Survey on Advanced Text Extraction Techniques Using OCR and Resume-Based Job Recommendation Systems for Sponsorship Platforms

Tejas Adagale[1] Abhishek Dhage[2] Onkar Shivarkar[3] Vikramsinh Khalate[4] Prof. Y. R. Khalate[5]
[1,2,3,4]Student [5]Professor
[1,2,3,4,5]Department of Computer Science & Engineering
[1,2,3,4,5]SVPM's College of Engineering, Malegaon (Bk), Maharashtra, India

Abstract — This review examines cutting-edge research and machine learning techniques that support platforms, along with optical character recognition (OCR) technology for effective text extraction. The main objective of the platform is to bridge the gap between students and organizations, offering a shared space for students to receive assistance and for organizations to identify essential roles and skills. Based on this research, the paper illustrates how OCR, iterative parsing, and hybrid consensus models collaborate to enhance the accuracy of matching students with advocates. Contemporary OCR methods transform physical documents into text formats that can be analyzed through machine learning algorithms. Furthermore, this paper investigates a hybrid strategy that merges collaborative and content-based filtering to strengthen the connection between student activities and required sponsors. This approach holds significant potential for developing platforms that not only harness the project's capabilities but also align learners' skills with the specific demands of institutional support.

Keywords: OCR, Text Capture, Resume Analysis, Sponsorship Platform, Job Offer, Cosine Similarity, Machine Learning, Deep Learning, Natural Language Processing (NLP)

I. Introduction

Automatic elements such as text mining, text mining conflict resolution and various recommendation mechanisms have been effectively adopted in the area of providing services to students. Before an optical character recognition (OCR) system appeared it was not practical as the primary and efficient means of supporting effectively of the film represents a document which OCR technology shifts the system with multiple request processing. Each frame emphasis from image to documents of different types and serves to build a support system. Finally, because of the recommendation techniques, it is easier to create narratives, skill, and knowledge support needed by the students and the institution. In this regard this section explains how the support platform operates and presents a few cases on the example of OCR technology being used for waste activities and what proved to be successful hybrid recognition systems. The purpose of the platform is to study the proposal of the student, identify the goals and skills of the project, and provide the right sponsors in relation to some parameters like the nature of the support model in the past and its relationship to the sponsors objective. With this introduction, the paper outlines all possible components of the platform and how they contribute to enabling competition.

II. Literature Survey

A. Text Extraction using OCR

OCR technology has become essential for managing digital documents, enabling the conversion of scanned images into searchable formats, especially through open-source tools like Tesseract. Over time, OCR systems have advanced, with deep learning-based engines now offering enhanced accuracy, even for complex documents. The text discusses how OCR interacts with various documents, accommodating different fonts, layouts, and image qualities. This section illustrates the significance of OCR technology for sponsorship platforms, as it helps sponsors extract vital information from large files while filtering necessary precision and pattern recognition capabilities to extract clear, structured text from proposals

B. Resume and Project Proposal Analysis for Sponsorship

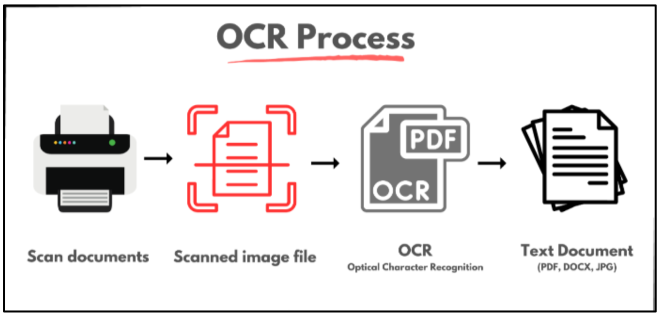
Resume and project proposal analysis are core functions of the sponsorship platform and provide a basis for comparing students with suitable sponsors. In this section, the survey examines techniques such as Naive Bayes and Cosine Similarity, which help identify students' resumes and jobs by comparing their skills, experiences, and descriptions of potential support needs. Techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) are used to vectorize the text and convert it to a digital format for further analysis. A similar cosine test is used to measure how well the student's skills and knowledge match the sponsor's criteria. Focusing on these text analytics models, the study shows how machine learning can classify student data to provide sponsors with a list of person matches that meet specific requirements.

