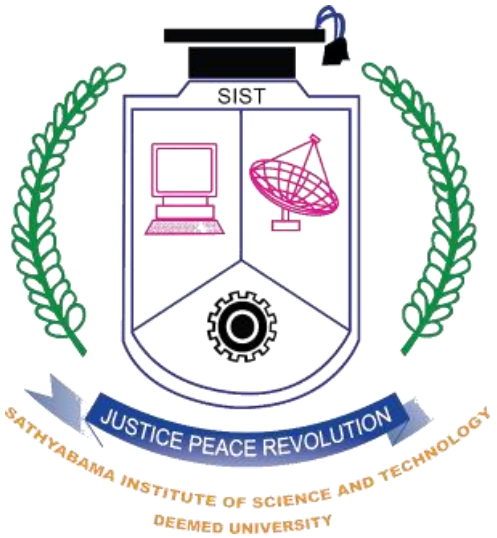
Extraction of Data from Documents

Using AWS Textract

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering
by

ARIGONDA MAHESH (Reg. No - 39110078)
BALINA AKHILESH VISHNU (Reg. No – 39110118)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHISALAI,
CHENNAI - 600119

APRIL - 2023

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with —A‖ grade by NAAC Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **ARIGONDA MAHESH (3911078) and BALINA AKHILESH VISHNU (39110118)** who carried out the Project Phase-1 entitled **"Extraction of Data from Documents Using AWS Textract"** under my supervision from December 2022 to April 2023.

Internal Guide

Dr. JANCY, MCA., M.B.A., M.Tech., Ph.D

Head of the Department

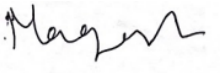
Dr. L. LAKSHMANAN, M.E., Ph.D.

Submitted for Viva voce Examination held on 24.04.23

Internal Examiner **External Examiner**

DECLARATION

I, **ARIGONDA MAHESH (Reg.No- 39110078),** hereby declare that the Project Phase-1 Report entitled **"Extraction of Data from Documents Using AWS Textract"** done by me under the guidance of **Dr.JANCY, MCA., M.B.A, M.Tech.,Ph.D** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.



DATE: 24.04.23
PLACE: Chennai **SIGNATURE OF THECANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph. D**, **Dean**, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.,** Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.Jancy, MCA., M.B.A, M.Tech., Ph.D**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-1 project work.

I wish to express my thanks toall Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

In the digital era of twenty first century, everything is becoming automated, and information is stored and transfer in digital forms. But there are many situations where data is not stored in digital form and it is essential to extract text from those hardcopies to store in digitized form. The latest technology such as Text recognition software has completely changed the process of text extraction using Optical Character Recognition. Therefore, this paper introduces the concept of OCR technology, explains the process of extraction using Amazon Textract tool and current research in the area. Detailed information and working methodology of Amazon Textract. Its comparison with other OCR tools and its scope. This paper will help other researchers in the field to get an overview of the technology.

Optical Character Recognition (OCR) can open up understudied historical documents to computational analysis, but the accuracy of OCR software varies. This article reports a benchmarking experiment comparing the performance of Tesseract, Amazon Textract, and Google Document AI on images of English and Arabic text. English-language book scans (n = 322) and Arabic-language article scans (n = 100) were replicated 43 times with different types of artificial noise for a corpus of 18,568 documents, generating 51,304 process requests. Document AI delivered the best results, and the server-based processors (Textract and Document AI) performed substantially better than Tesseract, especially on noisy documents. Accuracy for English was considerably higher than for Arabic. Specifying the relative performance of three leading OCR products and the differential effects of commonly found noise types can help scholars identify better OCR solutions for their research needs. The test materials have been preserved in the openly available “Noisy OCR Dataset” (NOD) for reuse in future benchmarking studies.

Keywords: Amazon Textract, Optical Character Recognition, Machine Learning, Google AI

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Optical Character Recognition (OCR) is the machine learning tool use for conversion of text or making a digital copy of the text. It takes resources through handwritten documents, printed text, or from natural images. Optical character recognition is a science that enables us to translate various types of documents or images into analysable, editable and searchable data. The objective of this research paper is to study the use of Optical Character Recognition with Amazon Textract tool and its comparison over other OCR tools available in the market. In this Research paper, we collected and analyzed research articles on the topic of OCR technology and closely related topics which were published between year 2015 to 2021. Articles were searched using keywords, forward reference searching and backward reference searching in order to search all the articles related to the topic. OCR is a field that has applications in pretty much every other field like in Health care division, Data Extraction, Data Storage, Banking segment, etc.

