International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

SETU: A Fairness-Aware AI Framework for Optimizing Student-Internship Matching in Large-Scale National Schemes

Meet Bhuva[1], Aditya Kumar Gautam[2], Dharambir Singh Sidhu[3], Dr. NB Prakash[4], Kumar Tanmay[5]

[1,2,3,4,5]School of Computing Science and Engineering (SCOPE) VIT Bhopal University Bhopal, India

Abstract

Internship programs are pivotal for bridging the gap between academic knowledge and industry demands. In India, the Prime Minister Internship Scheme (PMIS) aims to provide one crore internships over five years, yet it faces significant challenges, including a mere 5% conversion rate from application to participation, skill-opportunity mismatches, and systemic biases. This paper introduces SETU (Smart Employment and Training Unification), a novel AI-driven framework designed to overhaul the PMIS matching process. SETU leverages a multi-faceted approach, integrating Natural Language Processing (NLP) for deep resume and job description analysis, advanced embedding models for semantic skill matching, and a predictive analytics module to estimate a candidate's likelihood of joining. A core contribution of our work is the integration of a fairness-aware optimization layer, specifically designed to mitigate geographic and demographic biases, ensuring equitable access for students from underrepresented backgrounds. We propose a scalable, cloud-based architecture that can handle millions of users while providing personalized, fair, and efficient internship recommendations. This system aims to significantly increase the conversion rate, enhance the overall impact of the PMIS, and ensure that the right student is matched with the right opportunity on a national scale.

Recommender Systems, Natural Language Processing, Fairness-Aware Machine Learning, Internship Matching, Predictive Analytics, Skill Extraction, PMIS.

INTRODUCTION

Workforce readiness is a critical determinant of a nation's economic growth. Internships serve as a crucial conduit for transitioning students from academia to the professional world, equipping them with practical skills and industry exposure. Recognizing this, the Government of India launched the ambitious Prime Minister Internship Scheme (PMIS), with the goal of facilitating 10 million internships between 2024 and 2029[4].

However, despite its noble vision, the scheme's initial implementation has revealed deep-seated inefficiencies. In 2024, out of 600,000 applicants for 120,000 available positions, only 30,000 students ultimately joined, translating to a staggering 95% wastage rate and a dismal 5% conversion rate[5]. Analysis of this systemic failure points to several core problems:

- **Skill-Opportunity Mismatch:** A significant crisis where job requirements do not align with student skills, leading to underutilization of human capital[6]. Reports indicate that only 8.25% of Indian graduates find jobs that match their education[7].

- **Information Asymmetry & Crowd Effect:** Students are often unaware of suitable opportunities, and qualified candidates get lost in massive, undifferentiated applicant pools[8].
- **Systemic Biases:** A pronounced geographic bias where urban students disproportionately secure opportunities, leaving rural talent marginalized[9].
- **Operational Bottlenecks:** Inefficient and manual screening processes that cannot scale to handle the high volume of applications[1].

To address these multifaceted challenges, we propose **SETU (Smart Employment and Training Unification)**[2]. SETU is an intelligent platform designed to transform the PMIS from a manual, inefficient system into a smart, fair, and scalable ecosystem. Our framework makes the following key contributions:

- **An Advanced NLP Pipeline:** For parsing unstructured resumes and job descriptions to extract and standardize structured data like skills, projects, and educational qualifications.
- **A Semantic Matching Engine:** That goes beyond simple keyword matching by using contextual embeddings to measure the true alignment between a candidate's profile and an internship's requirements.
- **A Fairness-Aware Optimization Module:** That actively works to counteract systemic biases by re-ranking recommendations to ensure equitable opportunity distribution across different demographics.
- **A Predictive Joining-Likelihood Model:** To identify candidates who are most likely to accept an offer, optimizing the allocation of opportunities and reducing dropouts[3].

This paper details the architecture, methodology, and technological underpinnings of the SETU framework. We will discuss the related work in job recommendation systems, outline our proposed methodology, and present a plan for evaluation.

LITERATURE REVIEW

The domain of job recommendation is well-established, with commercial platforms like LinkedIn and Indeed serving millions of users[4]. These systems primarily rely on a combination of two techniques:

Content-Based Filtering: This method recommends items (jobs) that are similar to items the user has previously shown interest in. For job matching, this involves comparing the text of a user's profile/resume with job descriptions, often using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) to find keyword matches.

Collaborative Filtering: This approach recommends jobs based on the behavior of similar users. If User A and User B have similar profiles and application histories, a job that User A liked is recommended to User B. Modern platforms use matrix factorization or deep neural networks for this[2].

