IJARASEM

ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 3, May 2023

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 6.551

Multi-Doc Parser for scholarship System using Google Cloud Platform

Ms.Sowmya, Rakshith Acharya, Raksha B Kottari, Shreya S

Asst. Professor, Department of Computer Science & Engineering, Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, Karnataka, India

Student, Department of Computer Science & Engineering, Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, Karnataka, India

Student, Department of Computer Science & Engineering, Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, Karnataka, India

Student, Department of Computer Science & Engineering, Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, Karnataka, India

ABSTRACT: The scholarship web application is a user-friendly platform that provides students with the opportunity to apply for scholarships easily and efficiently. The platform allows students to submit their applications online and upload the necessary documents required for their application. The application process is straightforward, with step-by-step guidance provided throughout the process. The platform also features an admin panel that allows scholarship administrators to manage the application process efficiently. The admin panel gets a structured list of all the applicants, and the administrator can access each application and view the necessary information from the documents submitted by the applicants. That is when the admin opens an application, they will see all the relevant information they need to decide on the scholarship application without having to look through the documents themselves. The structured format makes it easier and more efficient for the admin to review the applications and manage the scholarship application process.

KEYWORDS: Google Cloud Platform, Optical Character Recognition, Document AI, Machine Learning, Postman, Virtual Machine.

I.INTRODUCTION

Scholarships are financial aid awarded to students to support their education. They are based on criteria such as academic merit, diversity and inclusion, athletic ability, and financial need, and reflect the values and goals of the donor. The SAR tool helps economically disadvantaged students apply for scholarships. It provides efficient communication, tracks feedback, and offers solutions through excellent administration [1].

Scholarship administration involves not only awarding scholarships to eligible students but also verifying their eligibility criteria and required documents. The staff responsible for administering scholarships often have to go through a lengthy and time-consuming process to ensure that all the information provided by the applicants is accurate and meets the scholarship criteria. They may have to verify academic records, financial needs, and other criteria before awarding scholarships. This process can be challenging, especially when dealing with a large number of applications. Even online scholarship management systems also focus on the other aspect of these issues [6]. However, it is essential to ensure that scholarships are awarded fairly and to deserving students who meet the criteria set by the donor.

The proposed scholarship website aims to simplify the scholarship application process by utilizing cloud storage and document processing technology. Applicants can enter their information and upload their documents in PDF or image format, which will be stored securely in Google Cloud Storage as well as in the database.

The Google Cloud Platform also offers a Document AI service that utilizes Optical Character Recognition to extract required data from uploaded documents. The processed data can then be used for various purposes, including verifying eligibility criteria, reducing manual effort, and saving time for scholarship administrators. This technology can improve the efficiency and accuracy of the scholarship administration process, ultimately benefiting both applicants and administrators.

II.LITERATURE REVIEW

The proposed student scholarship system is based on an online platform that utilizes two algorithms, collaborative filtering, and content-based filtering, to categorize applicants into three categories: caste-wise, academic-wise, and poor status. The Admin is responsible for adding student scholarships into the appropriate category, while students can register and apply for scholarships online. The system includes a filtering process based on caste, and students can check their application status to see if their scholarship has been paid or unpaid. The system aims to simplify the scholarship application process and make it more accessible to students while improving the efficiency and accuracy of scholarship administration. The proposed system also emphasizes the importance of scholarship students' involvement and the role of their Alma Mater in their academic and professional development[1].

Although the accuracy of OCR software varies, optical character recognition can make understudied historical materials available for computational research. The performance of Tesseract, Amazon Textract, and Google Document AI on pictures of English and Arabic text was benchmarked in this paper. For a corpus of 18,568 documents. The best outcomes were produced by Document AI, and the server-based processors (Textract and Document AI) outperformed Tesseract significantly, especially for noisy documents. Arabic has significantly lower accuracy than English. Anyone with knowledge can find better OCR solutions for their research needs by describing the relative performance of these three prominent OCR systems and the differential effects of regularly encountered noise types[2].