Transcribing text from standard documents still represents a significant proportion of manual work in most companies. By using cloud-based OCR tools (AWS Textract, for example), we can intelligently extract key data fields from documents like invoices and other bills. The cloud-based, API-enabled OCR services offer a great balance between price and performance for most companies. The ability to work with multiple formats enables companies to accurately automate away manual transcription work and save man-hours. API-based cloud services are also highly scalable, meaning that any solution that integrates them can scale from a few hundred requests to thousands of requests without any loss in response time.

when the driver is getting lazy. Different considerations have been proposed that around 20% of all street mishaps are fatigue-related, up to 50% on certain streets. Driver weakness could be a critical reason in the large number of vehicle mishaps. Later measurements assess that yearly 1,200 passing's and 76,000 wounds can be credited to weariness related crashes. The improvement of technologies for recognizing or avoiding laziness at the wheel could be a major challenge within the field of accident evasion frameworks. Due to the risk that laziness or fatigue presents on the street, strategies have to be shaped for neutralizing its influences. Both driver tiredness and diversion, in any case, might have the same impacts, i.e., diminished driving execution, longer response time, and an expanded hazard of crash inclusion. Based on Procurement of video from the camera that's before driver, it performs real-time preparing of an approaching video stream in order together the driver‘s level of weariness on the off chance that the laziness is estimated at that point it'll deliver the caution by detecting the eyes, mouth and headposture.

CHAPTER 2

LITERATURE SURVEY

For what reason to involve Amazon Artificial Intelligence and Machine Learning administrations for report robotization?

By utilizing Amazon Web Services Artificial intelligence and Machine Learning administrations to control report handling helps associations and reduce work to each shape type: Amazon Textract gets it and understands records and structures without requiring any broad pre-work to get the structure's design. All things considered, the AI-based approach gets the substance in view of the actual design, in any event, separating the information held in tables or structures and planning that into machine meaningful constructions to demonstrate what has been written in each piece of a structure by planning those qualities to their individual information fields.

- Increase and down depending on the situation: Business tasks are frequently tested by overseeing tops sought after, for instance, during application cut-off times or during occasions like the COVID-19 pandemic. Adaptability and current serverless cloud models are critical, which help rapidly increase to handle enormous volume of reports and afterward proportional down, limiting the continuous expenses right away.
- Join human and AI mastery to affirm or address information passage all the more effectively and rapidly: Tightly incorporated expanded AI banners to a human analyst the parts of structures which the AI couldn't peruse without hesitation. The blend of AI and a human cooperating conveys an exceptionally hearty way to deal with productively mechanizing an archive work process.
- Perceive maintainability benefits: Organizations can lessen the carbon and energy consumed in actually moving huge loads of actual paper reports among locales, and afterward putting away something very similar in actual documents. Moving to electronic archive handling, with advanced sorting rooms ingesting and filtering the media, frees the labor force away from being truly co-situated with the records. A more extensive shift for the labor force that AI brings, is the capacity to depend on the AI for the commonplace

errands and permit the human labor force to zero in on more worth adding undertakings that require interestingly human abilities.

- Remove additional worth from information to further develop processes: Using ML methods likewise increases current standards on how much worth can be extricated from records. Amazon Rekognition is utilized to recognize and separate pictures or charts installed inside archives, saving time and manual exertion by distinguishing and editing out pictures. The text inside reports is handled through administrations Amazon Translate, making it conceivable to help 55 dialects and variations from Afrikaans to Vietnamese, without expecting in-house interpreters. Amazon Comprehend utilizes normal language handling strategies to assist with getting a report. This is much of the time used to emergency inbound correspondence by getting the idea of the solicitation, and guiding the undertaking to the best work line. These can be taken care of straightforwardly into mechanical cycle mechanization driven work processes to some extent or completely attempt work that would require human groups.
- Assemble information bits of knowledge to further develop administrations: Extracted information can be driven into a diagram data set, like Amazon Neptune, for ensuing organization examination. This approach distinguishes application misrepresentation where organizations of partners, locations, and organizations are distinguished from the diagram that may be generally extremely difficult to perceive.

