Optimizing Resume Compatibility for Automated Screening: Strategies for Enhanced Job Placement Success

Pat Chimtanoo, Meihui Guo, Nawat Swatthong

Abstract

This project aims to strategically improve job placement success rate by refining resume compatibility within automated screening systems. The primary objective is to identify key factors that increase a resume's visibility and provide guidance on optimizing applicants' resumes. Our approach involves developing a resume matching score system, which is based on three components: technical skills and soft skills extraction (SkillNER), academic qualification extraction (NER), and a network analysis of industry-specific skills and academic fields, derived from an extensive review of over 2000 data scientist and data analyst job postings. The results indicate a significant differentiation in resumes, especially regarding technical skills and academic qualifications. The project's major contribution is the development of a quantifiable system that evaluates the impact of various resume components on job compatibility, offering precise guidance for applicants to tailor their resumes effectively. However, the current challenge is validating the scoring system's accuracy and practical applicability, as it is more theoretical at this stage. Future work will focus on obtaining labeled data to correlate our theoretical findings with real-world job application outcomes, thereby closing the gap between academic research and practical job market demands.

Key words: Resume Matching, Resume Scoring System, NER, SkillNER
Github link: https://github.com/nawatswatthong/resume_job_posting_matching

Introduction

Traditional resume screening process involves manually reviewing candidates' experience and qualifications, followed by shortlisting suitable candidates for interviews. The ineffectiveness and time consuming of this process leads to the adoption of recruiting software such as Application Tracking System (ATS), which are currently adopted by more than 90 percent of employers (Handerson, 2023), to improve the efficiency of the hiring process.

The objective of this project is to investigate and identify the factors that make a resume more likely to be noticed during job applications. The team aims to formulate strategies that enhance resume compatibility, thereby increasing its likelihood of a higher rank and successful job placement in automated screening processes. By understanding the resume screening process, our team can create strategies to equip job applicants with essential tools, enhancing their job placement prospects.

Data Preprocessing

The following data sets have been preprocessed and used in this project:

1) Job Posting Data Set, compiled by an anonymous reddit user (2022), features 2663 job postings for data scientist and data analyst stored in a html format. It contains position title, company names, and detailed job description. To make the data more accessible, we processed it using the htmlaundry library, removing the majority of HTML tags and language for clearer analysis.

2) Resume Corpus (Jiechieu and Tsopze, 2020) contains a set of resume files with extension .txt. For our project, which concentrates on developing a matching algorithm for data scientist and data analyst roles, we selectively extracted resumes specifically tailored to these positions. This was achieved by filtering out resumes that did not include the terms "data scientist" or "data analyst." From this process, we obtained 672 targeted resumes. To facilitate information extraction, we then organized each resume into three distinct sections: summary, education, and experience.

3) The University Rank Data, available on Kaggle and sourced from the Center for World University Rankings in Saudi Arabia (O'Neill, 2014), catalogs the top 1000 universities worldwide. We utilized this dataset to assess university performance, a factor that contributes to evaluating the qualifications of candidates.

Methodology

The algorithms for resume screening in ATS software are highly varied, and developers often do not disclose their specific workings, making it challenging to replicate these algorithms accurately for the actual screening process. The team decided to develop a resume matching algorithm to replicate the screening process, utilizing job requirements and necessary skills extracted from available job postings. Although this project will only focus on data scientist and data analyst roles, the same framework could be extended and applied to other positions.

In general, the framework for constructing the algorithms, as outlined by Mohamed et al. (2018), involves leveraging resume ontology to categorize key components and highlight its features prior to information extraction. Typically, a resume comprises education and work experience sections. Accordingly, our team has proposed an aggregated resume matching score formula as described in Figure 1, consisting of 3 score components: academic qualifications extraction, work experiences and skills Extraction, and Network analysis.

