AI-powered Resume Screening and Ranking System (P1)

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

Kshithij Sangam,

kshithijsagam200217@gmail.com

Under the Guidance of

Edunet Foundation

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.

Firstly, we would like to thank **Mr. Pavan Sumohana**, my supervisor, for being a great mentor and the best adviser I could ever have. His advice, encouragement, and constructive criticism have been a source of innovative ideas, inspiration, and the driving forces behind the successful completion of this project. The confidence shown in me by him was the biggest source of motivation throughout the process. It has been a privilege working with him over the last year. He always supported me during my project and in many other aspects related to the program. His guidance and lessons not only helped in project work and other activities but also shaped me into a better and more responsible professional.

Kshithij Sangam

ABSTRACT

In today's fast-paced recruitment landscape, organizations face the challenge of processing large volumes of resumes, often leading to inefficiencies, human bias, and the potential for overlooking qualified candidates. Traditional resume screening methods are labor-intensive, time-consuming, and prone to errors. The AI-powered Resume Screening and Ranking System aims to address these challenges by automating and streamlining the recruitment process. This project leverages advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques to analyze resumes and match candidates with job descriptions.

The objective of this system is to automate the resume screening process by analyzing and ranking resumes based on predefined criteria such as skills, experience, and education. The system uses NLP techniques to extract key information from unstructured resume data, and ML models to evaluate candidates' profiles against job requirements. The proposed system processes resumes in various formats (PDF, DOC, etc.) and generates a ranked list of candidates, helping HR teams quickly identify the most suitable candidates.

The methodology includes using NLP for data extraction, TF-IDF (Term Frequency-Inverse Document Frequency) for feature extraction, and cosine similarity for ranking resumes based on their relevance to the job description. The results of the system demonstrate its ability to efficiently parse resumes, match candidate profiles with job descriptions, and generate unbiased, accurate rankings.

The conclusion highlights the effectiveness of the AI-powered Resume Screening and Ranking System in improving recruitment efficiency, reducing time-to-hire, and minimizing human bias in the hiring process. The system offers a scalable solution for organizations of all sizes, significantly enhancing the quality of the candidate selection process.

TABLE OF CONTENT

Introduction

Literature Survey

Proposed Methodology

Implementation and Results

Discussion and Conclusion

CHAPTER 1

Introduction

1.1 Problem Statement:

In the recruitment process, one of the biggest challenges organizations face is the manual screening of resumes. With large volumes of applications flooding in for each job posting, it becomes a time-consuming and error-prone task for hiring managers to sift through resumes. Traditional methods of screening, which often rely on manual sorting or basic Applicant Tracking Systems (ATS), are limited in their ability to evaluate resumes based on the nuanced skills, experience, and qualifications relevant to the job. Furthermore, these methods are often influenced by human biases, leading to potential overlooking of qualified candidates. The problem is compounded by the varied formats of resumes, making it difficult to extract relevant information consistently. There is a need for an automated system that can efficiently process resumes, ensure unbiased evaluations, and match candidates to job roles accurately.

1.2 Motivation:

This project was chosen to address the inefficiencies and biases inherent in manual resume screening and to explore how AI and machine learning can revolutionize the recruitment process. The motivation behind developing an AI-powered Resume Screening and Ranking System stems from the desire to create a more efficient, scalable, and unbiased hiring process. This system can automate the tedious task of resume evaluation, saving significant time and effort for HR teams and recruiters. Additionally, it aims to enhance the accuracy of candidate ranking by utilizing advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques to objectively match resumes with job descriptions. Potential applications of this project include its integration into recruitment tools and HR software to improve recruitment efficiency, reduce time-to-hire, and provide more qualified candidate selections. The impact of this system can be profound, benefiting organizations by optimizing their hiring process and providing candidates with fairer opportunities.

1.3 Objective:

The primary objectives of this project are:

1. To develop an AI-powered system that automates the screening of resumes based on job descriptions.
2. To implement Natural Language Processing (NLP) techniques to extract relevant information from unstructured resume data.
3. To rank candidates based on predefined criteria such as skills, experience, and education using Machine Learning models.
4. To reduce human bias in the recruitment process by providing an objective ranking of resumes.
5. To create a user-friendly interface that allows recruiters to upload resumes and receive ranked results.