This paper briefly reviews some of the representative models, tasks, and benchmark datasets. The acceleration of digitization has become a key part of the success of digital transformation. Document analysis techniques based on heuristic rules, algorithms, and models based on traditional statistical machine learning. However, traditional rule-based methods often require large labor costs. There are Document AI products, from Microsoft, Amazon, and Google Documents. For instance, forms are usually displayed in the form of key-value pairs. Specifically, according to the text bounding boxes obtained by OCR, the algorithm first gets the coordinates of the text in the document. After converting the corresponding coordinates into virtual coordinates, the model calculates the representation of the coordinates corresponding to the four embedding sublayers of x, y, w, and h[3].

The massive production of documents in portable document format (PDF) format has motivated research on the automated extraction of data contained in these files. This work is mainly focused on extractions of natively digital PDF documents, made available in large repositories of educational exams. For this, the educational tests applied at Evade were used. The results of the extractions point out some limitations concerning the diversity of layout in each year of application. The extracted data provide useful information in a wide variety of fields, including academic examination and support for students and teachers. We are developing a model in such a way that it will be able to extract data from images or scanned papers etc. We will be achieving this with the help of optical character recognition technology[4].

Key information extraction (KIE) from document images requires understanding the contextual and spatial semantics of texts in two-dimensional (2D) space. This paper tackles the problem by going back to the basics: the effective combination of text and layout. They have 2 categories: entity extraction (EE) and entity linking (EL). The EE task identifies sequences of text blocks that represent desired target texts (header, question, and answer). The EL task connects key entities hierarchically (such as their name, unit price, amount, and price.) They have BROS, which focuses on modeling text and layout features for effective key information extraction from documents[5].

III.METHODOLOGY

The scholarship system is a website that streamlines the scholarship application process. It leverages Google Cloud Platform's Document AI service, which is based on optical character recognition technology, to parse and extract required information from submitted documents. This structured data is then presented to the administration for further processing, reducing the need for manual intervention and minimizing the possibility of errors.

By allowing applicants to electronically submit their documents and automatically extracting relevant data, the Scholarship System offers several benefits. It simplifies the application process, saves time and resources, and helps ensure that the information submitted is accurate and complete. This innovative approach can transform how scholarships are administered and make the process.

Google Cloud Platform (GCP) is a suite of cloud computing services that allows developers to build, test, and deploy applications on Google's infrastructure. One of the services provided by GCP is Document AI, an optical character recognition (OCR) based service that helps extract information from unstructured data, such as scanned documents or PDFs.

OCR is the process of converting text within scanned documents into a machine readable format. Modern OCR tools are fairly advanced and use steps such as document pre-processing, feature extraction followed by character/word/document classification and post processing. Document AI uses machine learning algorithms to recognize text and extract key information such as dates, names, and addresses.

The Scholarship System utilizes APIs for various functions, including login and signup processes, as well as application filling. APIs are also used to fetch data from user sections and display them to the administration. Additionally, the Scholarship System leverages APIs to save the information entered by applicants in a MongoDB database.

APIs act as an essential component of our Scholarship System's functionality, enabling seamless communication and integration between various components. By leveraging APIs, the Scholarship System can provide a user-friendly experience for applicants and administrators alike, it also ensures that all relevant information is stored securely and accurately in the database.

