AI-powered Resume Screening and Ranking System

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning
with
TechSaksham – A joint CSR initiative of Microsoft & SAP

by

Aryan Sachin Shitole, aryan0543f@gmail.com

Under the Guidance of

Saomya Chaudhary

ACKNOWLEDGEMENT

We would like to take this opportunity to express our deep sense of gratitude to all individuals who helped us directly or indirectly during this thesis work.

First and foremost, we extend our heartfelt thanks to our project supervisor Saomya Chaudhary, for their invaluable guidance, support, and expertise throughout the development process. Their insights and encouragement were instrumental in shaping the direction of this project.

Special thanks to the organizations and individuals who provided access to datasets and resources, enabling us to train and validate our models effectively. Your willingness to share valuable information has significantly enhanced the depth of our research.

I am grateful to the academic and professional communities whose research and publications laid the groundwork for our project. The insights gained from existing literature on machine learning, natural language processing, and recruitment technologies were invaluable in guiding our approach.

ABSTRACT

AI-Powered Resume Screening and Ranking System: Summary

Problem Statement:

The traditional resume screening process is often time-consuming and prone to biases, leading to suboptimal hiring decisions. This project aims to address these issues by developing an AI-powered system that enhances the efficiency and fairness of candidate evaluation.

Objectives:

- To create a system that utilizes advanced AI algorithms for resume screening and ranking.

- To reduce biases in the hiring process by implementing machine learning techniques.

- To improve the accuracy of candidate assessments in relation to job descriptions.

Methodology:

- **Data Collection:** Gather a diverse dataset of resumes and job descriptions.

- **Model Development:** Employ machine learning techniques, including natural language processing (NLP) and stacked models, to analyze and rank resumes based on contextual relevance.

- **Evaluation:** Test the system's effectiveness by comparing its performance against traditional screening methods, focusing on accuracy and bias reduction.

Key Results:

- The AI system demonstrated a significant improvement in the ranking accuracy of resumes compared to conventional methods.

- A reduction in bias was observed, leading to a more equitable selection process.

- The system facilitated quicker decision-making, allowing recruiters to focus on the most qualified candidates.

Conclusion:

The AI-powered Resume Screening and Ranking System represents a significant advancement in recruitment technology. By leveraging machine learning, it not only streamlines the hiring process but also promotes fairness, ultimately enhancing the overall quality of candidate selection for employers while providing job seekers with a more equitable opportunity. Integrating AI into recruitment can lead to better hiring outcomes, making it a valuable tool for organizations seeking to optimize their talent acquisition strategies.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Problem Statement:

The problem being addressed is the inefficiency and bias present in traditional resume screening processes. Recruiters often face overwhelming volumes of applications, leading to a reliance on manual screening methods that can be time-consuming and subjective. Many systems primarily use keyword matching, which can overlook qualified candidates whose resumes may not contain the exact terms used in job descriptions. This approach can result in the exclusion of talented individuals and perpetuate biases based on factors such as gender, ethnicity, or educational background, ultimately affecting the diversity and quality of the candidate pool.

Significance:

- Affects the quality of hires and organizational performance

- Impacts employer branding and reputation

- Limits diversity and innovation

- Results in unfair treatment of candidates

1.2 Motivation:

This project was chosen due to the pressing need for innovation in the recruitment process, which has remained largely unchanged despite advancements in technology. Traditional methods are often inefficient, biased, and unable to keep pace with the growing volume of applications in today's job market. By developing an AI-powered Resume Screening and Ranking System, the project aims to address these challenges, streamline the hiring process, and promote fairness in candidate evaluation.

Potential Applications:

1. **Corporate Recruitment:** Organizations can implement the system to enhance their hiring processes, ensuring that they identify the best candidates efficiently and equitably.

2. **Recruitment Agencies:** Agencies can utilize the technology to improve their service offerings, providing clients with faster and more accurate candidate recommendations.