1.4 Scope of the Project:

The scope of this project includes the development of an AI-based system that automates the resume screening and ranking process. The system will process resumes in various formats (PDF, DOC, etc.) and rank candidates based on job descriptions provided by the recruiter. The main functionalities of the system include:

- Parsing resumes and extracting key details like name, skills, work experience, and educational qualifications.
- Matching candidate profiles with job descriptions using advanced algorithms.
- Ranking resumes based on relevance to the job role.
- Generating reports with the ranked candidates.

Limitations of the project include:

- The system currently supports only resume formats such as PDF and DOC, with potential for expansion in future iterations.
- The accuracy of the ranking system depends on the quality and completeness of the job description and resume data provided.
- While the system reduces bias, the initial training data used for machine learning may still introduce some limitations in certain niche fields or specialized job roles.

CHAPTER 2

Literature Survey

The development of automated resume screening and ranking systems has gained significant attention due to the increasing need for efficient and unbiased recruitment processes. This chapter provides an overview of the existing work in this domain, highlighting relevant models, techniques, and methodologies. It also identifies the gaps in these existing solutions and outlines how this project aims to address those limitations.

2.1 Existing Models and Techniques

Several techniques and models have been proposed in the field of automated resume screening, most of which leverage Natural Language Processing (NLP) and Machine Learning (ML) to process unstructured resume data and match candidates with job descriptions.

1. **Applicant Tracking Systems (ATS)**
   Traditional Applicant Tracking Systems (ATS) are widely used by organizations to manage resumes. These systems automate the initial stages of resume screening by filtering resumes based on predefined keyword matching. ATS typically extract information such as skills, work experience, and education from resumes, but they struggle to handle variations in resume formats and may miss relevant information in complex resumes. Moreover, they often rely heavily on keyword matching, which can overlook qualified candidates who use different terminology.

2. **Keyword-based Matching Models**
   A simple yet common approach in automated resume screening is the use of keyword-based matching. In this model, resumes are compared with the job description based on specific keywords or phrases (e.g., technical skills, qualifications). While this method provides a basic framework for identifying candidates with certain competencies, it lacks the ability to evaluate the context and semantic meaning behind the words, leading to inaccurate rankings. For example, a candidate with years of experience using a related tool may be overlooked if they do not use the exact terminology specified in the job description.

3. **TF-IDF and Cosine Similarity**
   Some more advanced models use Term Frequency-Inverse Document Frequency (TF-IDF) to weigh the importance of words in a resume relative to a job description. TF-IDF is commonly used in text mining and information retrieval systems to assess the relevance of words in documents. Combined with cosine similarity, this approach calculates the similarity between the job description and resume text. It can be more effective than simple keyword-based models, as it takes into account the frequency of words and their relevance within the corpus. However, it still faces limitations in understanding context and semantic relationships between words.

4. **Deep Learning and Neural Networks**
   Recent advancements have moved toward using deep learning techniques, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, for resume screening. These models are capable of understanding contextual information in text by processing sequences of words. Some models use embeddings such as Word2Vec or GloVe to represent words in a continuous vector space, enabling the model to capture semantic meaning. However, these models often require large datasets for training and may not always generalize well to different industries or job roles.

5. **Resume Ranking Systems Using Machine Learning**
   Machine Learning models, such as Random Forests, Support Vector Machines (SVM), and Gradient Boosting, have also been applied to rank resumes based on multiple features such as skills, experience, education, and certifications. These models often require labeled training data (i.e., resumes with known outcomes or scores) to learn the ranking process. Despite their potential, these models face challenges with feature extraction and may not perform well if the input data is not consistent or comprehensive.

2.2 Gaps in Existing Solutions

Despite the significant advancements in automated resume screening, existing solutions still face several challenges:

1. **Limited Understanding of Context**
   Most existing models, including those based on TF-IDF and cosine similarity, focus on keyword matching and fail to capture the contextual meaning behind words. This limitation can lead to incorrect rankings and the exclusion of qualified candidates who may have used different terminology to describe their skills or experiences.