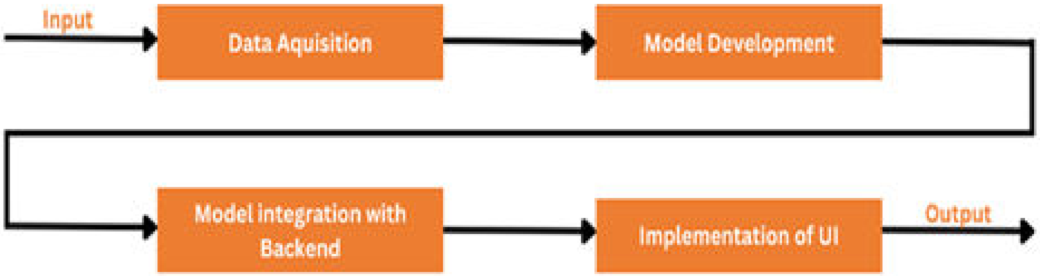


Figure 1: Detailed Methodology

1. Data Acquisition:

Data acquisition involves collecting important documents from applicants such as Aadhaar card, income certificate, and marks cards. This data is used to train the Google Cloud Document AI model, which automatically extracts relevant information such as applicant name, address, education details, and income information. This ensures a streamlined and efficient application process, with minimal errors and accurate data extraction.

It's also important to speak with representatives from various scholarship organizations to understand their unique application procedures and selection criteria, which can aid in a variety of ways with the success-oriented process moving forward.

2. Model Development:

Document AI who's API Should be enabled where they provide a lot of pre-built templates for a model. But here customized models are built for all the documents which need to be parsed and extract text from the documents. The customized model contains the buckets where all the training and testing documents are to be stored which should be imported to the Document AI.

Next is the training process where necessary data is identified from the document and trained it to the model. Each document's data position and the data should be marked and should be labeled. All the documents in testing and training should be marked. A key- value pair where the labeling the data is a key and the value is nothing but the marked data. So here completes the training process.

Next the model should be deployed. Once it is done the testing of the model can be done. The testing can be done manually too. But to train and test the model minimum ten documents are required.

3. Model Integration with Backend:

Integrating a model provided by Google Cloud Services with the backend typically involves using the API provided by the model to communicate with the backend. The database which is used is MongoDB, which is used to store data.

Node JS is used as a backend programming language which is an open source cross platform JavaScript runtime environment. With this integration, data is sent from the backend to the model, which processes it and sends the processed data back to the backend. The documents are parsed once and then directly stored in the database.

4. Implementation of UI:

Design files must be used to create a usable user interface before the UI can be implemented. Usually, this entails creating HTML, CSS, and JavaScript code to develop the UI's structure, design, and functionality.

React JS, an open source JavaScript package, is used to design user interfaces, which speeds up development and improves functionality. It is important to ensure that the UI is responsive, accessible, and optimized for various devices and browsers. Front-end development must include testing to make sure the user interface is intuitive and performs as planned.

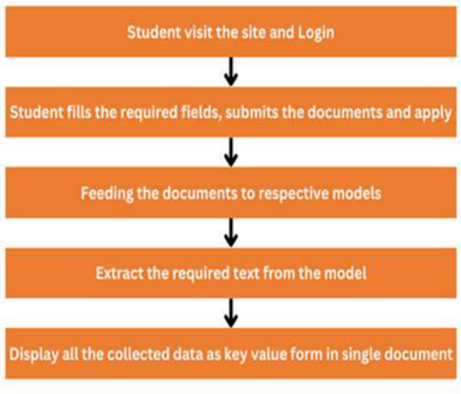

Student visit the site and Login
Student fills the required fields, submits the documents and apply
Feeding the documents to respective models
Extract the required text from the model
Display all the collected data as key value form in single document

Figure 2: Overview of the system

The system process consists of two modules, user and admin where both play an important role to build the architecture of this system.

In the user module, the user/applicant has to undergo all of the criteria and applying process. After browsing the website, the user has to register themselves at first if they are the new user to use this website, if he is already a user then he has to sign in by providing the necessary fields.

A home page with all the necessary information and the criteria will be shown to the user. The user then must click the button for the further applying process. A scholarship application arises which has been divided into different parts as personal data, educational data and other necessary documents. After the completion of the applying process, the provided data by the user will be saved in Google’s cloud storage and in the database.