3. **Human Resource Management Systems (HRMS):** Integration with existing HRMS platforms can enhance their functionality, allowing for automated resume screening and ranking.

4. **Diversity and Inclusion Initiatives:** Companies focused on improving diversity can use the system to minimize biases in hiring, helping to create a more inclusive workforce.

5. **Job Boards and Platforms:** Online job platforms can adopt the technology to improve the matching process between candidates and job openings.

1.3 Objective:

1. Enhance Screening Efficiency:

Develop an automated system that significantly reduces the time and effort required for resume screening, allowing recruiters to process a larger volume of applications quickly.

2. Improve Candidate Assessment Accuracy:

Utilize advanced machine learning algorithms and natural language processing (NLP) to analyze resumes in relation to job descriptions, ensuring a more accurate evaluation of candidate qualifications and fit.

3. Reduce Bias in Hiring:

Implement techniques that minimize biases in the screening process, promoting fairness and equity in candidate selection regardless of gender, ethnicity, or educational background.

1.4 Scope of the Project:

1. Integration:

The project will include the development of APIs or plugins to facilitate integration with existing Human Resource Management Systems (HRMS) and applicant tracking systems (ATS).

2. Data Handling:

The system will handle a diverse dataset of resumes and job descriptions across various industries to ensure broad applicability and effectiveness.

3. Bias Mitigation:

The project will implement techniques aimed at reducing bias in the screening process, promoting fairness in candidate evaluation.

4. User Interface:

A user-friendly interface will be developed for recruiters to easily interact with the system, view candidate rankings, and access insights.

5. Target Users:

The system is designed for use by corporate recruiters, recruitment agencies, and HR professionals seeking to streamline their resume screening processes.

CHAPTER 2
Literature Survey

2.1 Review relevant literature or previous work in this domain.

Gangoda, Nikethani, et al. "Resume Ranker: AI-Based Skill Analysis and Skill Matching System." *2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS)*. **IEEE, 2024.**

2.2 Mention any existing models, techniques, or methodologies related to the problem.

1. Natural Language Processing (NLP) Techniques:

- Text Parsing and Tokenization: Techniques for breaking down resumes into manageable components (tokens) to analyze content effectively.
- Named Entity Recognition (NER): Used to identify and classify key information such as skills, education, and work experience from resumes.
- Word Embeddings: Models like Word2Vec, GloVe, and FastText that convert words into vector representations, capturing semantic relationships and contextual meanings.

2. Machine Learning Models:

- Support Vector Machines (SVM): Often used for classification tasks, SVM can be applied to categorize resumes based on their relevance to job descriptions.
- Random Forests: An ensemble learning method that can be used for ranking candidates based on multiple features extracted from resumes.
- Gradient Boosting Machines (GBM): Another ensemble technique that can improve prediction accuracy by combining weak learners.

3. Deep Learning Approaches:

- Recurrent Neural Networks (RNNs): Useful for processing sequential data, RNNs can analyze resumes as sequences of words to capture context and relationships.
- Convolutional Neural Networks (CNNs): While traditionally used for image processing, CNNs can also be adapted for text classification tasks, including resume screening.
- Transformers: Models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have shown great promise in understanding context and semantics in text, making them suitable for resume analysis.

4. Skill Matching Algorithms:

- Cosine Similarity: A technique used to measure the similarity between two non-zero vectors, often applied to compare the skill sets in resumes against job descriptions.
- TF-IDF (Term Frequency-Inverse Document Frequency): A statistical measure used to evaluate the importance of a word in a document relative to a collection of documents, useful for identifying key skills in resumes.

5. Bias Mitigation Techniques:

- Fairness-Aware Algorithms: Approaches that incorporate fairness constraints into the model training process to reduce bias in candidate selection.
- Adversarial Debiasing: A technique that uses adversarial training to minimize bias in machine learning models by training them to be invariant to sensitive attributes.