AWS Textract is rethinking the way that organizations cycle records in a Digital World Contemplate the last time you opened a financial balance, applied for protection, or renegotiated your home. It was most likely done on paper. The quantity of reports in a home loan bundle alone is more than 100 pages in length. How would you manage all that paper? For some organizations across an assortment of businesses, including monetary administrations, medical services, and assembling, it is careful to handle these archives. It's manual, slow, costly, and blunder inclined, and information is in many cases spread across unique sources. Subsequently, making and dealing with a record handling pipeline stays a test for some organizations. As indicated by Ritu Jyoti from IDC, "Supporting archive handling requires an AI-local stage that further develops precision, execution, dexterity and adaptability while supporting an expansive arrangement of record types. Fake Processing of Scanned Documents

utilizing AWS Textract Dept of IT, SSGMCE, Shegaon Page 8 Intelligence (AI), can assist with smoothing out archive computerization giving better business results, further developed ROI, and decrease manual efforts."[AWS has sent off an answer for assist associations with separating bits of knowledge and robotize handling records of various arrangements (PDF, Word, crude message) and designs (shots, records) utilizing Amazon Comprehend. This new send-off joins the force of regular language handling (NLP) and Optical Character Recognition (OCR) to assist with decreasing how much pre-handling or post-handling expected to deal with records. You can now utilize specially named substance acknowledgment (NER) on greater archive types without expecting to change your records over to crude text. AWS has been advancing in the wise report handling (IDP) space for a really long time to change over information in records into usable data for archive driven processes. AWS sent off AI administrations like Amazon Textract, Amazon Comprehend, and others to assist with the computerization of separating experiences from reports. Since the sendoff of those administrations, upgrades in precision and speed have been ten times. These administrations offer new APIs like particular help for solicitations and receipts, penmanship and language support, in addition to enhancements in inertness.

AWS Textract is reclassifying the way that organizations interaction archives in a Digital World

Multi-modular transportation is probably the greatest advancement in the coordinated factors industry. There has been an effective coordinated effort across various transportation accomplices in inventory network cargo sending for a long time. Yet, there's as yet an impressive upward of desk work handling for every leg of the outing. A huge number of archives are handled in sea cargo sending alone. Utilizing physical work to deal with these archives (buy orders, solicitations, bills of filling, conveyance receipts, and that's only the tip of the iceberg) is both costly and blunder inclined. We really want to robotize the record handling in the operations business.

Supporting residents through COVID-19: Arizona State University Cloud Innovation Center (CIC)

The Arizona State University Cloud Innovation Center (CIC) assembled an open-source resource for refine the archive handling innovation of Amazon Textract for

service bill and driver's permit information extraction. This arrangement was as of late utilized by Wildfire, a state relationship for Community Action Agencies, and Prefix Health Technologies (Prefix), an AWS Partner Network (APN) Technology Partner, to assist with giving help to residents during the COVID-19 pandemic. The Arizona benefits gateway permits COVID-19 affected families to prescreen and apply for help with lease, contract, gas, electric, and water. Candidates can join record pictures to the advantage applications utilizing their cell phone camera. Amazon Textract catches the information from the pictures and populates or confirms the information entered which takes out the requirement for manual confirmation and paces up the handling time. Generally speaking, not entirely settled at the place of passage and assets are credited to the client's record with practically zero deferral. For extra subtleties on the arrangement created read, "A smoothed out, portable first way to deal with administration conveyance for districts and states" where the arrangement they produced for Arizona came about in 49% of the applications to be consequently supported, in this way lessening the time expected to confirm and appropriate assets.

2.1 INFERENCES FROM LITREATURE SURVEY

Amazon Textract enables applications to integrate with SDK APIs so that the documents or images with textual data from various representations of text in form of raw text, forms, tables are easily extractable. Textract also provides the confidence level / percentage of the extracted text making it a choice for the integrating of the applications to either consider it or neglect it. It has a Scalability issues as human reviews slow down processing.

2.2 OPEN PROBLEMS IN EXISTING SYSTEM

- Requires training large datasets for accuracy.
- Has scalability issues as human reviews slow down processing.
- Using multiple technologies adds to complexity.

Inability to Extract Custom Fields: There could be multiple data fields in a given invoice, say Invoice ID, Due Date, Transaction Date etc. These fields are something that are common in most invoices. But if we want to extract a custom field from an invoice, say, GST number or bank information, Textract does a poor job.

Integrations with upstream and downstream providers: Textract doesn't allow you to integrate with different providers easily, say, for example, we'll have to build an RPA pipeline with a third-party service; it would be difficult to find appropriate plugins that suit Textract.