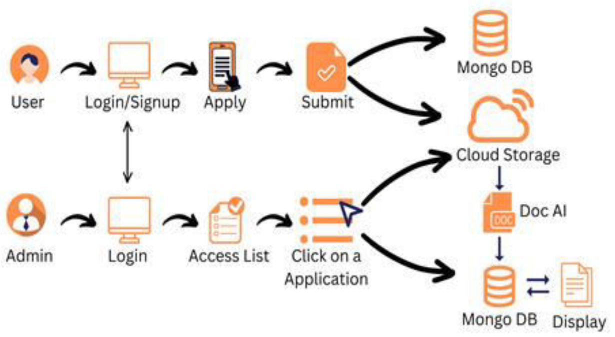


Figure 3: System Process Diagram

Next is the admin module, where the admin has to log-in using the same website. When the admin login with the appropriate details, a list of applicant names who have submitted the application will appear. When the admin clicks on a particular applicant name for the first time, it takes to the information page of that particular applicant.

Here the admin can view all the documents uploaded by that applicant. When the parse document button is clicked it directly evokes the Document AI models five processors to parse and extract the needed data from the documents of that applicant and process the structured data which contains all the other provided data with the extracted data as a single document which is stored in the database. When the admin clicks on the view data button the data will be directly displayed from the database. When the admin wants to get the structured data for the second time he/she should view it by using the view data button. For a particular applicant, the parse document button is only used once.

IV.RESULT

The result section of the scholarship website will provide valuable insights into the GradPeny website of both admin and user modules where the applicants data are collected, with all documents securely stored on the Google Cloud platform. The platform's powerful Document AI service, based on Optical Character Recognition, will extract required data, such as student names, academic records, and personal information. This processed data will be used to evaluate applicant eligibility.

The exactness, accuracy, support, and F1 score are commonly used to assess the performance of a model's predictions. Totally, five processors namely SSLC, PUC, Aadhar Card, Income Certificate and BE are created. Each processor gives the distinct F1 score which is the measure of model accuracy on a dataset.

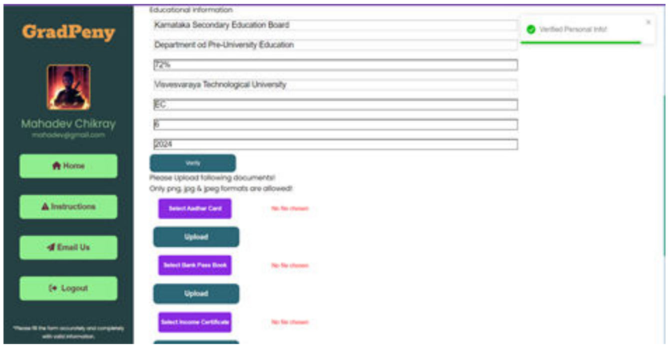
| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

Processors	F1 Score	Percentage
SSLC	0.928	92.8%
PUC	0.921	92.1%
Cast Certificate	1	100%
Aadhar Card	1	100%
BE	0.961	96.1%

The outcome of GradPeny user module are as follows:

GradPeny is a website that offers a scholarship opportunity. When users visit the website, they are welcomed with a home page that provides information about GradPeny's motivation and why it is a great choice for a scholarship. Users can also find a contact option to reach out to the administration for any queries. If users wish to register, they can click on the register button located on the right end of the home page. The registration form requires users to provide their full name, email address, password, and confirm the password.



Educational Information

Karnataka Secondary Education Board

Department od Pre-University Education

72%

Visvesvaraya Technological University

EC

6

2024

Please Upload following documents!

Only png, jpg & jpeg formats are allowed!

Verified Personal Info!

*Please fill the form accurately and completely with valid information.

