Ai Powered Resume Screening

Team: Capsule Corporation

1. Introduction

The goal of this project was to develop a comprehensive HR candidate screening system that evaluates candidate profiles based on a variety of factors such as fraud detection, network strength, experience, and skill synergy. By utilizing advanced natural language processing (NLP) techniques and graph-based network analysis, this system provides a holistic evaluation of candidates, identifies potential risks, and supports data-driven HR decision-making.

The system is divided into three main components:

- Fraud Detection System (Risk Score Calculation)
- Network Connection Analysis (Network Connection Score)
- HR Decision Support Dashboard

The final output is an HR dashboard that provides a side-by-side comparison of candidates, highlights red flags, ranks candidates, and visually represents the network strength of each individual.

2. Methodology and Models

2.1. Using RAG LLM with Google Gemini for Our Project

In this project, we implemented a Retrieval-Augmented Generation (RAG) approach using Google Gemini to extract structured data from unstructured resumes. RAG combines information retrieval with generative AI, allowing us to process domain-specific data effectively.

Approach Overview :

Retrieval Layer: We used “pdfplumber” to extract resume text from PDFs, serving as input for the generative model.

Generation with Google Gemini: The extracted text was passed to Google Gemini, which generated structured JSON output containing key resume details like total years of experience and skills. A retry mechanism ensured accurate and valid output.

Augmentation: Custom JSON parsing and validation mechanisms cleaned the model's output and handled malformed data, ensuring consistency.

Benefits

- Gemini focused on relevant content from the retrieved text, ensuring high-quality structured output.
- The RAG approach allowed flexible handling of resumes with varying formats and technical terms.
- This method automated large-scale resume processing, reducing time and effort while improving robustness.

2.1. Fraud Detection System (Risk Score Calculation)

The fraud detection system identifies inconsistencies and potential risks in candidate profiles by calculating a Risk Score based on three main components:

1.Text Similarity Score:

- Measures the similarity between the candidate's resume and their recommendation letters using TF-IDF embeddings and Cosine Similarity.
- Weight: 50% of the Risk Score.

2.Vague Words Score:

- Detects the presence of vague or generic terms (e.g., "hard-working," "great potential") in the recommendation letters.
- Indicates a lack of specificity and lowers recommendation reliability.
- Weight: 30% of the Risk Score.

3.Reciprocal Detection Score:

- Analyses mutual endorsements or recommendation reciprocity between candidates.
- High reciprocal endorsement scores can indicate biased or manipulated recommendations.
- Weight: 20% of the Risk Score.

The Risk Score was calculated as a weighted sum of the three components:

$$\text{Risk Score} = 0.5 \times \text{Text Similarity Score} + 0.3 \times \text{Vague Words Score} + 0.2 \times \text{Reciprocal Detection Score}$$

Since the data was not labelled, a supervised learning approach could not be used for calibration. The weights were selected based on domain knowledge and expert judgment. If labelled data had been available, we would have optimized these weights using techniques such as k-fold cross-validation and metrics like the F1 Score.

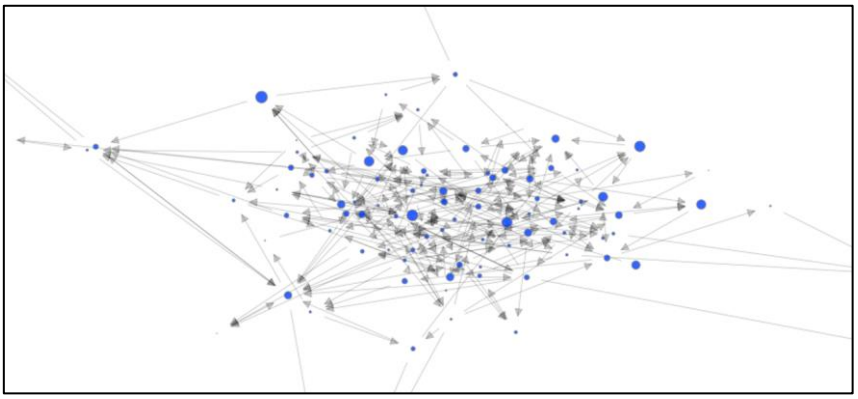
4.Outlier Detection:

The Risk Score was used to label resumes as fraudulent or non-fraudulent using boxplot-based outlier detection. Resumes with Risk Scores significantly higher than the median were marked as potential fraud cases.