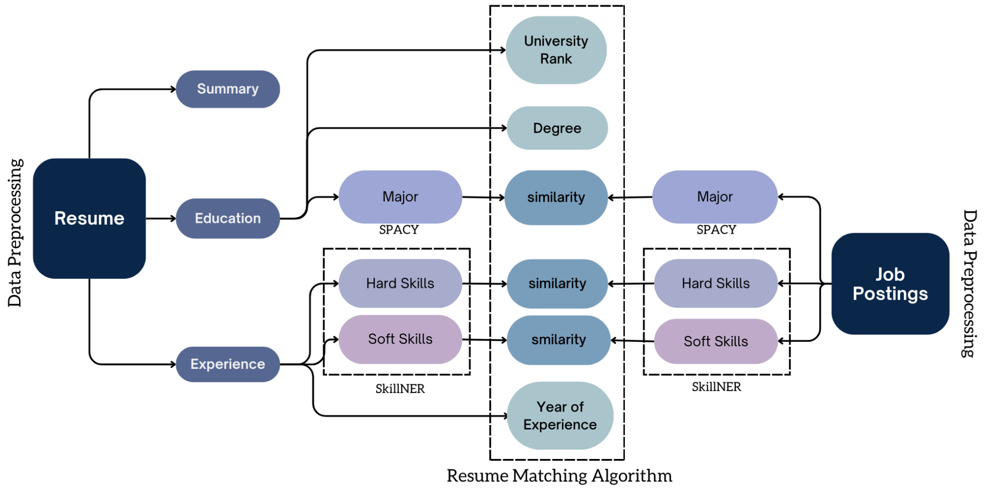


Figure 1: Resume Matching Algorithm Flowchart

Academic Qualification Extraction

Given that the majority of resumes in our dataset are in plain text format, a multi-step preprocessing approach is essential for effective information extraction. This involves the removal of stop words, as well as the application of word stemming and lemmatization techniques. These steps aim to tokenize the entire text of each resume, allowing for the extraction of key details such as educational background, grades, work experience, and skill sets (Roy, Chowdhary & Bhatia, 2020). The information on academic qualifications has been tokenized utilizing the Name Entity Recognition (NER) method from Spacy library, which is the process of locating named entities in unstructured text and then classifying them into predefined categories, such as person names, organizations, locations, monetary values, percentages, and time expressions.

After preprocessing, we used NER to extract details such as field of study preferences, degrees, and institution names from resume datasets. For the job posting , the only information available to use would be the relevant fields of study. These fields of study will contribute to calculating similarity scores between job applicants and a network of preferred study areas, as derived from data scientist job descriptions. In our matching algorithm, we will integrate these similarity scores along with university rankings and degree information to enhance the precision of candidate-job alignment.

Work Experiences and Skills Extraction

Extracting skills from work experience presents a significant challenge, as tokenization only marginally aids the process. Skills are complex entities and uncommon in nature and it requires more than recognizing language properties for an effective detection. To tailor this task, skill extraction should focus on contextual analysis and the identification of key indicators (Fareri et al, 2021).

Another challenge arises from the existence of a skills ontology, which outlines the interrelationships between skills and experiences within the job search context (Guo, Alamudun & Hammond, 2016). Considering potential challenges, we employ the SkillNER library, an enhancement of NER that integrates a skills taxonomy, to extract both technical and soft skills from job descriptions and resumes.

Furthermore, we calculate years of experience by extracting and summing the durations of work experiences, listed in 'Month, Year' format in resumes. Similar to handling academic qualifications, this information will feed into our matching algorithm, where a network of skills is constructed to calculate similarity score.

Network Construction

The conventional similarity scores, like Jaccard or Cosine, are adept at identifying memberships but fail in capturing the prevalence and relational dynamics between different entities. To address this gap, we aim to develop a Network-Based Similarity Score (NBSS) to improve the precision of our resume matching algorithm.