While effective at scale, these traditional systems have limitations in the context of a national social scheme like PMIS. Firstly, they often struggle with the "cold start" problem for new users with no interaction history. Secondly, their optimization goals are typically centered on maximizing engagement (clicks, applications), not societal goals like fairness, inclusivity, or ensuring a high final joining rate.

Recent academic research has focused on leveraging advanced NLP models to overcome the limitations of keyword-based matching. Researchers have used Word2Vec, GloVe, and more recently, transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) to create rich, contextual embeddings of resumes and job descriptions. By representing both candidate and opportunity in a shared vector space, a semantic similarity score (e.g., cosine similarity) can be computed, leading to more accurate matches.

A crucial gap that our work addresses is the integration of fairness. The field of fairness-aware machine learning has gained prominence, highlighting how algorithms can inherit and amplify existing societal biases. Research in this area proposes techniques like algorithmic re-ranking, constrained optimization, and adversarial de-biasing to ensure that recommendation outcomes are equitable across protected attributes like gender, race, or, in our case, geographic location. SETU integrates these principles directly into its matching core, a feature largely absent in conventional job portals.

SYSTEM ARCHITECTURE AND METHODOLOGY

The SETU framework is designed as a modular, end-to-end system. The architecture can be broken down into five primary stages as illustrated by the logical flow of our system.

Data Ingestion and Preprocessing

The system ingests data from two primary sources: **Student Data:** Students upload their resumes in various formats (PDF, DOCX)[7]. **Internship Data:** Postings are aggregated from government portals and other sources using APIs or lightweight web scrapers built with tools like BeautifulSoup[8].

The first step is parsing. We use libraries like PyMuPDF to extract raw text from resume files[7]. This raw text is then fed into a sophisticated NLP pipeline built using spaCy and transformer models[6]. This pipeline performs several tasks:

- Text Cleaning: Removal of special characters, stop words, and irrelevant formatting.
- Named Entity Recognition (NER): A custom-trained NER model identifies and extracts key entities such as skills (e.g., Python, Machine Learning), educational institutions, degrees, project titles, and certifications.
- Structuring: The extracted entities are organized into a structured JSON format representing the student's profile.

A similar process is applied to internship descriptions to extract required skills, experience level, location, and responsibilities.

Semantic Matching Engine

At the heart of SETU is the matching engine, which moves beyond keyword search to understand semantic meaning.

Embedding Generation: We use a pre-trained Sentence-BERT (SBERT) model, which is optimized for generating semantically meaningful sentence embeddings. The structured profile of a student (concatenating skills, project descriptions, etc.) and the structured description of an internship are fed into the SBERT model to generate fixed-size vectors for the student and for the internship.

Similarity Calculation: The alignment between a student and an internship is calculated as the cosine similarity of their respective vectors. This score, ranging from -1 to 1, represents the semantic relevance of the internship to the student's profile. A list of internships is then ranked for each student based on this score.

Fairness and Optimization Layer

This module is our primary contribution to ensuring equitable access. After an initial ranking is generated by the matching engine, the fairness layer adjusts it.

Bias Detection: The system analyzes the distribution of top-ranked opportunities across different demographic and geographic groups (e.g., rural vs. urban, tier-1 vs. tier-3 colleges).

Fair Re-ranking Algorithm: We employ a fair re-ranking algorithm. This algorithm takes the initial ranked list and a set of fairness constraints as input. For example, a constraint could be that in any set of

top 20 recommendations, the proportion of opportunities from non-metro areas should not be below a certain threshold. The algorithm then re-orders the list to satisfy these constraints while minimally impacting the overall relevance score. This ensures that students from underrepresented backgrounds see relevant and accessible opportunities[5].

Joining Likelihood Predictor

To tackle the 95% wastage rate, we built a model to predict the probability that a student will accept an offer if extended. This is a binary classification problem.

Model: A Gradient Boosting model (like LightGBM or XGBoost) is trained on historical data.

Features: The model uses features such as:

- The semantic similarity score from the matching engine.
- Geographic distance between student's location and internship location.
- Stipend information.
- The student's field of study vs. the internship's domain.
- The application density for that particular internship.

The output is a probability score P(join). The final recommendation score is a weighted combination of the similarity score and the joining likelihood, allowing the system to prioritize high-potential matches.

IMPLEMENTATION AND TECHNOLOGY STACK

The SETU platform is architected as a scalable web application.

