AI-Powered Sponsorship Recommendation System for Students using TF-IDF, SpaCy NLP, and Rule-Based Skill Mapping

Tejas Adagale[1] Abhishek Dhage[2] Vikramsinh Khalate[3] Onkar Shivarkar[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 paper presents an intelligent and interpretable recommendation system designed to effectively match students with potential sponsors, fostering a mutually beneficial relationship between emerging talent and organizations seeking innovative contributors. The platform utilizes a hybrid approach combining natural language processing (NLP) techniques and vector space models to analyze and compare student resumes with job or sponsorship descriptions. Core components of the system include text preprocessing using SpaCy—where stopwords, punctuation, and irrelevant content are removed and words are lemmatized to their base form—followed by the transformation of text into TF-IDF vectors to capture the relative importance of terms. Cosine similarity is then used to compute the degree of alignment between resumes and sponsor-defined criteria. Additionally, the system features a rule-based skill extraction module that identifies technical proficiencies from resumes using a curated skill list, and a recommendation engine that matches these extracted skills against predefined job roles through intersection logic. This combined methodology ensures accurate and explainable results, helping organizations identify suitable candidates while empowering students to access meaningful opportunities based on merit. The platform serves as a step forward in democratizing access to sponsorships and enhancing visibility for skilled students across diverse domains.

Keywords: Resume Matching, TF-IDF, Cosine Similarity, Spacy NLP, Skill Extraction, Rule-Based Recommendation, Sponsorship Platform, Student Recommendation System, Text Preprocessing, Natural Language Processing, Talent Discovery, AI-Based Matching, Resume Parsing, Job Role Recommendation, Skill Mapping, Intelligent Ranking System, Educational Sponsorship, Career Guidance Platform, Candidate Scoring System, Semantic Similarity

I. Introduction

In the rapidly evolving landscape of education and industry collaboration, there exists a significant disconnect between students seeking opportunities and organizations looking for fresh, skilled talent to invest in. While students often possess exceptional skills and innovative project ideas, their access to sponsorships or funding opportunities is limited due to lack of visibility and structured discovery mechanisms. On the other side, companies, research institutes, and NGOs are continuously searching for competent individuals who can contribute meaningfully to their projects, internships, and research initiatives. However, manual evaluation of resumes or project proposals is time-consuming, inefficient, and prone to human bias.

To bridge this gap, this project presents a novel, AI-powered sponsorship recommendation system specifically designed to empower students and assist sponsors in discovering relevant talent. The system automates the resume analysis and matching process by integrating multiple technologies and techniques from natural language processing (NLP), information retrieval, and rule-based logic.

At the core of the system lies the use of TF-IDF (Term Frequency–Inverse Document Frequency) and cosine similarity, which transform resumes and job descriptions into mathematical vectors to compute how closely a student's profile matches a sponsor's requirement. Prior to this, all textual content is cleaned, lemmatized, and standardized using SpaCy, an advanced NLP library. This ensures that the language is normalized and free of unnecessary noise like stopwords and punctuation.

The system further strengthens its capability through rule-based skill extraction, wherein resumes are parsed to detect relevant technical and non-technical skills by comparing them with a predefined skill dictionary. Once skills are identified, the recommendation engine matches them against sponsor-defined job roles or funding criteria using set-intersection logic to determine suitability.

By combining these methodologies, the platform not only recommends the most appropriate sponsorships to students but also ranks and highlights the best-fit candidates to sponsors. This intelligent and interpretable framework helps democratize access to funding and mentorship, especially for students from underrepresented or remote backgrounds, while simultaneously offering organizations a streamlined way to identify and invest in potential future leaders.

II. Literature Survey

Several research efforts have explored the application of machine learning and natural language processing in recruitment and recommendation systems. Traditional approaches for job matching primarily relied on keyword-based search or manual filtering, which often failed to capture semantic relevance and contextual meaning. Recent advancements introduced vector-based models such as TF-IDF and cosine similarity to improve textual comparisons between resumes and job descriptions. Works like those by Kulkarni et al. (2023) and Yan et al. (2020) have demonstrated the effectiveness of NLP in extracting relevant information from unstructured text for decision-making systems. Additionally, rule-based methods for skill extraction have proven efficient and interpretable for structured environments. However, most existing systems cater to corporate hiring and lack customization for academic sponsorship or student-oriented platforms. This project builds upon such foundations and addresses this gap by focusing on intelligent, transparent, and skill-based recommendation

systems tailored specifically for student sponsorship discovery.

III. Methodology

The proposed sponsorship recommendation system follows a structured, modular pipeline to process input data (resumes and job descriptions), extract relevant features, compute similarity, and recommend suitable sponsorships. The methodology comprises the following key steps:

A. Step 1: Text Preprocessing

- Tool Used: SpaCy (en_core_web_sm model)
- Purpose: Clean and normalize textual data from resumes and job descriptions to prepare it for vectorization.
- Operations Performed:
  - Lowercasing all text
  - Removing punctuation
  - Removing stopwords (e.g., "and", "is", "in")
  - Lemmatization (e.g., "running" → "run")

This results in standardized, tokenized text that improves the quality of vector representations in the next stage.