Ability to define table headers: For table extraction tasks, textract doesn't allow you to define table headers. Therefore, it would be not easy to search or find a particular column or a table in a given document.

No Fraud Checks: Modern OCRs are now able to find if a given document is original or fake by validating dates and finding pixelated regions. AWS Textract doesn’t come with this, its only job is to pick all the text from an uploaded document.

No Vertical Text Extraction: In some of the documents, invoice numbers or addresses can be found in a vertical alignment. At present, AWS only supports horizontal text extraction with a slight in-plane rotation.

Everything’s Cloud: Any document processed with Textract goes into the cloud, only supporting a few regions. More information here. However, some companies

might not be interested in taking their documents to the cloud for reasons like confidentiality or legal requirements. Still, unfortunately, AWS Textract does not support any on-premise deployment for document processing.

Language Limit: Amazon Textract supports English, Spanish, German, French, Italian, and Portuguese text detection. Amazon Textract will not return the language detected in its output.

CHAPTER 3

REQUIREMENT ANALYSIS

3.1 FEASIBILITY STUDIES/RISK ANALYSIS OF THE PROJECT

Cloud security at AWS is the highest priority. As an AWS customer, you benefit from a data center and network architecture that are built to meet the requirements of the most security-sensitive organizations. If the AWS Management Console tells you that you're not authorized to perform an action, then you must contact your administrator for assistance. Your administrator is the person that provided you with your user name and password.

3.2 SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT

- PROCESSOR: Intel Core i5 or a better processor
- OPERATING SYSTEM: Windows, Linux – Ubuntu 22.04.
- MEMORY: 8 GB RAM or higher
- HARD DISK SPACE: Minimum 30 GB or higher
- DATABASE: MySQL database
- CLOUD RESOURCES: AWS Textract
- INTERNET: Must have a good internet connection.

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

4.1 SELECTED METHODOLOGY OR PROCESS MODEL

Other than text and outline data about an archive, one most significant component that should Be separated from reports is information. Manual information extraction is monotonous cycle That could be overwhelming and mistake inclined, also the hardships when archives are checked as pictures and not text.

By and large, information isn't put away in passengers or sentences, yet In even structures and key-esteen matches (KVPs), which are basically two connected information things, key and worth, where the key is utilized as a special identifier for the worth (i.e., Name: Jhon) In particular, while managing records, for example, shapes, these information types

Step 1: Scan the Document

The first step is to scan the document from which the data has to be extracted. Below is the list of some types of documents, but not limited to, from which data can be extracted:

Customary Invoices and Bills

- Monetary Documents
- Clinical Documents
- Manually written Documents and letters

Pay slips or Employees Documents

Make sure paper is put in place properly before scanning the document. Amazon Textract may fail to recognize some part of the document if it is left out of the scanning area.

Step 2: Reading the Data

After the document is appropriately placed for scanning, Amazon Textract starts a virtual scan of the document. The tool basically reads the data. This helps to extract

and map the data at the later stages. This process is almost instantaneous and happens quite quickly.

Step 3: Identifying Key Information

Once a thorough scan is done of the document, Amazon Textract automatically identifies key and vital information that has to be extracted and stored. Since it is based on a deep-learning technology, the identification of the information is very accurate.

Step 4: Matching & Data Integration

Using the JavaScript Object Notation (JSON) format, the data is then extracted and stored. JSON is a standard file and data exchange format that helps the human-readable text to be stored on web servers. Since Amazon Textract is a product of Amazon Web Services (AWS), data can be integrated with other AWS products such as Amazon Comprehend, Amazon DynamoDB and so on.