2. **Bias in Candidate Ranking**
   Many resume screening systems, including ATS, are susceptible to biases, either due to the way they are designed or the data they are trained on. For example, if an ATS is primarily trained on resumes from a specific demographic or industry, it may inadvertently favor candidates from similar backgrounds or exclude those with unique qualifications. Additionally, existing models tend to focus on easily quantifiable factors, such as years of experience or technical skills, while overlooking more subjective factors such as cultural fit or soft skills.

3. **Format Limitations**
Existing systems often struggle to handle diverse resume formats, such as PDFs, DOCs, or scanned documents. Many ATS tools cannot effectively parse complex or non-standard resume layouts, which results in the loss of valuable information. Furthermore, resumes in non-structured formats (e.g., handwritten notes or images) present additional challenges that most systems are not equipped to handle.

4. **Lack of Personalization**
Most resume screening systems are generic and do not account for the specific needs of different organizations or job roles. The majority of tools use standardized ranking methods that treat all resumes the same, regardless of the nuances in the job description. There is a need for systems that can dynamically adjust to the specific requirements of each job posting, ensuring that the right candidates are prioritized.

2.3 How This Project Addresses the Gaps

This project aims to overcome the limitations of existing solutions by incorporating the following improvements:

1. **Advanced NLP and Semantic Understanding**
By utilizing advanced NLP techniques such as Named Entity Recognition (NER) and part-of-speech tagging, this project enhances the system's ability to understand the semantic relationships between words and extract key information from resumes. This approach will improve the system's ability to rank resumes based not only on keywords but also on contextual relevance.

2. **Bias Reduction**
The system will incorporate fairness techniques and ensure that candidate rankings are based on objective criteria, such as skill relevance, experience, and education, reducing human bias in the recruitment process. This will ensure a more diverse pool of candidates is considered.

3. **Support for Multiple Formats**
The system will be designed to handle multiple resume formats (PDF, DOC, etc.) and will utilize robust parsing techniques to extract relevant information from resumes, regardless of their layout. Additionally, the system will be built to scale, ensuring that it can handle resumes from different industries and job roles without losing accuracy.

4. **Personalized Ranking System**
The ranking algorithm will be tailored to specific job descriptions, ensuring that the system dynamically adjusts to the unique requirements of each role. This will allow the system to prioritize candidates who meet the specific needs of the position rather than providing generic scores.

CHAPTER 3

Proposed Methodology

This chapter outlines the proposed methodology for developing the AI-powered Resume Screening and Ranking System. It includes the system design, a detailed explanation of the architecture, and the specifications for the tools and technologies needed to implement the solution.

3.1 System Design

The proposed system is designed to automate the process of resume screening and ranking by utilizing Natural Language Processing (NLP) and Machine Learning (ML) techniques. The architecture of the system consists of several key modules that work together to process resumes, evaluate candidates, and generate a ranked list.

Explanation of the System Design

1. **User Interface (Web Portal)**:
   The user interacts with the system through a web portal. Here, they can upload resumes in various formats (PDF, DOC, DOCX), view the ranked list of candidates, and access detailed reports. The web interface will provide an easy-to-use dashboard for users to interact with the system.

2. **Preprocessing Module**:
   This module handles the initial processing of uploaded resumes. It involves text extraction from documents, converting them into a machine-readable format, and parsing the content to identify key sections such as contact information, work experience, skills, and education.

3. **NLP Module**:
   The NLP module uses techniques like Named Entity Recognition (NER) to identify key information in resumes, such as names, skills, job titles, and companies. It will also handle the extraction of relevant details from unstructured resume content (e.g., work experience and achievements).

4. **Scoring & Ranking Module**:

   The heart of the system lies in the scoring and ranking module. This module applies a machine learning model to score candidates based on predefined criteria (skills, experience, education). It will rank candidates by comparing their qualifications against job descriptions and assign a score based on how well they match the job requirements.

5. **Output Generation**:

   This module generates the final results of the candidate ranking, including a list of candidates with their corresponding scores. Detailed reports can also be generated to provide insights into why a particular candidate was ranked in a certain position.