B. Step 2: Feature Extraction using TF-IDF

- Tool Used: TfidfVectorizer from scikit-learn
- Purpose: Convert cleaned text into numerical vectors representing term importance.
- Formulas:

1) *Term Frequency (TF):*

$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

2) *Inverse Document Frequency (IDF):*

$$IDF(i) = \log_2\left(\frac{\text{Total documents}}{\text{documents with term i}}\right)$$

3) *TF-IDF Score:*

$$\text{TF-IDF} = \frac{\text{term frequency}}{\text{total words}} \times \log\left(\frac{\text{corpus size}}{1 + \text{documents with term}}\right)$$

Each document (resume or job description) is thus converted into a high-dimensional TF-IDF vector.

C. Step 3: Similarity Computation using Cosine Similarity

- Tool Used: cosine_similarity() from sklearn.metrics.pairwise
- Purpose: Quantify the similarity between a student's resume and a sponsor's job description.
- Formula:

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2}$$

$$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2}$$

The result is a similarity score between 0 and 1, where higher scores indicate greater textual relevance.

D. Step 4: Skill Extraction (Rule-Based)

- Tool Used: Custom logic with predefined skill list (skill_mapping.txt)
- Purpose: Identify technical skills mentioned in the resume.
- Process:
  - Compare each token in the preprocessed resume text against the predefined skill list.
  - If a match is found, the skill is extracted and added to the user's skill set.

Example:
Resume: "Experienced in Python, Machine Learning, and Pandas"
Extracted Skills: {Python, Machine Learning, Pandas}

E. Step 5: Role Recommendation (Rule-Based)

- Tool Used: Set-based matching logic
- Purpose: Recommend suitable sponsorship roles based on overlapping skills.
- Formula:

$$\text{Match Score} = \frac{|\text{Extracted Skills} \cap \text{Required Skills}|}{|\text{Required Skills}|}$$

- Decision Logic:
  - If the match score ≥ threshold (e.g., 0.5), the job is considered a good fit and recommended.
  - Multiple roles can be recommended based on match scores.

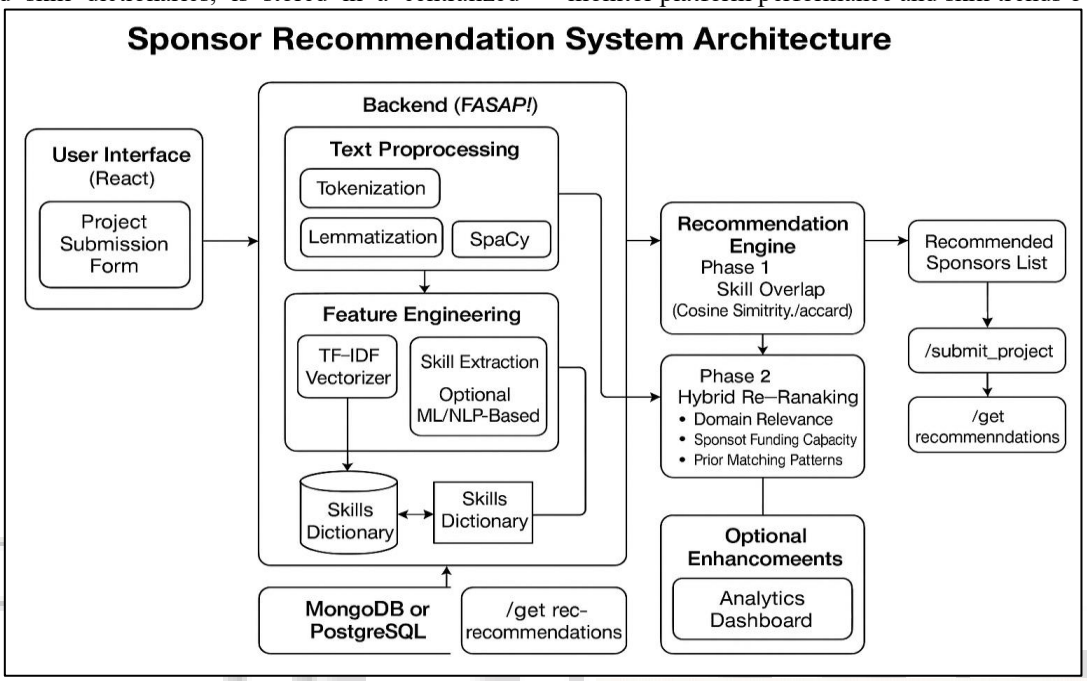
F. Step 6: Output Ranking

- Resumes are ranked per sponsor by cosine similarity.
- Sponsorships are recommended to students based on skill overlap and ranking.
- A top-N list of recommendations is generated and displayed.