4.2 ARCHITECTURE / OVERALL DESIGN OF PROPOSED SYSTEM

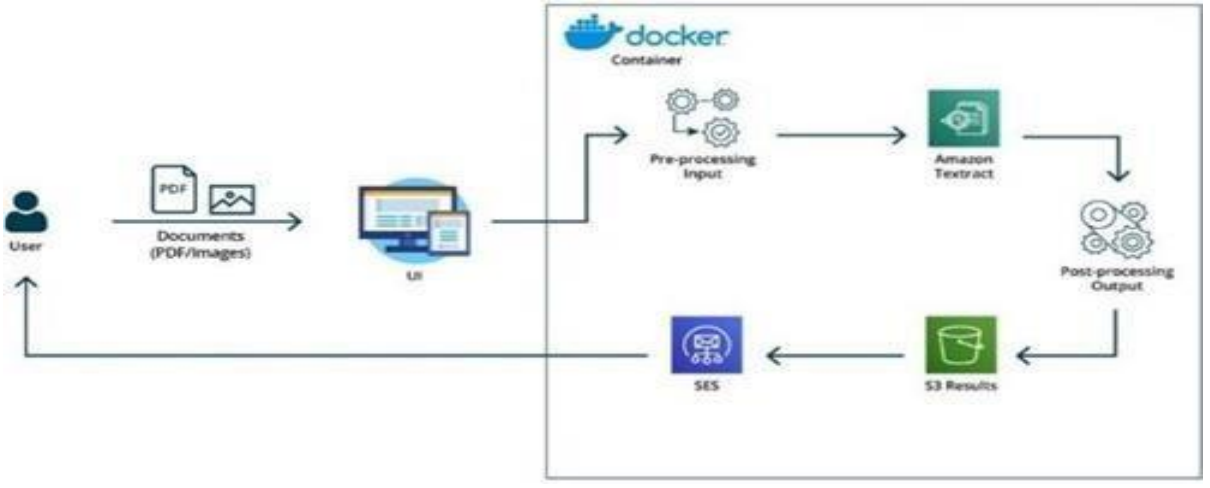


Fig 4.2.1 : System Architecture

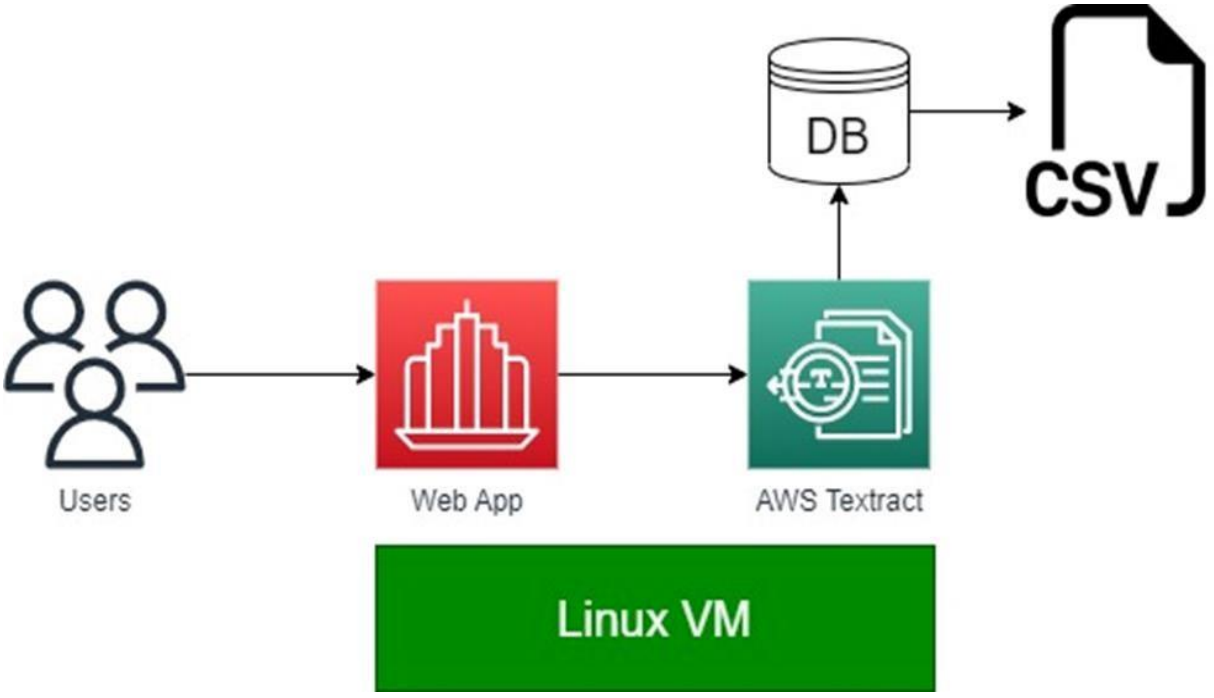


Fig:4.2.2 Architecture of Data from Document

4.3 DESCRIPTION OF SOFTWARE FOR IMPLEMENTATION AND TESTING PLAN OF THE PROPOSED MODEL/SYSTEM

A Laravel PHP based web application hosted on a Linux VM integrates with AWS Textract via APIs to perform intelligent character recognition at scale.