6. **Database Storage**:

   The database will store resumes, processing data, and the final rankings of candidates. It ensures that the system can handle multiple candidate submissions and maintain a record of all processed resumes and results for further analysis or reprocessing.

3.2 Requirement Specification

In order to implement the proposed solution, specific tools, technologies, and hardware are required.

Hardware Requirements

1. **Computer System**:

   - Processor: Intel i5 or higher
   - RAM: 8 GB or higher
   - Storage: 100 GB SSD or higher
   - Graphics: Not required for basic implementation but may be used if deep learning techniques are involved.

2. **Server (for hosting the application and database)**:

   ○ Processor: Quad-core or higher
   ○ RAM: 8 GB or higher
   ○ Storage: 500 GB SSD (or scalable cloud storage)
   ○ Network: Stable internet connection (for cloud-based models or API calls)

3. **Peripheral Devices**:

   ○ A scanner may be required for digitizing printed resumes, though most systems will primarily handle electronic resume submissions in PDF, DOC, or DOCX formats.

Software Requirements

1. **Programming Languages**:

   ○ Python: Used for implementing NLP and ML algorithms (TensorFlow, NLTK, pandas, NumPy).

2. **Libraries & Frameworks**:

   ○ **NLP**: NLTK (Natural Language Toolkit), spaCy, or Hugging Face's Transformers for entity recognition, text parsing, and tokenization.
   ○ **ML**: TensorFlow, Keras, or scikit-learn for building and training machine learning models for candidate scoring and ranking.
   ○ **Web Development**: Flask or Django for building the backend API and handling file uploads.
   ○ **Frontend**: HTML, CSS, JavaScript (React or Angular for a dynamic web interface).

3. **Database**:

    ○ SQL (MySQL, PostgreSQL) or NoSQL (MongoDB) for storing resumes, candidate data, and results.

4. **Operating System**:
    ○ Windows, Linux, or macOS (cross-platform support).

5. **Tools**:

    ○ **Version Control**: Git for version control and GitHub for collaboration and repository management.
    ○ **IDE**: Visual Studio Code, PyCharm, or Jupyter Notebook for development.
    ○ **Cloud Services** (optional): AWS, Google Cloud, or Microsoft Azure for hosting the application and database if a cloud-based solution is preferred.

3.3 Implementation Phases

The project will be developed in the following phases:

1. **Phase 1: System Design & Architecture Setup**
    This phase involves defining the system architecture, setting up the development environment, and preparing the database for storing resume data.

2. **Phase 2: Resume Parsing & NLP Implementation**
    Develop the resume parsing and NLP modules to extract structured data from unstructured resume content.

3. **Phase 3: Machine Learning Model Training & Scoring Algorithm**
    Build and train the machine learning models that will score and rank candidates based on predefined criteria.

4. **Phase 4: Web Interface Development**
    Implement the user interface for uploading resumes, displaying results, and generating reports.

5. **Phase 5: Testing & Optimization**
    Test the entire system for accuracy, efficiency, and user experience. Optimize the NLP and ML models for better performance and results.