Our approach involves constructing a network of relevant skills and preferred academic backgrounds specifically for data scientist and data analyst roles. This network is built based on a count matrix C, which tabulates the frequency of various skills and fields of study as mentioned in job descriptions. Each time a set of elements appears in X_k, an extracted information of job description k, the count for each item in the count matrix C_{ij} is incremented, where i and j represent the index element in X_k as shown in Equation 1. As we process subsequent job descriptions, we continue to aggregate these counts.

$$C_{ij} = \begin{cases} \Sigma_{k=1}^{K} \mathbb{I}\left(x_i \in X_k \wedge x_j \in X_k\right) & \text{if } i \neq j \\ \Sigma_{k=1}^{K} \mathbb{I}(x_i \in X_k) & \text{if } i = j \end{cases}$$

Equation 1: Count Matrix for Skills and Field of Study Formula

The count matrix is then transformed to a normalized adjacency matrix to help reduce the frequency skewness using the following equation:

$$A_{ij} = \frac{1}{ln(C_{ij} + 1)}$$

Equation 2: Adjacency Matrix for Skills and Field of Study Formula

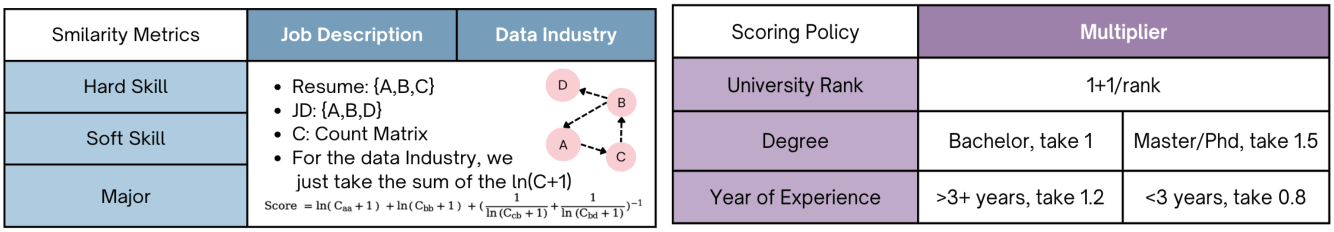
Next, we construct three undirected network graphs, each representing preferred fields of study, soft skills, and technical skills. These graphs are built from the adjacency matrices extracted from the job description data set in the previous section with normalized frequency as node attribute.

These comprehensive networks will not only illustrate the prevalence and interconnections of skills and academic fields within the industry but also enhance our matching algorithm. By utilizing network properties such as closeness centrality measures and node attributes, we can more effectively match candidates to job requirements, leveraging the nuanced relationships and significance of different skills and fields of study in our analysis.

Resume Similarity Score

The resume similarity score is constructed by integrating technical skills, soft skills, and academic qualifications from the resume with the network of industry specific requirements from the previous section. The base score is a measure of how closely candidates' qualifications align with job requirements, quantified using normalized frequencies of each skill from network graphs and can be calculated both for a specific job posting and overall match for data scientist position. When skill/major A appears both in a resume and in job posting K, the score assigned to that skill is its normalized frequency. In cases where skill/major A is not directly listed in a particular job posting, the candidate will receive a score determined by the inverse of the shortest path length, derived from the adjacency matrix from the previous section, between A and any skill/major that is present in the job description. This method reflects the core principles of hiring processes, emphasizing skill and background compatibility and allowing candidates to achieve higher scores by recognizing the transferable value of their experience.

Additionally, university rankings, the highest level of educational attainment, and years of experience are incorporated as multipliers to the base score. We believe these factors, while not mandatory, enhance a candidate's profile, acting as an advantage in the evaluation process. The summary of the resume scoring system is shown in Figure 2.