Frontend: The user interface is a responsive web app built with React.js and styled with Tailwind CSS, ensuring accessibility across devices[4]. It features a simple drag-and-drop interface for resume uploads[7].

Backend & API: The backend logic is implemented in JavaScript using Node.js. It exposes a set of RESTful APIs for communication with the frontend.

Machine Learning: The NLP and ML models are built using the PyTorch framework and libraries from the Hugging Face Transformers ecosystem. Scikit-learn is used for the predictive models.

Database: A PostgreSQL database stores structured user and internship data, while a vector database like FAISS or Elasticsearch is used for efficient nearest-neighbor search on the high-dimensional embedding vectors.

Deployment: The entire system is designed to be deployed on a cloud platform (e.g., AWS, GCP) using containerization (Docker) for scalability and maintainability, capable of handling millions of requests[4].

EVALUATION AND EXPECTED RESULTS

We propose a comprehensive evaluation plan with a multi-pronged approach.

Offline Evaluation:

- **Matching Quality:** Using a curated dataset of student profiles and internship descriptions, we will evaluate the recommendation engine using standard information retrieval metrics like Precision@k, Recall@k, and Normalized Discounted Cumulative Gain (NDCG).
- **Fairness Metrics:** We will measure fairness using metrics like Demographic Parity (ensuring different groups receive a similar number of recommendations) and Equal Opportunity (ensoring candidates who are actually qualified have an equal chance of being recommended, regardless of their group). We will compare the output of our fairness-aware model against a standard relevance-based ranking baseline.

Online Evaluation (A/B Testing): Once deployed, an A/B test would be the gold standard. A control group would receive recommendations from a baseline model, while the test group receives recommendations from the SETU system.

We would track key performance indicators (KPIs) such as application rate, offer acceptance rate, and the final joining/conversion rate. We hypothesize that the SETU system will demonstrate a statistically significant improvement in these KPIs over the baseline.

We expect our system to not only improve the raw number of successful placements but also to show a more equitable distribution of opportunities among students from diverse backgrounds, thereby fulfilling the true vision of the PMIS.

CONCLUSION AND FUTURE WORK

The Prime Minister Internship Scheme holds immense promise for India's youth, but its potential is currently stifled by profound systemic inefficiencies. The SETU framework presented in this paper offers a data-driven, intelligent, and fair solution to this critical national challenge. By combining advanced NLP, semantic matching, and a novel fairness-aware optimization layer, SETU is designed to dramatically increase internship conversion rates, reduce wastage, and ensure that opportunities are distributed equitably.

Our work provides a blueprint for leveraging AI for social good, transforming a high-volume, low-efficiency process into a smart and impactful system.

Future work will focus on extending the platform's capabilities[3]. We plan to develop an employer-facing dashboard to provide analytics on application trends and skill gaps. Further research will go into more sophisticated dropout prediction models and expanding the framework to encompass full-time job opportunities, creating a holistic career development ecosystem for India's youth.

REFERENCES

1. P. K. Swain and S. S. Sahoo, "Skill Mismatch in Indian Labour Market and its Impact on Productivity," SSRN, 2024. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4835942
2. H. Zhang, G. Gu, and J. Sun, "A Research of Job Recommendation System Based on Collaborative Filtering," in 2014 7th International Conference on Intelligent Computation Technology and Automation, 2014, pp. 241-244. Available: https://ieeexplore.ieee.org/document/7064250
3. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019. Available: https://arxiv.org/abs/1908.10084
4. S. K. Singh, M. K. Singh, and S. Singh, "Job Recommendation System Using Content-Based and Collaborative Filtering," in Innovations in Computer Science and Engineering, Springer, 2020. Available: https://link.springer.com/chapter/10.1007/978-981-15-5347-8_45
5. B. D. T. D. P. K. Singh, M. C. B. C. P. B. S. S. Verma, "Fairness in Recommendation Systems: A Survey," ACM Computing Surveys, vol. 54, no. 8, pp. 1–38, 2021. Available: https://dl.acm.org/doi/10.1145/3462722
6. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. Available: **https://arxiv.org/abs/1810.04805**

7. Sharma, S. Sharma, and P. Sharma, "Automated Resume Parsing and Skill Extraction Using Natural Language Processing," in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1198-1204. Available: **https://ieeexplore.ieee.org/document/9432247**
8. "Drishti IAS: Strengthening India's Innovation Landscape," Drishti IAS. Available: **https://www.drishtiias.com/daily-updates/daily-news-editorials/strengthening-india-s-innovation-landscape**
9. "Times of India: Why are students turning away from the PM internship scheme?" The Times of India. Available: **Reference**