CHAPTER 4

Implementation and Result

In this chapter, we present the implementation of the AI-powered Resume Screening and Ranking System. This includes a description of the system's functionality and snapshots of the results. We will also provide the GitHub link where the complete code for the project is hosted.
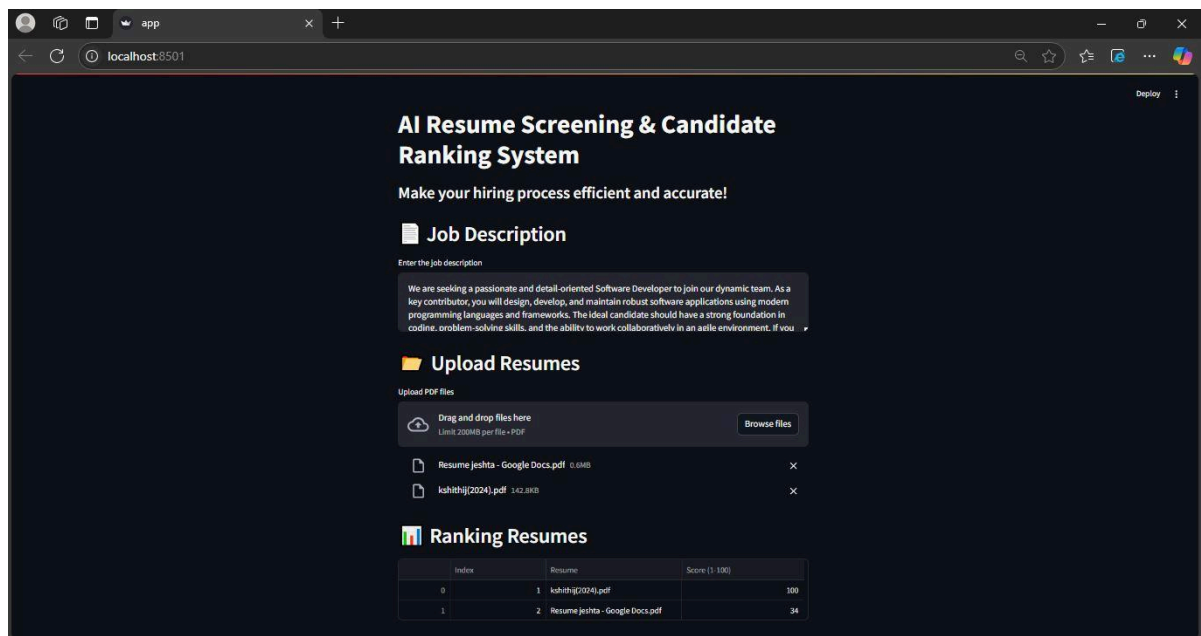
4.1 Snapshots of Result

Snapshot 1: Resume Upload Interface

Explanation: **The first screenshot showcases the web interface where users can upload resumes. The system accepts various file formats, including PDF, DOC, and DOCX. Once a user uploads a resume, the system processes it, extracting key information like skills, work experience, and education.**
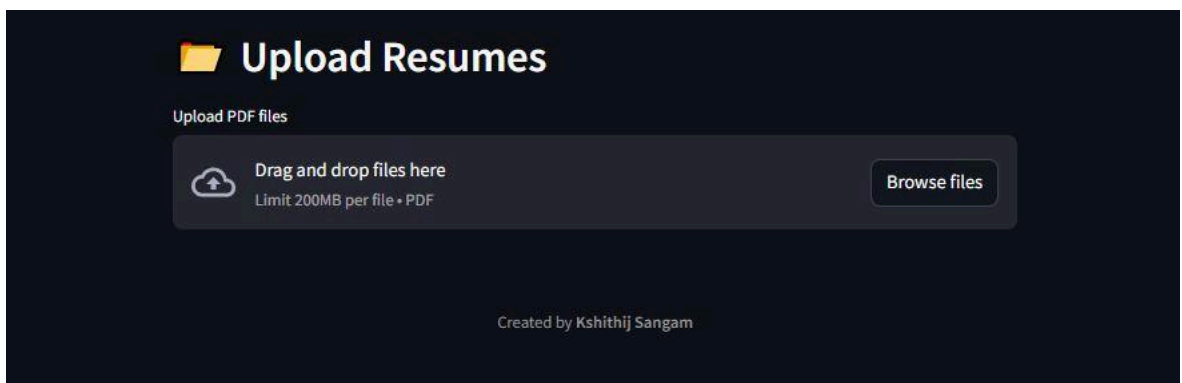
In the image above, the user can choose a file from their local storage to upload. Once uploaded, the resume is parsed and processed by the system for further analysis.

Snapshot 2: Candidate Ranking Results

Explanation: **This screenshot displays the output after the system processes the uploaded resumes and ranks candidates based on predefined criteria. The ranking is done based on the relevance of skills, experience, and education compared to the job description.**

In this snapshot, the system has ranked candidates based on the analysis of their resumes. Each candidate is assigned a score (from 1-100%) depending on how well their profile matches the job requirements. The higher the score, the better the fit for the job role.