C. Hybrid Recommendation Systems for Sponsorship Matching

Recommendations are crucial in shaping promotional campaigns by offering personalized suggestions through a blend of collaborative and content-based filtering. This section explores how hybrid consensus models merge insights from previous sponsors (collaborative filtering) with student interests (content-based filtering) to generate a tailored match list. The integration leverages historical data from past sponsorships to suggest sponsors for new student programs that align with similar previous opportunities. For instance, content-based filtering assesses the distinct learning traits of each student and aligns them with the sponsor's objectives and interests. By combining these approaches, hybrid recommendations can deliver personalized, precise, and meaningful insights to both students and sponsors, ensuring that support planning is highly relevant and well-aligned.

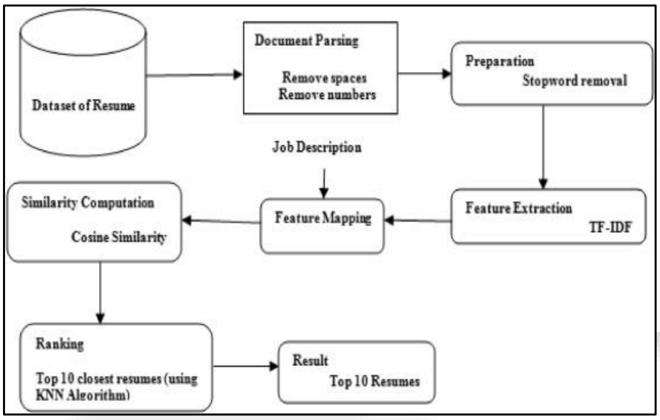
III. SYSTEM ARCHITECTURE

A. OCR System for Proposal Extraction



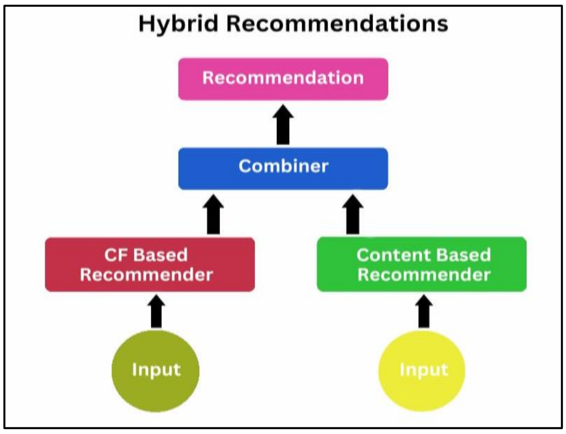
The platform employs OCR techniques to pull out essential information from student proposals. This process includes transforming scanned documents into text, which is then. analysed further. Specific extraction methods like zoning and segmentation help pinpoint key components such as project objectives, timelines, and funding requirements out student comments. Cutting-edge OCR methods, such as convolutional neural networks (CNNs), deliver the

B. Resume Analysis and Job Recommendation



In analysing CVs and project proposals, we can use methods such as feature extraction, TF-IDF vectorization, and cosine similarity for matching. This approach allows us to align a student's project with the specific requirements of the sponsor, taking into account the skills, experience, and objectives outlined in the application.

C. Hybrid Recommendation Matching



The Hybrid Recommendation System aims to create personalized matches between students and sponsors. It integrates collaborative filtering, which relies on historical support data, with content-based filtering to effectively evaluate the implementation plan's content and the organization's objectives.

IV. DATASET

A. OCR Dataset for Proposal Analysis

The OCR system is designed to process various documents by scanning proposals and study materials. This diversity enhances the accurate digitization of documents, accommodating different fonts, layouts, and image qualities. OCR machines trained on this type of data are typically adept at extracting information from actual documents submitted to support platforms.

B. Resume and Project Proposal Dataset

This section discusses the review of resumes using a comprehensive database of student resumes and descriptions gathered from various educational institutions. It outlines the skills, experiences, and key factors related to career advancement. When integrated with the support system, this data enables the training.

C. Sponsorship Matching Dataset

The recommendation model is developed using data that includes relevant background information such as previous work experience, grants received, and personal interests. This section highlights the effective support model by presenting a consensus framework that details the traits of an ideal supporter for particular students. This historical information significantly improves the platform's ability to predict and identify highly advantageous matches.

