learning (ML) have fueled a lot of analytics, it is largely restricted to structured data as it is well organized and is easy to work with. This tutorial presents the opportunities to unlock the potential in unstructured documents in financial domain. These forms of data are more challenging to interpret, but can deliver a more comprehensive and holistic understanding of the bigger picture. While there are challenges around processing such document, ability to quickly make decisions by leveraging such data can provide differentiated value propositions and competitive benefits. This tutorial start with select problems & challenges in Document AI space and use cases involving such documents, show the business opportunities present and describe the technical challenges involved. Subsequently, we discuss techniques and algorithms for several document processing requirements and real world applications.

KEYWORDS

Document Intelligence,Information Extraction,Table Struc- ture Recognition,Document Structure & Component Understanding, Tampering & Authenticity, Document Visual Question Answering

ACM Reference Format:

Sriranjani Ramakrishnan, Himanshu Sharad Bhatt, Sachin Raja, and C. V. Jawahar. 2024. Unlocking the Potential of Unstructured Data in Finance Through Document Intelligence. In *7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD) (CODS-COMAD 2024), January 04–07, 2024, Bangalore, India.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3632410.3633293

CODS-COMAD 2024, January 04–07, 2024, Bangalore, India

ACM ISBN 979-8-4007-1634-8/24/01…$15.00
https://doi.org/10.1145/3632410.3633293

1 INTRODUCTION

According to projections [16], [13], 80% of worldwide data will be unstructured by 2025. Financial services (FS) industry is no different, where most enterprises hold vast array of unstructured data which is largely under-analyzed. A huge amount of enterprise information flows through documents and thus, understanding documents and extracting relevant information is at the heart of digital transformation journeys for the organizations. Documents may be of different types and formats including native PDFs or scanned images, structured, semi-structured of unstructured which makes document processing and understanding an arduous task. An ability to automate document processing and understanding can deliver a more comprehensive and holistic benefits to several applications and use-cases involving manual handling of these documents. In this tutorial, we focus on Financial Services (FS) industry and how Document Intelligence i.e., AI powered automated analysis of documents, allows to tap into the opportunities by analyzing huge amount of information present in such documents.

In financial services industry, documents include financial statements, invoices, bank statements, policies, contracts, marketing creatives etc. Data residing in such documents can be of variety types including images, tables, figures, and text. While there are challenges around processing documents, ability to quickly make decisions by leveraging such data can provide differentiated value propositions and competitive benefits. These benefits include improved operational excellence, automated compliance, or regulatory workflows, discovered insights from mining/ matching disparate data sources and overall enhanced customer experience. However, the very nature of unstructured data prohibits the direct application of AI/ML techniques that can be seamlessly applied on the structured data. This talk will present the arts and sciences behind developing Document Intelligence solutions covering select use cases involving semi or unstructured documents, show the business opportunities present and describe the technical challenges involved. Subsequently, we provide an outline to develop various Document Intelligence solutions that can aggregate, query, analyse, and accelerate the understanding of such data to unveil deep insights across Financial Services use-cases.

2 CONTENT OUTLINE

The tutorial aims to impart knowledge and understanding of AI/ML techniques in the Document Intelligence space including discovery

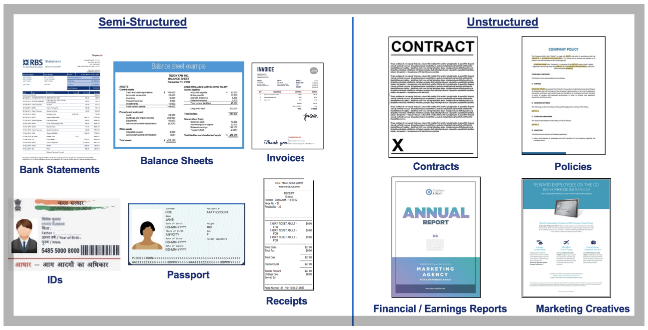

Figure 1: Categorization of documents into semi-structured and unstructured documents from the financial domain

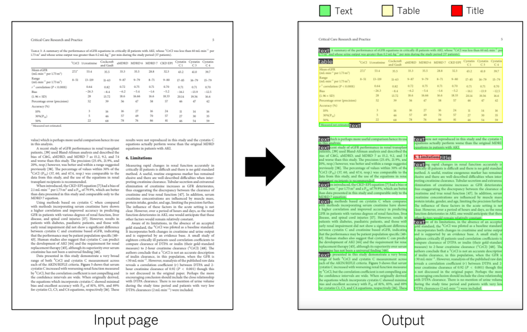

Figure 2: (a) is a table from PubTabNet dataset. In (b), red lines denote the predicted structure and blue lines depict the true structure.