2.2. Network Connection Score

The Network Connection Score evaluates the quality and strength of a candidate's professional network. It is calculated using a combination of centrality measures and community detection algorithms to quantify the candidate's influence and community integration

Largest Community (Community 6) Recommendation Network



Centrality Scores:

Centrality scores are used to measure the importance and influence of each candidate within the professional network. The following types of centrality were calculated for each candidate:

1.Betweenness Centrality:

- Measures the frequency with which a candidate appears on the shortest paths between other nodes in the network.
- High betweenness indicates that the candidate is an intermediary or bridge within the network, controlling the flow of information.

2.Closeness Centrality:

- Measures how close a candidate is to all other nodes in the network.
- Candidates with high closeness centrality can reach others quickly, making them well-connected in the network.

3.In-Degree and Out-Degree Centrality:

- In-Degree Centrality captures the number of incoming connections (e.g., endorsements or recommendations).
- Out-Degree Centrality captures the number of outgoing connections (e.g., the candidate endorsing others).

These measures indicate the candidate's visibility and influence. Each centrality score was normalized and given an equal weight:

$$\text{Centrality Score} = 0.2 \times \text{Betweenness} + 0.2 \times \text{Closeness} + 0.2 \times \text{In-Degree} + 0.2 \times \text{Out-Degree}$$

Community Score:

Communities were detected using the Louvain method, which groups nodes into highly connected clusters. The Community Score measures how strongly integrated the candidate is within their community. Candidates who are central in their communities receive a higher score.

Final Network Connection Score:

The final Network Connection Score is a combination of centrality scores and the community score, weighted as follows:

$$\text{Network Connection Score} = 0.8 \times \text{Centrality Score} + 0.2 \times \text{Community Score}$$

2.3. HR Decision Support Dashboard

The HR Dashboard provides a side-by-side comparison of candidates, highlighting their strengths and weaknesses using the following metrics:

1.Risk Score (30%):

- A higher weight is given to the Risk Score to penalize candidates who show signs of potential fraud or inconsistency.
- Candidates with a high Risk Score should be pushed down in the ranking.

2.Network Connection Score (30%):

- Network strength and community integration are critical indicators of a candidate's reputation and influence in their field.
- A higher weight is appropriate for positions that involve collaboration, leadership, or business development.

3.Experience Years (20%):

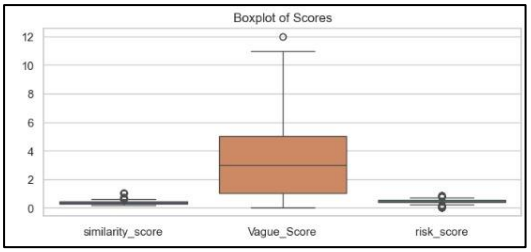
- Experience should be considered, but not given excessive weight. Too much emphasis on years alone may favour candidates with long tenures but outdated skills.
- Consider increasing this weight for roles that require extensive experience.

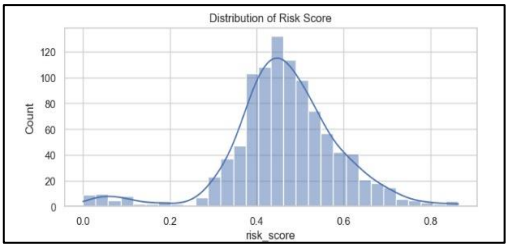
4.Synergy Score (20%):

- Measures the reliability of the candidate's skill set. High synergy scores should boost the candidate's ranking.
- Synergy is particularly valuable for technical roles where skill proficiency is critical

3. Key Findings and Performance Metrics

- Risk Scores successfully flagged candidates with high reciprocal endorsements and generic recommendation letters.
- Network Connection Scores provided insights into candidate influence and community integration. Candidates with strong connections were ranked higher.
- Skill Synergy Scores identified candidates who had strong alignment between their listed skills and practical experience, giving HR teams a clearer picture of skill reliability.
- The HR Dashboard enabled the HR team to make data-driven decisions, ranking candidates based on the calculated scores.





4. Discussions of Bias, Fairness, and Ethical Considerations

- Bias in Text-Based Models: Applied neutral language processing techniques and used bias detection tools to mitigate gender or racial biases in text similarity scores.
- Fairness in Network Analysis: Implemented a community detection step to ensure that candidates are compared within their peer groups.

5. Scalability and Optimizations

- Used parallel processing for text extraction and similarity calculations.
- Implemented efficient community detection algorithms that scale well for larger networks.

6. Conclusion

This project successfully developed a data-driven HR Candidate Screening System that integrates fraud detection, network analysis, and skill evaluation. The final HR Dashboard supports transparent and fair hiring decisions, ranking candidates based on their network strength, experience, and fraud risk.