V. RESULT ANALYSIS

Paper	Task	Accuracy	Key Findings
OCR-based Proposal Extraction	Text extraction from project proposals	Improved Accuracy	High accuracy in extracting structured information from various document formats, utilizing deep learning techniques to manage differences in fonts, layouts, and image quality.
Resume Analysis and Job Recommendation	Sponsorship matching	Precision: 0.99	Highly accurate matching through cosine similarity and machine learning models, connecting student resumes with sponsor requirements based on skills, experience, and project needs.
Hybrid Recommendation System	Sponsorship recommendation	F1 Score: 0.96	Improved matching accuracy by combining collaborative and content-based filtering, resulting in personalized recommendations that align with both student profiles and sponsor objectives. Proposals and sponsor needs.

VI. FUTURE SCOPE

The future scope of this research is extensive, focusing on enhancing the efficiency and accuracy of matching sponsors with students through improvements in Optical Character Recognition (OCR) and recommendation algorithms. As OCR technology evolves, the integration of more advanced deep learning techniques could enable the platform to extract detailed information from increasingly complex document formats with greater precision. This would facilitate quicker and more accurate parsing of student proposals and resumes, regardless of variations in font or layout. By capitalizing on this progress, the platform can manage a higher volume of applications while ensuring accuracy, thereby improving its scalability and making it accessible to a wider audience.

Another exciting avenue involves refining the recommendation system to boost real-time compatibility and relevance. By incorporating sophisticated machine learning models, such as hybrid recommendation systems that blend collaborative and content-based filtering, the platform could provide near-instantaneous recommendations. This system would leverage historical sponsorship data along with current user preferences and skills, creating a dynamic matchmaking process that adapts with each new interaction. Furthermore, enhancing this real-time adaptability could be especially advantageous in high-demand fields, were swift model to effectively differentiate between qualified candidates and those who may be less suitable for each type of support.
alignment between project needs and sponsorship interests is essential.

Finally, the platform could broaden its functionality to encompass a wider range of support areas, thereby creating sponsorship opportunities across various industries and academic disciplines. By expanding its reach, the platform can establish a comprehensive support network that connects students with sponsors for internships, research grants, and mentorship opportunities beyond conventional project funding. Additionally, future iterations of the platform may incorporate tools for ongoing project monitoring and performance feedback, enabling sponsors to track student progress and providing students with insights to better meet sponsor expectations.

VII. CONCLUSION

This survey outlines the importance of OCR systems, background checks, and hybrid recognition strategies in creating effective support. While OCR technology played a key role in the project, the background analysis and recommendations also shaped the relationship between students and sponsors. By combining collaborative filtering with content-based methods, both parties benefit as the platform enhances its ability to identify the most engaged students. Future advancements in this field could aim at enhancing real-time compatibility and broadening the platform's reach to additional support areas.

REFERENCES

[1] R. Mittal, A. Garg, "Text Extraction using OCR: A Systematic Review," IEEE, 2020.
[2] A. Mankawade, V. Pungliya, "Resume Analysis and Job Recommendation," IEEE, 2023.
[3] H. Zhang, J. Li, and X. Wei, "Text Extraction from Document Images using CNN and LSTM," IEEE XPLORE, 2023.
[4] A. Patel, R. R. Gupta, and K. Deshmukh, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review," IEEE XPLORE, 2020.
[5] M. Kumar, S. P. Singh, and S. Verma, "A Detailed Review on Text Extraction Using Optical Character Recognition," SCIENCEGATE, 2023.
[6] Y. Liu, F. Zhang, and X. Chen, "A Review on Recommendation Systems: Context-aware to Social-based," AR5IV, 2022.
[7] Rinaldi, R. Ferdiana, "A Review of Text-based Recommendation Systems in Text Mining," IEEE, 2022.
[8] R. S. Patel, M. Joshi, and N. K. Jain, "Hybrid Recommendation Systems for E-commerce and Beyond," IEEE XPLORE, 2023.
[9] J. M. Wilson, L. Zhang, and T. K. Patil, "A Comprehensive Guide to Information Extraction from Texts with Spark NLP," GITHUB, 2023.
[10] C. R. Li, Y. Wang, and M. Hu, "Information Extraction from Full Text Scientific Articles: Challenges and Techniques," IEEE XPLORE