of different types of document sources and their inherent potential, how to analyse & interpret unstructured documents such as financial & bank statements, marketing documents, invoices, receipts etc. We will cover specific cutting-edge techniques to perform information extraction on documents, including semi-structured or unstructured documents, which helps transform the information into a structured format. This tutorial will introduce multiple real-world applications that are being powered by Document Intelligence such as credit ratings, under writings, improve the competitive positioning, mitigate compliance and regulatory risks. The talk will also demonstrate tangible benefits and financial business impacts driven by some of the document intelligence solutions in the financial world. The tutorial will also provide an opportunity to the participants to get familiar with some of the open questions, research challenges, and opportunities in this space. Overall, we expect that our tutorial will educate the participants and will empower them to find answers to the following questions in this context:

(1) What are different types of documents and their associated challenges & potential for the financial industry?
(2) What are the Document AI/ML techniques leveraged by organizations to efficiently process and understands documents?
(3) What is the real business value/impact that the information from these documents can deliver for the organizations?
(4) What are some of the upcoming trends and open challenges?

To answer the first question, we categorize different documents types as 1) Semi-structured and 2) Unstructured documents and further analyse documents in financial services industry using the above taxonomy, as shown in Figure 1. Document which follow a generic layout or template such as invoices, receipts, bank statements, payslips, KYC/IDs, tax forms etc. are examples of semi-strcutured documents. On the other hand, documents which are lengthy and comprises a lot of textual information such as contracts, financial statements, policies, marketing creatives etc. falls in the unstructured document category. In this tutorial, we will discuss about different document types, their challenges and opportunities associated with each one of those.

Subsequent section of the tutorial will cover how organizations address the document AI requirements either using existing vendor or cloud based solutions or manually. Specifically, this tutorial provides an overview of existing document handling techniques for several real-world applications in finance domain.

3 CHALLENGES WITH INTELLIGENT DOCUMENT PROCESSING

This section of the tutorial will highlight some of the challenges for large document processing. We also present recent advances to addresses the associated challenges in the space of document intelligence for financial services. This section aims to motivate the audience with a high-level literature review and then go deep into one or two research directions and methods to address different problems.

3.1 Document Structure & Component Understanding

Documents can be of different types and formats including native PDFs, scanned images, web-pages, emails and can follow different templates, which makes document processing and understanding an arduous task. Techniques which localize and identify components [24], [18] such as tables, figures, title, text and lists can facilitate better understanding of documents while parsing it into a structured format. Figure 2 shows an illustrative example of the components identified in an unstructured document. The tutorial will cover some of the recent advances in object detection algorithms and graph neural networks that are lately being used for document component understanding. The tutorial will present some of the open research directions to further include the textual information and neighbourhood or proximity to enhance the document component detection.

3.2 Document Information Extraction

Extracting information from unstructured documents, specially financial documents, in a structured format (i.e. key-value pairs) and analysing them unlocks the opportunities to make faster and confident decisions, improve operational efficiency by making extracted structured information available to downstream applications. Financial documents are very different from traditional natural language English corpus. These could be semi-structured such as bank statements, balance sheets or unstructured such as earnings report, annual reports etc. Therefore, traditional information extraction techniques do not generalize well on financial documents which are loaded with statements conveying facts and claims using a vocabulary rich in currency, integral or floating-point values, and date-time types etc. Mostly, organizations end up employing human resources to examine and analyze these financial documents. Since financial documents could be lengthy (anywhere between 200-300 pages) and captures a huge amount of factual or numeric details, it becomes very time-consuming, error-prone and laborious task to manually extract and analyse such information.

There is a need for an efficient and automated technique that can extract relevant information from financial documents in a structured manner i.e., an entity (value) and its description (key) as key-value pairs. Such an automated solution could have two practical benefits. Firstly, a highly accurate system, which extracts important entities and their descriptions from financial documents, for currency, integral or floating-point values, and date-time types, will minimize the chances of missing important information that could lead to incorrect analysis and decision making which could lead to possible compliance and regularity risk. Secondly, a document tagged with relevant entities and their descriptions will improve the analyst's speed to interpret and analyse the documents for specific business requirements.

The tutorial will cover technique for information extraction from both semi-structured and unstructured documents. It will cover advances in modelling interactions between text and layout information along with image features [20] [21] for semi-structured documents to extract relevant information as key-value pairs. The tutorial also covers presents information extraction as a question answering framework [19], [15] for unstructured documents. Finally, the tutorial will also refer and point the audience to some of the datasets which are based on publicly available financial documents. The dataset consists of paragraphs with tagged values (entities), and their keys (descriptions). We believe it will further initiate new research directions, especially in the financial domain.