IV. System Architecture

The system architecture for the sponsor recommendation platform consists of several interconnected components designed to process project descriptions and recommend suitable sponsors effectively. At the frontend, a React-based user interface allows students to submit their project details, including title, description, and technologies used. Once submitted, the data is sent to the backend, built with FastAPI, where the text undergoes preprocessing using SpaCy for tokenization, lemmatization, stopword removal, and named entity recognition. The cleaned text is then transformed into numerical features using TF-IDF vectorization. A rule-based skill extraction module identifies relevant skills by matching the text against a predefined dictionary of technical terms. These extracted skills are fed into a two-stage recommendation engine. The first stage calculates skill

overlap using similarity measures like cosine similarity or Jaccard similarity to find initial sponsor matches. In the second stage, a re-ranking mechanism refines the results by considering additional filters such as project domain relevance, sponsor funding capacity, and previous matching patterns. The final output is a ranked list of recommended sponsors. All data, including user submissions, sponsor profiles, and skill dictionaries, is stored in a centralized database (MongoDB or PostgreSQL). The system also exposes RESTful API endpoints for actions like project submission and recommendation retrieval. Optionally, the architecture can incorporate an embedding store (e.g., FAISS) for semantic search and a feedback loop for users to rate recommendations, enabling iterative model improvement. An analytics dashboard can also be added to monitor platform performance and skill trends over time.



V. INPUT DATA

The proposed system processes two primary types of input provided by end users:

A. Student Data:

Students submit resumes in either .pdf or .txt format. These resumes contain unstructured information about the student's academic background, technical and soft skills, project experience, certifications, and internships. This data serves as the basis for skill extraction and profile vectorization.

B. Sponsor Data:

Sponsors or organizations provide job descriptions or sponsorship opportunity listings, which detail the required skill sets, responsibilities, qualifications, and domain of work. These descriptions act as reference documents for computing textual similarity with student resumes.

Both inputs are pre-processed using NLP techniques to ensure consistency and cleanliness before entering the matching pipeline. Certainly. Here's a more concise version of the Loss Function section suitable for a research paper:

VI. LOSS FUNCTION

Although the current system uses rule-based and similarity matching techniques, it is designed to support future integration with machine learning models. In such cases, a cosine distance-based loss function can be used to train the model by minimizing the dissimilarity between student resumes and sponsor descriptions.

$$\text{Loss} = 1 - \frac{A \cdot B}{|A| \cdot |B|}$$

Here, AA and BB are TF-IDF vectors of the resume and job description, respectively. This loss function ensures that similar pairs have higher similarity scores and dissimilar pairs are penalized, improving overall matching accuracy during training.

VII. FUTURE SCOPE

The sponsor recommendation platform holds significant potential for future development and real-world impact. One of the primary areas for advancement lies in enhancing the recommendation engine using deep learning techniques. By integrating transformer-based models like BERT or T5, the system can better understand contextual relationships in project descriptions and sponsor profiles, thereby improving the precision of skill extraction and matching. Moreover, implementing a semantic search feature using embedding stores such as FAISS or Pinecone can enable faster and more accurate retrieval of sponsors based on project similarity. Real-time recommendation refinement through user feedback and adaptive learning mechanisms will make the system more dynamic and intelligent over time.

Beyond technical improvements, the platform can be extended into a full-fledged ecosystem connecting students, educational institutions, and industry sponsors. Features like a sponsor dashboard, funding history analytics, success stories, and project tracking can increase engagement and transparency. Integration with LinkedIn APIs or institutional databases could automatically pull project achievements and student credentials, streamlining the onboarding process. Additionally, the system could be expanded to include international sponsors and enable cross-border collaborations, thereby creating a global innovation bridge. Such scalability and adaptability ensure that the platform can evolve into a cornerstone for educational funding and talent recognition in the academic and professional world.

VIII. Conclusion

This project presents an AI-driven sponsorship recommendation system that effectively bridges the gap between talented students and potential sponsors. By leveraging NLP techniques like TF-IDF and SpaCy, along with rule-based skill extraction and similarity scoring, the system provides accurate, interpretable, and skill-matched recommendations. It enhances student visibility and streamlines talent discovery for organizations, creating a mutually beneficial ecosystem for academic and professional growth.

References

[1] Durgesh Kumar Kushwaha; Lipakshi Arora; Abhinav Anand; Riya; Darshit Sheth; Shanu Khare "AI Powered Job Recommendation System" IEEE 2024.
[2] Thansidh Kirthan Nadar; S.S. Athulya; S.S. Rahul; K. Meenakshi "Job Recommendation by Content Filtering using TF-IDF and Cosine Similarity" IEEE 2024.
[3] Deepanshi Gupta; Preetvanti Singh "Development of a Recommendation System for Resume Tracking and Job Suggestions" IEEE 2024.
[4] Derick Prince B; Madhan K; Vishwa K; Yamunathangam D "Job and Course Recommendation System using Collaborative Filtering and Naive Bayes algorithms" IEEE 2023.
[5] Wentao Yan; Hongsheng Yu; Weixing Lu; Jianwen Yin "An Improved Text Similarity Calculation Method Combining TF-IDF and LexRank" IEEE 2024.
[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