Figure 4: Application Form Page

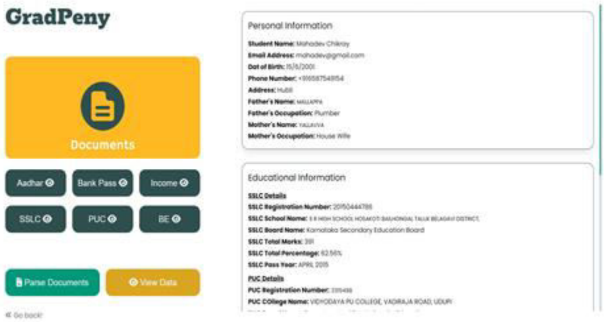
Returning users can simply log in using their email address and password. Once logged in, users can access the application form, which consists of personal and educational sections. Users must verify their inputs for each section before proceeding.

Additionally, there is a document section where users can upload the required documents. The application form page also features a user profile and four buttons for navigating to the home, instruction, email, and log out sections.

Next is the outcome of GradPeny admin module are as follows:

The website serves both users and administrators, with separate login options for each. The admin can access the user login page by clicking on "register" and then switching to the admin login. To log in as an admin, the correct email address and password are required.

Once logged in, the admin sees their profile on the left where the right side contains the list of applicants, along with a list of applicant names it also contains email addresses, application dates, processing status, and the option to delete applicants. Clicking on an applicant's name reveals a page displaying the uploaded documents, which can be viewed individually.



Personal Information

Student Name: Mahadev Chikray
Email Address: mahadev@gmail.com
Dat of Birth: 15/5/2001
Phone Number: +916587548154
Address: Hubli
Father's Name: MALLAPPA
Father's Occupation: Plumber
Mother's Name: YALLAVVA
Mother's Occupation: House Wife

Educational Information

SSLC Details
SSLC Registration Number: 20150444788
SSLC School Name: S R HIGH SCHOOL HOSAKOTI BAILHONGAL TALUK BELAGAVI DISTRICT,
SSLC Board Name: Karnataka Secondary Education Board
SSLC Total Marks: 391
SSLC Total Percentage: 62.56%
SSLC Pass Year: APRIL 2015
PUC Details
PUC Registration Number: 235438
PUC College Name: VIDYODAYA PU COLLEGE, VADIRAJA ROAD, UDUPI

Figure 14: Display of the Applicant Input with Parsed Data

Clicking on the "parse document" button, it extracts and parses information from the documents and stores it in the database. Parsing can only be done once per applicant. However, the admin can view the applicant's details, including both the original input and parsed data, by clicking on the "view data" button, which can be done multiple times.

V. CONCLUSION

In conclusion, scholarships are a valuable form of financial aid that can support students in achieving their educational goals. The criteria for awarding scholarships can vary, including academic merit, diversity and inclusion, athletic ability, and financial need. However, scholarship administration can be a challenging process, requiring extensive verification of applicants' eligibility criteria and required documents. To simplify this process, the proposed scholarship website utilizes cloud storage and document processing technology, allowing applicants to easily upload their documents in PDF or image format. The technology can also extract necessary data from uploaded documents, improving the efficiency and accuracy of scholarship administration. Ultimately, the goal of reducing the manual work in the scholarship administration will be reduced.

REFERENCES

1. T.Sowndhariyaa, T.M.Nithyav, "Advanced Application System for Student Scholarship using Content Based Filtering Technique," International Journal for Modern Trends in Science and Technology 2021, 7, 0708041, pp. 110-115 August 2021.
2. Thomas Hegghammer, OCR with Tesseract, Amazon Textract, and Google Document AI, Journal of Computational Social Science volume 5, oages861-882, 2022.
3. Lei Cui, Yiheng Xu, Tengchao Lv, Furu Wei, "Document AI: Benchmarks, Models and Applications", 16 Nov 2021.
4. Automated Data Extraction from PDF Documents: Application to Large Sets of Educational Tests**,** International Conference on Enterprise Information Systems**,** May 2021.
5. Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, Sungrae Park, "BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents", Apr 2022
6. Governance Knowledge Center: Online Scholarship Management, Researched and Documented: oneworld.net, owsa@oneworld.net.