6. Existing Systems and Frameworks:

- Applicant Tracking Systems (ATS): Many ATS platforms incorporate basic keyword matching and ranking algorithms, serving as a foundation for more advanced AI-driven solutions.
- AI-Powered Recruitment Tools: Tools like HireVue and Pymetrics use AI to assess candidates through video interviews and gamified assessments, providing insights into candidate fit beyond resumes.

2.3 Highlight the gaps or limitations in existing solutions and how your project will address them.

Gaps/Limitations in Existing Solutions:

1. **Over-Reliance on Keyword Matching:**

   - Many existing systems prioritize exact keyword matches, failing to capture contextual relevance or synonymous terms (e.g., "ML" vs. "machine learning"). This leads to qualified candidates being overlooked.

2. **Bias in Training Data:**

   - Models trained on historical hiring data often inherit biases (e.g., gender, ethnicity, or educational bias), perpetuating unfair hiring practices.

3. **Lack of Contextual Understanding:**

   - Basic NLP models struggle with nuanced language, industry-specific jargon, or unconventional resume formats, reducing accuracy.

4. **Limited Integration with HR Systems:**

   - Many AI tools operate in isolation and lack seamless integration with existing Applicant Tracking Systems (ATS) or HR platforms.

5. **Static Models:**

   - Traditional systems do not adapt to evolving job market trends or new skill requirements, leading to outdated evaluations.

6. **Transparency and Explainability:**

   - "Black-box" AI models provide limited insights into how decisions are made, reducing recruiter trust and accountability.

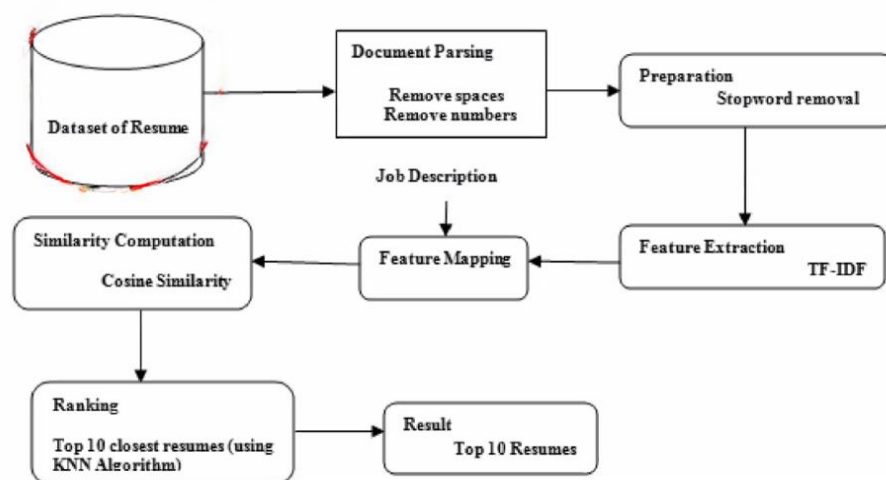7. **Handling Diverse Resume Formats:**

   - Many tools fail to parse non-standard resume layouts (e.g., infographics, non-linear career paths) effectively

CHAPTER 3
Proposed Methodology

3.1   System Design

System Architecture

3.2   Requirement Specification

Programming Languages:

- Python: The primary language for developing machine learning models, data processing, and implementing NLP techniques.

Machine Learning and Deep Learning Frameworks:

- TensorFlow: An open-source library for building and training machine learning models, particularly deep learning architectures.
- PyTorch: Another popular deep learning framework that is user-friendly and widely used for research and production.
- Scikit-learn: A library for traditional machine learning algorithms, useful for tasks like classification and regression.

Natural Language Processing (NLP) Libraries:

- NLTK (Natural Language Toolkit): A comprehensive library for working with human language data, useful for text processing and analysis.

4. Data Handling and Storage:

- Pandas: A powerful data manipulation library for handling structured data, particularly useful for preprocessing resumes and job descriptions.

- NumPy: A library for numerical computations, often used in conjunction with Pandas for data analysis.