3.3 Table Structure Recognition

Table are the most commonly used structural representation that organizes data into rows and columns. It captures structural and geometrical relationships between different elements and attributes in the data. Table extraction refers the process of converting table regions in unstructured format (such as PDF, image, etc.) to a structured format (such as csv, excel, JSON, etc.). An intermediate representation in the table extraction algorithms is the row and cell separations and associating the text to the correct cell, as shown in Figure 3 & 4. This further leads to a structured format which could be easily consumed by the downstream applications involving table analysis and understanding. This tutorial will cover techniques for major tasks in table extraction [10], [14], [22] i.e. i) detecting the presence of a table in an image ii) localizing the table in the image iii) decoding the structural relational among table cells and iv) understanding the text inside each cell.

One of the important, however relatively less explored, aspects in table extraction is the availability of a standard measure to report

	Hazard Ratio (HR) * (95% CI)	P value
Age (≥ 64 vs. < 64)	1.97 (0.87-4.44)	0.105
Sex (male vs. female)	1.13 (0.55-2.36)	0.738
Hypertension (yes vs. no)	1.22 (0.56-2.69)	0.619
Diabetes mellitus (yes vs. no)	2.16 (1.04-4.49)	0.039
Dyslipidemia (yes vs. no)	0.97 (0.46-4.05)	0.940
Smoking (yes vs. no)	0.74 (0.33-1.70)	0.465
Time from stroke onset to admission (> 24 hours vs. ≤ 24 hours)	1.57 (0.73-3.39)	0.248
Qualifying event (Stroke vs. TIA)	1.22 (0.29-5.15)	0.783
NIHSS at admission (≥ 4 vs. < 4)	1.72 (0.75-3.91)	0.198
mRS at discharge (2-5 vs. 0-1)	2.06 (0.84-5.09)	0.117
IMCAD (occlusion vs. ≥ 50% stenosis)	1.09 (0.51-2.35)	0.819

Figure 3: A sample table from a medical tables dataset.

	Hazard Ratio (HR) * (95% CI)	P value
Age (≥ 64 vs. < 64)	1.97 (0.87-4.44)	0.105
Sex (male vs. female)	1.13 (0.55-2.36)	0.738
Hypertension (yes vs. no)	1.22 (0.56-2.69)	0.619
Diabetes mellitus (yes vs. no)	2.16 (1.04-4.49)	0.039
Dyslipidemia (yes vs. no)	0.97 (0.46-4.05)	0.940
Smoking (yes vs. no)	0.74 (0.33-1.70)	0.465
Time from stroke onset to admission (> 24 hours vs. ≤ 24 hours)	1.57 (0.73-3.39)	0.248
Qualifying event (Stroke vs. TIA)	1.22 (0.29-5.15)	0.783
NIHSS at admission (≥ 4 vs. < 4)	1.72 (0.75-3.91)	0.198
mRS at discharge (2-5 vs. 0-1)	2.06 (0.84-5.09)	0.117
IMCAD (occlusion vs. ≥ 50% stenosis	1.09 (0.51-2.35)	0.819

Figure 4: Intermediate representation with row & cell separation.

the performance for different algorithms. Unlike traditional machine learning problems where the output is a class (classification) or number (regression), the outcome of a table parsing algorithm is always a structure. There needs to be a way to compare one structure against another structure and define some measure of "similarity/distance" to evaluate different methods. A number of measures quantifying this "distance" have been proposed in literature and multiple competitions. However, existing metrics, such as Adjacency Relation (Text) [8], Adjacency Relation (IOU) [6] & Tree-Edit-Distance Based Similarity with IOU *(TEDS-IOU)*, evaluate the performance of table extraction algorithms using both the structural and textual information. This tutorial will discuss the limitations of existing metrics due to their dependence on the textual information. We highlight that considering textual information introduces additional dependency on the OCR (text detection/recognition), which is a separate area in itself and should not be included in evaluating how good is the detected table structure.