Smilarity Metrics	Job Description	Data Industry
Hard Skill	• Resume: {A,B,C} • JD: {A,B,D} • C: Count Matrix • For the data Industry, we just take the sum of the ln(C+1) $\text{Score} = \ln(C_{aa}+1) + \ln(C_{bb}+1) + (\frac{1}{\ln(C_{cb}+1)} + \frac{1}{\ln(C_{bd}+1)})^{-1}$	
Soft Skill		
Major		

Scoring Policy	Multiplier	
University Rank	1+1/rank	
Degree	Bachelor, take 1	Master/Phd, take 1.5
Year of Experience	>3+ years, take 1.2	<3 years, take 0.8

$$\text{Total Score} = \sum S_i * R * D * \text{YoE} \quad \text{where } i \in \{\text{Hard\_skill, Soft\_skill, Major}\}$$

Figure 2: Resume Matching Scoring System

Results and Discussions

This section delves into the analysis and interpretation of data derived from our resume matching algorithm. We focus on how integrating network graphs with our scoring system provides a nuanced insight into the connection among skills and the specific demands of the job market, particularly in the fields of data science and data analytics.

Resume Matching Score

In our evaluation of the scoring system, we selected two distinct resumes for analysis. These resumes differ significantly in academic and professional backgrounds. Figure 3 illustrates the capability of our scoring system to differentiate effectively between these resumes, particularly in the terms of technical skills and academic fields, which are reflected in the overall scores.

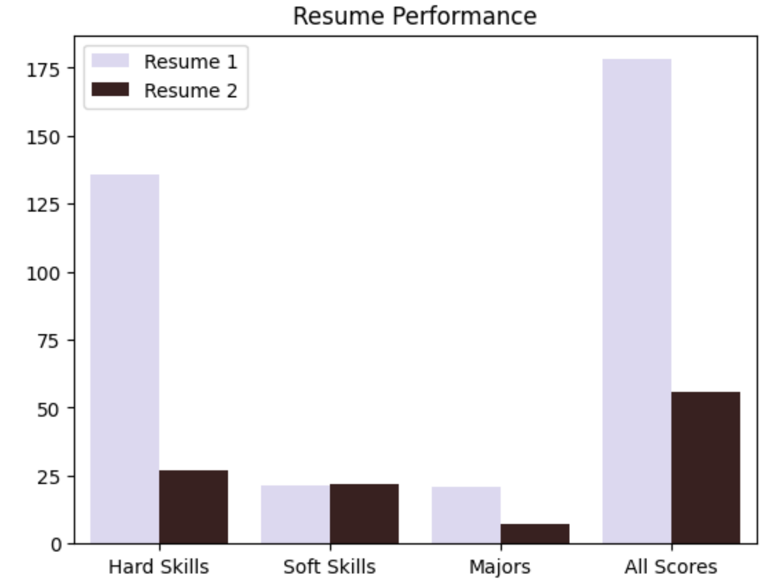


Figure 3: Resume Performance Comparison

A deeper examination of the technical skill scores is presented in Figure 4. Here, we detail the skills and their respective sub-scores for both resumes. It is evident that Resume 1, consisting of key skills like Python, SQL, and analytics, achieves a higher score compared to Resume 2.