3.2.1   Hardware Requirements:

1. **Development Workstations:**

- Processor (CPU):
  - Minimum: Intel Core i5 or AMD Ryzen 5
  - Recommended: Intel Core i7 or AMD Ryzen 7 for better performance during model training and data processing.

- **Memory (RAM):**
  - Minimum: 16 GB
  - Recommended: 32 GB or more, especially for handling large datasets and running multiple applications simultaneously.

- **Storage:**
  - Minimum: 512 GB SSD for faster data access and application performance.
  - Recommended: 1 TB SSD or larger, especially if working with large datasets of resumes and job descriptions.

- **Graphics Processing Unit (GPU):**
  - Minimum: NVIDIA GTX 1060 or equivalent for basic deep learning tasks.
  - Recommended: NVIDIA RTX 2060 or higher (e.g., RTX 3080) for more intensive model training and faster processing of neural networks.

2**. Server Requirements (for Deployment):**

- Processor (CPU):
  - Minimum: Multi-core processor (e.g., Intel Xeon or AMD EPYC) for handling concurrent requests.
  - Recommended: High-performance multi-core server CPUs for better scalability.

- Memory (RAM):
  - Minimum: 32 GB
  - Recommended: 64 GB or more, depending on the expected load and number of concurrent users.

- Storage:
  - Minimum: 1 TB SSD for fast read/write operations.
  - Recommended: 2 TB SSD or larger, with additional HDD storage for backups and archival data.

- Graphics Processing Unit (GPU):
  - Optional: Depending on the deployment needs, a dedicated GPU (e.g., NVIDIA Tesla or A100) may be beneficial for real-time processing and inference of deep learning models.

3. **Networking:**

- Network Interface Card (NIC):
  - Minimum: Gigabit Ethernet for reliable and fast network connectivity.
  - Recommended: 10 Gigabit Ethernet for high-throughput applications, especially if the system will handle large volumes of data or serve multiple users simultaneously.

- Firewall and Security Appliances:
  - Hardware firewalls to protect the server and data from unauthorized access and cyber threats.

4. Backup and Redundancy:

- External Backup Drives:
  - For regular backups of data and application states to prevent data loss.
- Uninterruptible Power Supply (UPS):
  - To ensure continuous operation during power outages and protect hardware from power surges.

3.2.2 Software Requirements:

1. Operating System:

- Windows, macOS, or Linux:
  - Recommended: Ubuntu or another Linux distribution for server deployment due to its stability, security, and compatibility with many development tools.

2. Development Tools:

- Integrated Development Environment (IDE):
  - PyCharm: For Python development, offering powerful features for code management and debugging.
  - Visual Studio Code: A lightweight, versatile code editor that supports multiple programming languages and extensions.
- Version Control System:
  - Git: For tracking changes in the codebase and facilitating collaboration among team members.
  - GitHub or GitLab: Platforms for hosting Git repositories, managing project documentation, and enabling collaboration.

3. Machine Learning and Deep Learning Frameworks:

- TensorFlow: An open-source library for building and training machine learning models, particularly deep learning architectures.

- PyTorch: A flexible deep learning framework that is widely used for research and production.
- Scikit-learn: A library for traditional machine learning algorithms, useful for tasks like classification and regression.

4. Natural Language Processing (NLP) Libraries:

- spaCy: An efficient library for NLP tasks, including tokenization, named entity recognition, and dependency parsing.
- NLTK (Natural Language Toolkit): A comprehensive library for working with human language data, useful for text processing and analysis.
- Transformers (by Hugging Face): A library that provides pre-trained models like BERT and GPT for advanced NLP tasks.

5. Data Handling and Storage:

- Pandas: A powerful data manipulation library for handling structured data, particularly useful for preprocessing resumes and job descriptions.
- NumPy: A library for numerical computations, often used in conjunction with Pandas for data analysis.
- SQL/NoSQL Databases:
    - PostgreSQL or MySQL: For structured data storage and management.
    - MongoDB: For handling unstructured data, such as resumes in various formats.