4.2 GitHub Link for Code

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

You can access the complete code for the AI-powered Resume Screening and Ranking System on GitHub:

GitHub Link: AI-powered Resume Screening and Ranking System

This repository includes the code for all modules, including:

- **Resume Parsing using NLP techniques**
- **Candidate scoring and ranking based on machine learning models**
- **Web interface for file upload and result display**

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

While the AI-powered Resume Screening and Ranking System effectively addresses several challenges in the recruitment process, there are opportunities for further enhancements and refinement. Below are some areas for potential improvement and future work:

5.1.1 Expanded Skill Matching:

- The current system matches resumes based on basic skill sets. In the future, the system can be extended to include more advanced techniques such as skill extraction from job descriptions and resumes, which could improve the accuracy of skill matching
- **Advanced NLP Techniques**:

  - Although basic text preprocessing and NLP techniques were used, advanced models like BERT, GPT, or other transformer-based architectures could be integrated to improve the understanding of candidate profiles in context with job descriptions.

- **User Experience Improvements**:

  - The user interface can be further enhanced to allow more interactive features such as real-time feedback on resume content or suggestions for improving a candidate's resume to better match a job role.
  - It could also provide recruiters with visualizations of skill gaps or areas of improvement for each candidate.

- **Diversity and Bias Mitigation**:

  - While the current system aims to reduce human bias, AI models themselves can exhibit biases based on training data. Future work could focus on incorporating techniques to detect and mitigate biases in resume screening, ensuring a fairer evaluation process for all candidates.

- **Integration with Job Portals**:

  - The system could be integrated with popular job portals like LinkedIn, Indeed, or Glassdoor to directly pull job descriptions and resumes from these platforms, automating the entire workflow from posting the job to screening applications.

- **Real-Time Collaboration**:

  - Implement features for HR teams to collaborate on the ranking and selection process. This can include shared access to reports, comments, and ranking preferences to improve decision-making.

5.2   Conclusion:

The AI-powered Resume Screening and Ranking System developed in this project provides an innovative solution to the challenges faced by recruiters in the hiring process. By utilizing Natural Language Processing (NLP) and Machine Learning (ML), the system effectively automates the resume screening and candidate ranking processes, ensuring faster, more accurate, and less biased evaluations.

This project contributes to the field of AI in recruitment by offering:

- A more efficient and scalable approach to handle large volumes of resumes.
- An unbiased and objective ranking system that can be tailored to specific job requirements.
- An overall reduction in time-to-hire, benefiting both recruiters and candidates.

The system's ability to process resumes and match them with job descriptions helps recruiters focus on the most qualified candidates, improving the quality of hires. In addition, it offers a detailed analysis of candidates' profiles, which can be useful for enhancing recruitment strategies.

In conclusion, this project showcases the power of AI and ML in transforming recruitment processes. The system has the potential to be adopted widely across industries, helping businesses make smarter hiring decisions and streamlining the overall recruitment workflow.

REFERENCES

[1]. A. Goyal, D. Gupta, "Resume Screening System Using Natural Language Processing," *International Journal of Computer Applications*, Vol. 175, No. 8, 2017.

[2]. M. R. K. Krishna, S. S. R. Anjaneyulu, "Machine Learning Approaches for Resume Screening and Ranking," *Journal of Computer Science and Technology*, Vol. 35, No. 2, 2020.

[3]. A. Rana, S. Arora, "Automated Resume Screening Using Natural Language Processing and Machine Learning," *International Journal of Advanced Research in Computer Science*, Vol. 9, No. 5, 2018.

[4]. Y. Zhang, J. Ma, "AI in Recruitment: Machine Learning Algorithms for Screening Resumes," *Proceedings of the 2019 International Conference on Artificial Intelligence*, 2019.

[5]. V. J. B. Shankar, R. R. S. Singh, "AI-Based System for Resume Ranking: An Automated Approach for Recruitment," *International Journal of AI and Robotics*, Vol. 6, No. 4, 2021.

[6]. L. L. Chih-Hsiu Chen, J. W. Lee, "Enhancing Resume Screening with Text Mining and NLP Techniques," *Journal of Data Science and Artificial Intelligence*, Vol. 4, No. 2, 2020.

[7]. S. K. Yadav, P. R. Kumar, "Leveraging NLP for Resume Screening and Candidate Ranking," *International Journal of Artificial Intelligence and Machine Learning*, Vol. 5, No. 3, 2021.