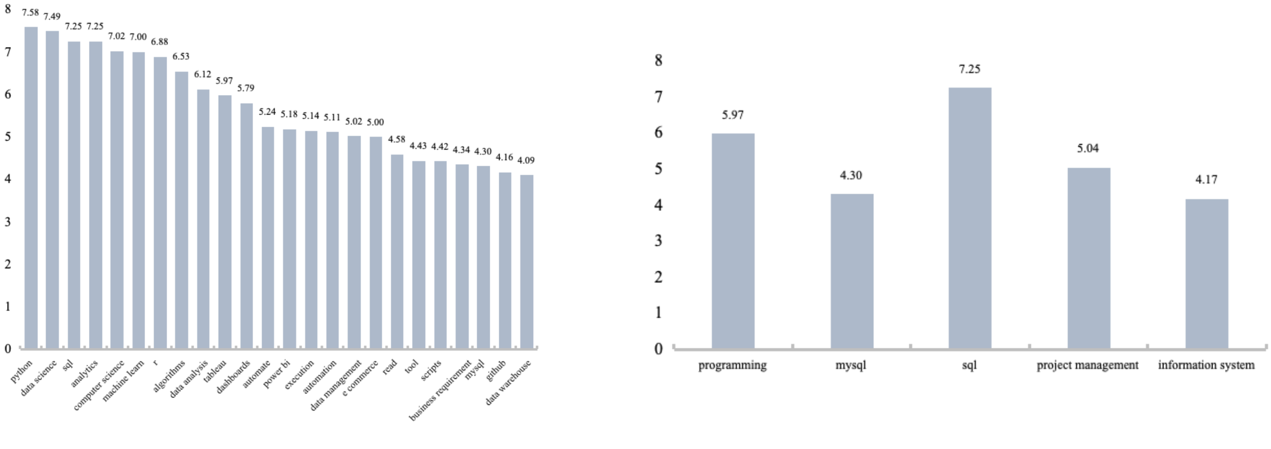


Figure 4: Detailed Breakdown of Technical Skill Scores in the Resumes

Further, we analyze the scores across three network domains – technical skills, soft skills, and academic majors – and the total scores for each resume when applied to a specific job role. This is visualized in Figure 5, which plots the distribution of these elements across 672 resume samples. Our findings highlight the dominance of technical skills in contributing to the overall score in data science and analytics roles. While variations in soft skills and academic majors are subtler, they are not insignificant. This suggests a stronger emphasis on technical prowess in the hiring process. Interestingly, while the highest score exceeded 700, the bulk of scores hovered around 100, indicating a general alignment with job requirements. Nevertheless, to truly stand out, candidates should consider enhancing their resumes with additional pertinent skills.