6. Security and Compliance:

- Encryption Libraries: Tools for securing sensitive data, such as user information and resumes, to comply with data protection regulations (e.g., GDPR).
- OAuth or JWT: For implementing secure authentication and authorization mechanisms.
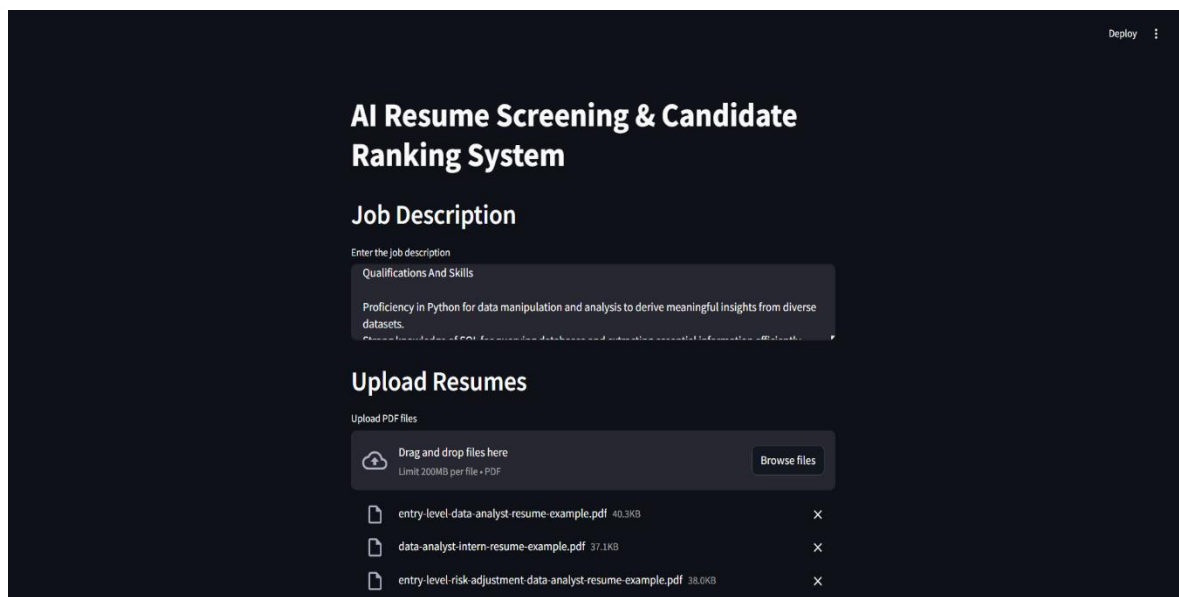
7. User Feedback and Improvement:

- Survey Tools (e.g., Google Forms, Typeform): For collecting user feedback on the system's performance and usability.
- A/B Testing Frameworks: Tools for testing different versions of the application to optimize user experience
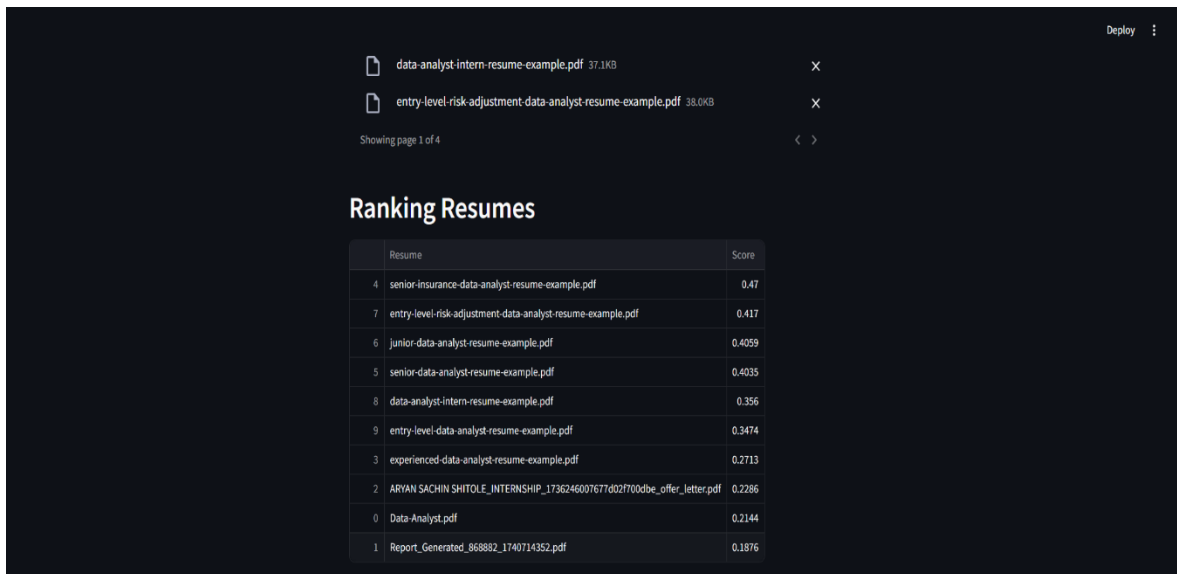
CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:

The app has a clean, dark-themed interface with these main components:

1. A title section "AI Resume Screening & Candidate Ranking System" at the top

2. A "Job Description" section where you can enter qualifications and skills (I can see you've added Python data manipulation skills)

3. An "Upload Resumes" section with a drag-and-drop interface for PDF files (with a 200MB limit per file)

The main "Ranking Resumes" table displays:

- A list of candidate resumes sorted by their match scores

- Each resume file name is shown with its corresponding score

- The scores range from 0.47 (highest) to 0.1876 (lowest on this page)

The top-ranked resumes are:

- senior-insurance-data-analyst-resume-example.pdf (0.47)

- entry-level-risk-adjustment-data-analyst-resume-example.pdf (0.417)

- junior-data-analyst-resume-example.pdf (0.4059)

4.2 GitHub Link for Code:

https://github.com/aryan0543f/AI-powered-Resume-Screening-and-Ranking-System-.git

CHAPTER 5

Discussion and Conclusion

5.4   Future Work:

Improving the AI-powered Resume Screening and Ranking System and addressing unresolved issues can significantly enhance its effectiveness and user satisfaction. Here are several suggestions for future work:

1. Enhanced Contextual Understanding:

Incorporate Advanced NLP Techniques:

Utilize more sophisticated models, such as transformer-based architectures (e.g., GPT-4, T5), to improve the system's ability to understand context, nuances, and industry-specific jargon in resumes and job descriptions.

3.   Bias Mitigation Strategies:

Implement Continuous Bias Audits:

Regularly evaluate the model for biases in its predictions and rankings. Use techniques like adversarial debiasing and fairness constraints to ensure equitable treatment of candidates from diverse backgrounds.

4.   Integration of Soft Skills Assessment:

Incorporate Soft Skills Evaluation:

Develop methods to assess soft skills and cultural fit, possibly through supplementary assessments (e.g., personality tests, situational judgment tests) or by analyzing language patterns in resumes and cover letters.

5.2 Conclusion:

The AI-Powered Resume Screening and Ranking System represents a significant advancement in the recruitment process, addressing the inefficiencies and biases inherent in traditional hiring methods. By leveraging cutting-edge technologies such as natural language processing, machine learning, and deep learning, the system enhances the accuracy and efficiency of candidate evaluation, ensuring that organizations can identify the best talent more effectively.

REFERENCES

[1]. Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, "Detecting Faces in Images: A Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume. 24, No. 1, 2002.

[2]. Gangoda, Nikethani, et al. "Resume Ranker: AI-Based Skill Analysis and Skill Matching System." *2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS)*. IEEE, 2024.