We will test the application and the OCR accuracy with some sample invoices in a variety of formats

4.4 PROJECT MANAGEMENT PLAN

We will complete this project in the Agile mode with 4 bi-weekly sprints. The first sprint will consist of setting up the project and the required infrastructure and cloud resources.

The development of the web application will be taken up as part of the second sprint, while the third sprint involves API integration and OCR functionality development (via AWS Textract integration).

The final sprint will be used to develop the CSV export functionality and the overall end-to-end testing of the application for stability and performance.

4.5 FINANCIAL REPORT ON ESTIMATED COSTING

The Linux VM will cost about INR 4000 per month, while AWS Textract will be charged at INR 1 per document. Meanwhile, PHP and MySQL are open source.

Overall, we estimate the project to take about INR 10,000 to develop, test and release.

4.6 TRANSITION/ SOFTWARE TO OPERATIONS PLAN

Since we will be hosting the application on the cloud, we will delete the test data to get it ready for go-live as soon as the final testing is concluded. We will set up limits in place to ensure fair usage and also avoid bill shocks.

REFERENCES

[1].Jamshed, Memon Maira, Sami, Rizwan Ahmed Khan and Mueen Uddin, “Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)”, School of Computing, Quest International University Perak, Ipoh 30250, Malaysia, Department of Computer Science, Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology, Karachi 75600, Pakistan. Faculty of IT, Barrett Hodgson University, Karachi 74900, Pakistan. Department of Software Engineering, Faculty of Science and Technology, Ilma University, Karachi 75190, Pakistan.

[2].Rishabh Mittal, “Text extraction using OCR: A systematic review”, Department of Computer Science and Engineering, Amity school for Engineering and Technology, Amity university Uttar Pradesh, Noida (UP), India.

[3].Thomas Hegghammer, “OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment.

[4].Alghamdi, Mansoor A., Alkhazi, Ibrahim S., & Teahan, William J. (2016). “Arabic OCR Evaluation Tool.” In 2016 7th International Conference on Computer Science and Information Technology (CSIT), 1–6. IEEE.

[5].Bieniecki, W., Grabowski, S., & Rozenberg, W. (2007). “Image Preprocessing for Improving OCR Accuracy.” In 2007 International Conference on Perspective Technologies and Methods in MEMS Design, 75–80. IEEE.

[6].Boiangiu, C.-A., Ioanitescu, Radu, & Dragomir, Razvan-Costin. (2016). Voting-Based OCR System. The Proceedings of Journal ISOM, 10, 470–86.

[7].Carrasco, R. C. (2014). “An Open-Source OCR Evaluation Tool.” In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 179–84.

[8].Colavizza, G. (2021). Is your OCR good enough? Probably so. Results from an assessment of the impact of OCR quality on downstream tasks. KB Lab Blog.

[9].Dengel, A., Hoch, R., Hönes, F., Jäger, T., Malburg, M., Weigel, A. (1997) “Techniques for Improving OCR Results.” In Handbook of Character Recognition and Document Image Analysis, 227–58.World Scientifc.

[10].Doush, I. Abu, A., Faisal, & Gharibeh, A. H. (2018). “Yarmouk Arabic OCR Dataset.” In 2018 8th International Conference on Computer Science and Information Technology (CSIT), 150–54. IEEE.

[11].Grant, P., Sebastian, R., Allassonnière-Tang, M., & Cosemans, S. (2021). Topic modelling on archive documents from the 1970s: global policies on refugees. Digital Scholarship in the Humanities, March. https://doi.org/10.1093/llc/fqab018

[12].Gupta, A., Gutierrez-Osuna, R., Christy, M., Capitanu, B., Auvil, L., Grumbach, L., Furuta, R., & Mandell, L. (2015). “Automatic Assessment of OCR Quality in Historical Documents.” In Proceedings of the AAAI Conference on Artifcial Intelligence. Vol. 29. 1.

[13].Hamdi, A., Jean-Caurant, A., Sidere, N., Coustaty, M., & Doucet, A. (2019). “An Analysis of the Performance of Named Entity Recognition over OCRed Documents.” In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 333–34. IEEE. https://ieeexplore.ieee.org/document/8791217.

[14].Hegghammer, T. (2021). Noisy OCR Dataset. Repository details TBC.

[15].Holley, R. (2009). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine, 15(3/4)