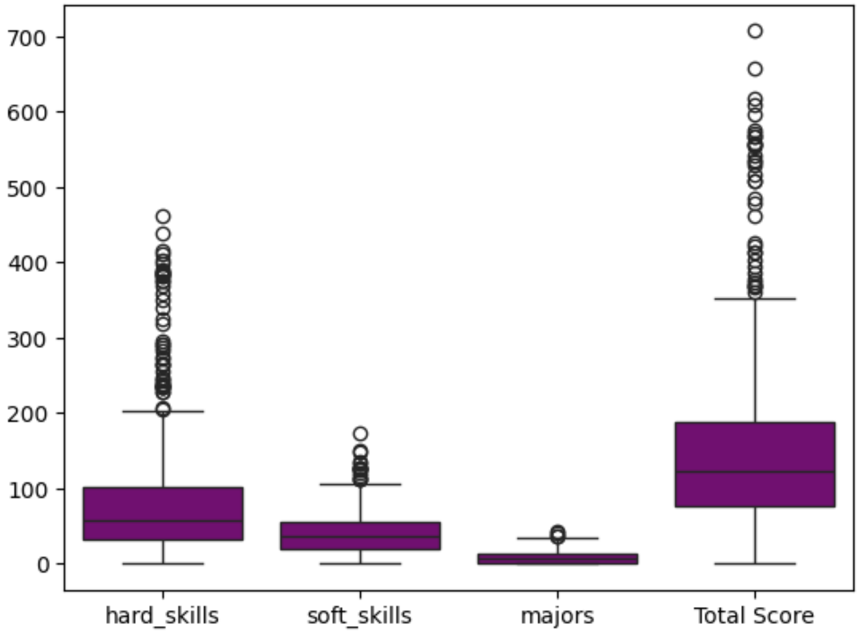
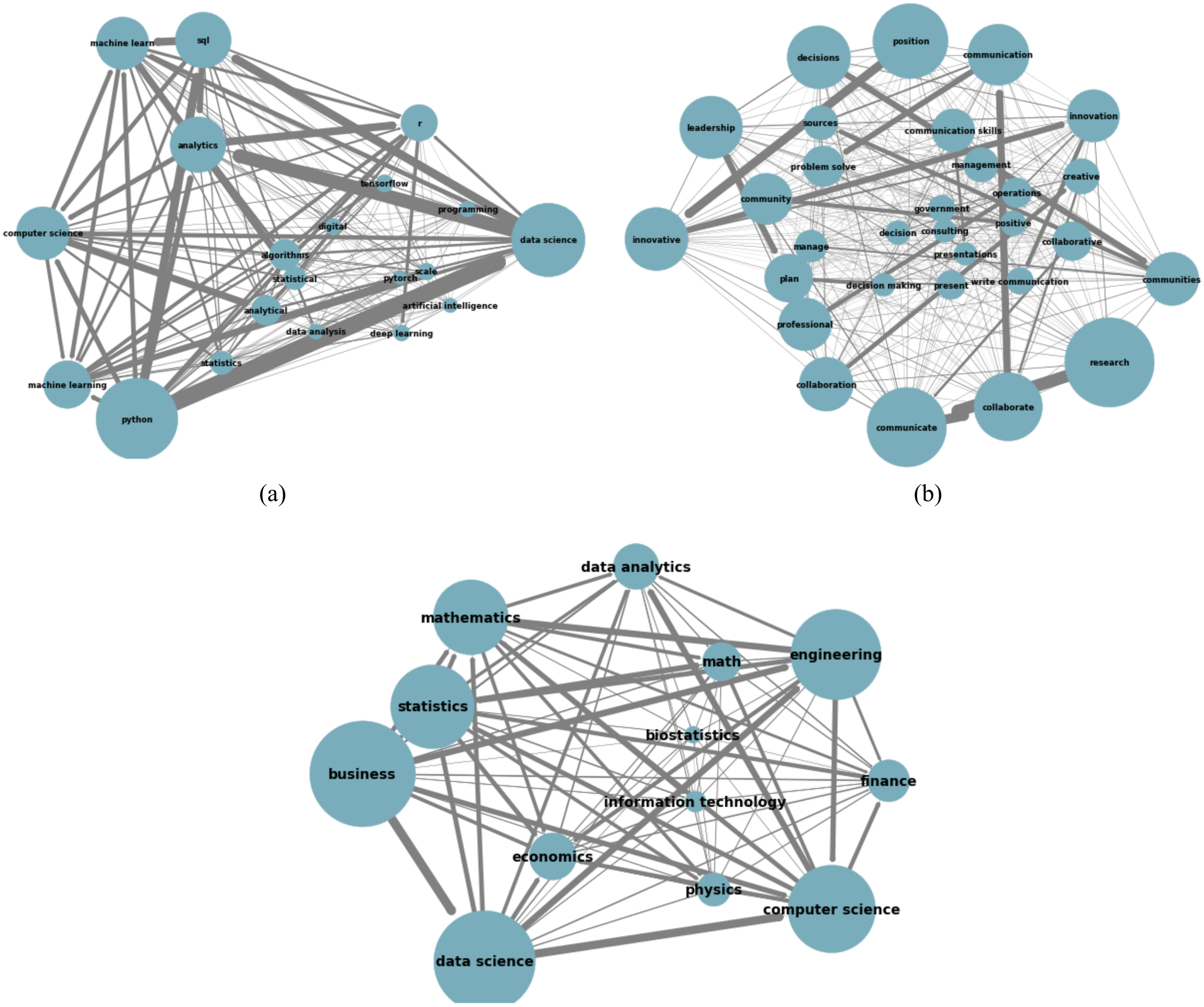


Figure 5: Distribution of all score components

Texts Network Analysis from Job Postings

To broaden our understanding of job market preferences, we analyzed the constructed networks from job postings. As shown in Figure 6, the networks highlight key attributes through larger nodes and connections. The technical skill network underscores the importance of “Python”, “data science”, and “SQL” in data scientist and analyst job postings. Conversely, the soft skill network emphasizes skills like “research”, “communication”, and “collaboration”. However, the centrality of these soft skills is less visible, appearing more as sequential chains than dominant nodes. The major network reveals a strong centrality in engineering, linked to key fields such as business, data science, computer science, and mathematics. Table 1 enumerates the most frequent terms in each network, based on their normalized occurrence.