3.4 Document Authenticity Verification

Verifying document authenticity [7], [23], [5] is an under-researched but highly important aspect of document analysis system. While several important decisions are based on the information extracted from the documents, it becomes highly imperative to ensure that the documents are genuine, legitimate and have not been tampered in anyway, especially the financial and ID documents. Researchers have proposed approaches to address the authenticity of handwritten text [2] and validate the forensic characteristics of the documents [3]. There are several approaches which proposed to validate the authenticity of the documents by detecting tampering or fraudulent characteristics in the documents [4], [9]. However, these approaches rely on the imperfections in font, style, and position of the text which is introduced due to tampering. There are also several off-the-shelf solutions available to validate the authenticity of KYC/ID documents [1] based on the bar-code information and validating native PDF document based on the PDF header information or the signature of the documents. However, due to deep penetration of mobile (scanning) devices, a huge fraction of the documents are scanned (image) documents where there are limited capabilities for authenticity verification. This tutorial will present a broad understanding of techniques and methods developed for verify document authenticity and gaps that the research community needs to emphasize on.

3.5 Document Visual Question Answering

The tutorial will present recent advances in an upcoming research area of document visual question answering [17], [11]. DocVQA [12] expands the horizon of document understanding from passing a document image through OCR to understanding all types of information conveyed by a document to enable high-level semantic tasks. The tutorial will cover the datasets available for this task and the techniques build to answer a natural language question on a single documents or a collection of digitised documents.

4 CHALLENGES

The tutorial will cover **challenges, expectations** and how it is been handled in finance industry with use-cases. With every new customer on-boarding and handling existing documents in organizations, the process leads to multiple challenges in intelligent document processing such as:

- Multiple types, layouts, format leading to complexity in classification, information extraction
- Noise in document image, rotation, document quality leading to handle image quality assessment or enhancement before processing it further
- Data can span across multiple pages complicating the retrieval data from documents
- Data can varying across languages handling multi-language documents with existing algorithms poses challenge
- OCR based output might not be accurate leading to lot of post-processing
- Security access, privacy issus, technical stack, cloud-based infrastructure leading deployment issues and delayed inference time in products.

Another global challenge for ubiquitous document analysis and recognition (DAR) system is the varying **international regulatory and compliance guidelines**. Data compliance and governance is a necessity for the financial services industry in order to maximize the potential of their data. With regulatory scrutiny due to GDPR, CCPA, PII, HIPAA, AML and KYC regulations, many institutes suffer penalties without being compliant. Hence business must be data compliant as per regulations, handle security and privacy of data using data governance, which includes organizational policies and process to manage data within the organization and for all AI models. For high risk decision making models, these becomes utmost important. This tutorial will cover on aspect on how data/model governance and compliance is handled using intelligent document processing. How document processing can help in handling compliant issues for many use-cases.

5 DISCUSSION ON OPEN RESEARCH DIRECTIONS

To conclude, the tutorial will present some of the open questions and potential research directions in document AI space. The tutorial will discuss about the next steps needed to move away from using traditional manual or rule based approaches for document understanding and extraction. The tutorial will compare and contrast some the available cloud-based document AI solutions from Google, Microsoft and Amazon to help the audience make conscious decisions about the choices for their problem. The tutorial will present some of the advanced research areas in document AI such as document tampering checks. Since a lot of important decision are being made based on the information extracted from the documents, it becomes a very important problem in document AI to ensure that the documents are legitimate and not tampered. This tutorial will touch upon some upcoming research direction which are at the intersection of object detection, image processing, natural language processing, machine Learning and deep learning. These include multi-object detection to efficiently segment document images scanned on the same page, de-noising document images and removing watermarks that could hamper the text extraction from documents, identifying font/style change detection on documents that could raise regulatory or brand concerns etc. The tutorial will also discuss on future trends on powering Generative AI with Intelligent Document Processing. This helps in increased efficiency, accuracy, enhanced data insights, data compliance, cost savings etc. This topic includes leveraging generative AI Large Language models for documents and some of its potential use-cases like synthetic document generation, document summarization, semantic search in documents, interacting with documents etc.

6 CONCLUSIONS

This tutorial will provide an opportunity to bring researchers in document AI space together to discuss about the opportunities in this space, to overcome the existing challenges for developing AI/ML solutions, and to unveil the potential value hidden in the huge pool of such unstructured documents sources. The tutorial will also focus on demonstrating some of the practical use-cases of document analysis systems the value it can draw potential financial business impacts. Finally, the tutorial will present some of the open questions, research challenges, and opportunities in this space.

REFERENCES

[1] [n. d.]. Valida Gradiant. https://www.gradiant.org/en/portfolio/valida/. Accessed: 2022-04-01.

[2] Ricardo da Silva Barbosa, Rafael Dueire Lins, Edson Da F. De Lira, and Antonio Carlos A. Camara. 2014. Later Added Strokes or Text - Fraud Detection in Documents Written with Ballpoint Pens. In *2014 14th International Conference on Frontiers in Handwriting Recognition*. 517–522. https://doi.org/10.1109/ICFHR.2014.93

[3] Ricardo da Silva Barboza, Rafael Dueire Lins, and Darlisson Marinho de Jesus. 2013. A Color-Based Model to Determine the Age of Documents for Forensic Purposes. In *2013 12th International Conference on Document Analysis and Recognition*. 1350–1354. https://doi.org/10.1109/ICDAR.2013.273

[4] Romain Bertrand, Oriol Ramos Terrades, Petra Gomez-Krämer, Patrick Franco, and Jean-Marc Ogier. 2015. A Conditional Random Field model for font forgery detection. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. 576–580. https://doi.org/10.1109/ICDAR.2015.7333827

[5] Francisco Cruz, Nicolas Sidère, Mickaël Coustaty, Vincent Poulain D'Andecy, and Jean-Marc Ogier. 2017. Local Binary Patterns for Document Forgery Detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 01. 1223–1228. https://doi.org/10.1109/ICDAR.2017.202

[6] Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. 2019. ICDAR 2019 competition on table detection and recognition (cTDaR). In *2019 ICDAR*. IEEE, 1510–1515.

[7] Utpal Garain and Biswajit Halder. 2008. On Automatic Authenticity Verification of Printed Security Documents. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*. 706–713. https://doi.org/10.1109/ICVGIP.2008.67

[8] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. 2013. ICDAR 2013 table competition. In *2013 ICDAR*. IEEE, 1449–1453.

[9] Hailey James, Otkrist Gupta, and Dan Raviv. 2020. OCR Graph Features for Manipulation Detection in Documents. *CoRR* abs/2009.05158 (2020). arXiv:2009.05158 https://arxiv.org/abs/2009.05158

[10] Saqib Ali Khan, Syed Muhammad Daniyal Khalid, Muhammad Ali Shahzad, and Faisal Shafait. 2020. Table Structure Extraction with Bi-directional Gated Recurrent Unit Networks. *CoRR* abs/2001.02501 (2020). arXiv:2001.02501 http://arxiv.org/abs/2001.02501

[11] Fen Liu, Guanghui Xu, Qi Wu, Qing Du, Wei Jia, and Mingkui Tan. 2020. *Cascade Reasoning Network for Text-Based Visual Question Answering*. Association for Computing Machinery, New York, NY, USA, 4060–4069. https://doi.org/10.1145/3394171.3413924

[12] Minesh Mathew, Ruben Tito, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2021. Document Visual Question Answering Challenge 2020. arXiv:2008.08899

[13] Paul Nelson. [n. d.]. Search and unstructured data analytics: 5 trends to watch in 2020. https://www.accenture.com/us-en/blogs/search-and-content-analytics-blog/search-unstructured-data-analytics-trends. Accessed: 2021-06-05.

[14] Shubham Paliwal, Vishwanath D, Rohit Rahul, Monika Sharma, and Lovekesh Vig. 2020. TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images. arXiv:2001.01469 [cs.CV]

[15] Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic Graphs for Generating Deep Questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1463–1475. https://doi.org/10.18653/v1/2020.acl-main.135

[16] Adam Rogers. [n. d.]. The 80% Blind Spot: Are You Ignoring Unstructured Organizational Data? https://www.forbes.com/sites/forbestechcouncil/2019/01/29/the-80-blind-spot-are-you-ignoring-unstructured-organizational-data/?sh=551891fd211c. Accessed: 2022-04-01.

[17] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: Knowledge-Aware Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 8876–8884. https://doi.org/10.1609/aaai.v33i01.33018876

[18] Fotini Simistira, Manuel Bouillon, Mathias Seuret, Marcel Würsch, Michele Alberti, Rolf Ingold, and Marcus Liwicki. 2017. ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 01. 1361–1370. https://doi.org/10.1109/ICDAR.2017.223

[19] Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 641–651. https://doi.org/10.18653/v1/N18-1059

[20] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug 2020). https://doi.org/10.1145/3394486.3403172

[21] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201

[22] Wenyuan Xue, Baosheng Yu, Wen Wang, Dacheng Tao, and Qingyong Li. 2021. TGRNet: A Table Graph Reconstruction Network for Table Structure Recognition. arXiv:2106.10598 [cs.CV]

[23] Lin Zhao, Changsheng Chen, and Jiwu Huang. 2021. Deep Learning-Based Forgery Attack on Document Images. *IEEE Transactions on Image Processing* 30 (2021), 7964–7979.

[24] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. PubLayNet: largest dataset ever for document layout analysis. *arXiv preprint arXiv:1908.07836* (2019).