(a)

(b)

(c)

Figure 6: Network analysis of the (a) technical skills, (b) soft skills and (c) academic majors

Network	Top 5 words
Technical Skills	“python”, “data science”, “sql”, “analytics”, “computer science”
Soft Skills	“research”, “communicate”, “position”,“collaborate”, “decisions”
Academic Majors	“business”,“data science”,“engineering”, “computer science”, “statistics”

Tabel 1: Top words in the three dimensions

Conclusions

In conclusion, our study provides an ability to discern between effective and ineffective resumes for specific job roles. Our methodology allows us to quantify the contribution of various parts of a resume, identifying which skills or qualifications most significantly impact the suitability score for a particular role. This insight enables us to offer targeted advice to applicants on which areas they should focus on improving to better align with the job requirements.

Furthermore, the potential of this model extends beyond individual job applications. It offers a framework for understanding the broader job market's requirements, guiding job seekers in prioritizing the skills and qualifications that are most valued across various roles. This could significantly enhance the efficiency of job preparation and the relevance of applicants in the competitive job market.

However, our study is not without limitations. The primary challenge lies in verifying the accuracy and real-world applicability of our scoring system. Currently, the system's effectiveness is hypothesized rather than empirically validated. Future work will focus on acquiring labeled data that indicates the success or failure of job applications. This data will be invaluable in training and refining our scoring system, allowing us to provide more accurate and effective recommendations. By doing so, we aim to bridge the gap between theoretical analysis and practical application, ensuring that our findings are not only insightful but also directly applicable to the real-world scenarios faced by job applicants.

References

Data Set of Job Description for Your Pleasure. (2022). Reddit. https://www.reddit.com/r/datasets/comments/w340kj/dataset_of_job_descriptions_for_your_pleasure/

Fareri, S., Melluso, N., Chiarello, F., & Fantoni, G. (2021). SkillNER: Mining and mapping soft skills from any text. Expert Systems with Applications, 184, 115544. https://doi.org/10.1016/j.eswa.2021.115544

Guo, S., Alamudun, F., & Hammond, T. (2016). RésuMatcher: A personalized résumé-job matching system. Expert Systems with Applications, 60, 169-182. https://doi.org/10.1016/j.eswa.2016.04.013

Henderson, R. (2023, May). What Is An ATS? 8 Things You Need to Know About Applicant Tracking Systems. Retrieved from https://shorturl.at/pqFPT

Jiechieu, K. F. F., & Tsopze, N. (2020). Skills prediction based on multi-label resume classification using CNN with model predictions explanation. Neural Computing and Applications. https://doi.org/10.1007/s00521-020-05302-x

Mohamed, A., Bagawathinathan, W., Iqbal, U., Shamrath, S., & Jayakody, A. (2018). Smart Talents Recruiter - Resume Ranking and Recommendation System. In 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS) (pp. 1-5). IEEE. https://doi.org/10.1109/ICIAFS.2018.8913392

O'Neill, M. (2014) World University Rankings. Kaggle. https://www.kaggle.com/datasets/mylesoneill/world-university-rankings

Roy, P. K., Chowdhary, S. S., & Bhatia, R. (2020). A Machine Learning approach for automation of Resume Recommendation system. Procedia Computer Science, 167, 2318-2327. https://doi.org/10.1016/j.procs.2020.03.284