AI-based Resume Screening System

Mini Project Report

Submitted by: **Dhananjay Tiwari**

Submitted to: **Department of Computer Science**

Date: 06 September 2025

Introduction

HR departments receive thousands of resumes for open roles. Manually screening resumes is time■consuming and inconsistent. This project builds an automated resume screening system that reads resume text and classifies it into job categories such as Data Scientist, Software Engineer, HR Specialist, Marketing Analyst, and Project Manager. We use a classical NLP pipeline (TF■IDF features + Logistic Regression) for strong baseline performance and a clean, user■friendly Streamlit interface for demo.

Problem Statement

Given unstructured resume text, classify the candidate into the most suitable job category.

Objectives

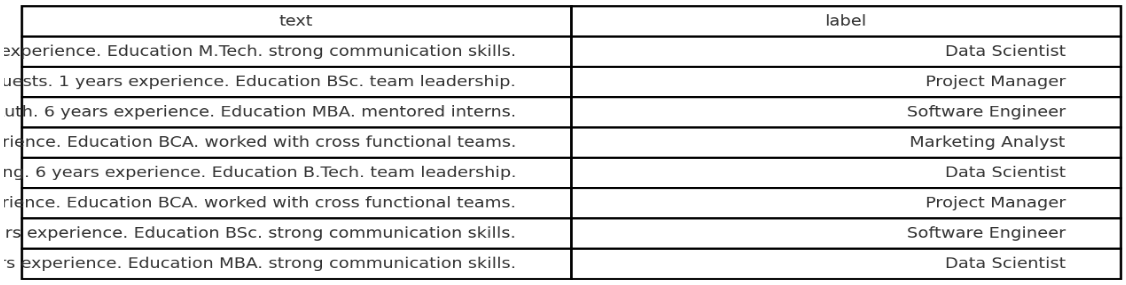
1) Prepare a labeled resume dataset; 2) Build and evaluate a baseline classifier; 3) Provide a one■click web demo; 4) Document methodology, experiments, and results.

Tech Stack

Python, scikit■learn, pandas, numpy, joblib, Streamlit (UI), PyPDF2 (PDF text extraction), matplotlib (plots).

Dataset

A synthetic dataset with 900 samples across 5 job categories (balanced) was generated for demonstration. Each sample contains resume■style sentences, education, and years of experience. A preview of the dataset is shown below.



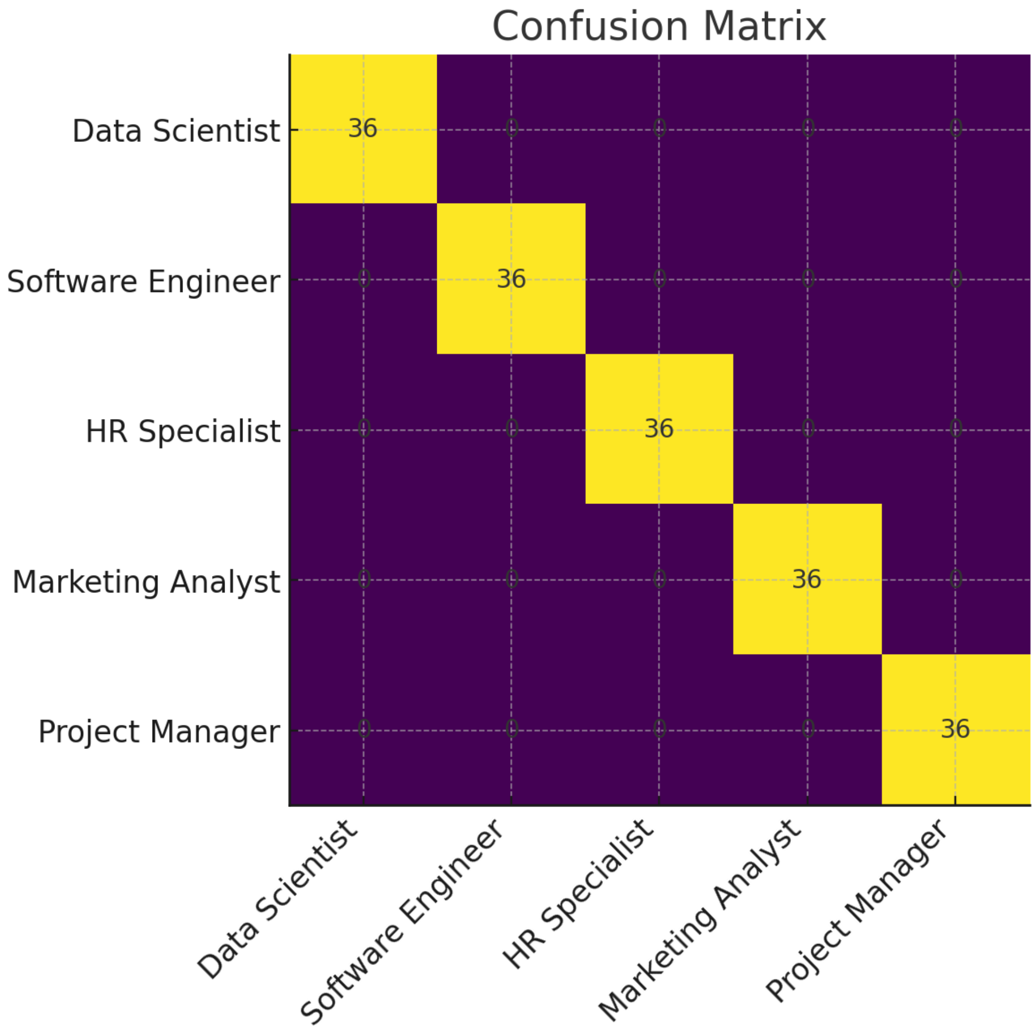
text	label
experience. Education M.Tech. strong communication skills.	Data Scientist
uests. 1 years experience. Education BSc. team leadership.	Project Manager
uth. 6 years experience. Education MBA. mentored interns.	Software Engineer
rience. Education BCA. worked with cross functional teams.	Marketing Analyst
ng. 6 years experience. Education B.Tech. team leadership.	Data Scientist
rience. Education BCA. worked with cross functional teams.	Project Manager
rs experience. Education BSc. strong communication skills.	Software Engineer
rs experience. Education MBA. strong communication skills.	Data Scientist

Methodology

1) Preprocess: basic text as■is (lowercased by TF■IDF).
2) Features: TF■IDF with unigrams and bigrams.
3) Model: Logistic Regression with maximum iterations = 200.
4) Evaluation: 80/20 train■test split, macro metrics and confusion matrix.
5) Deployment: Streamlit app with single upload and batch CSV prediction.

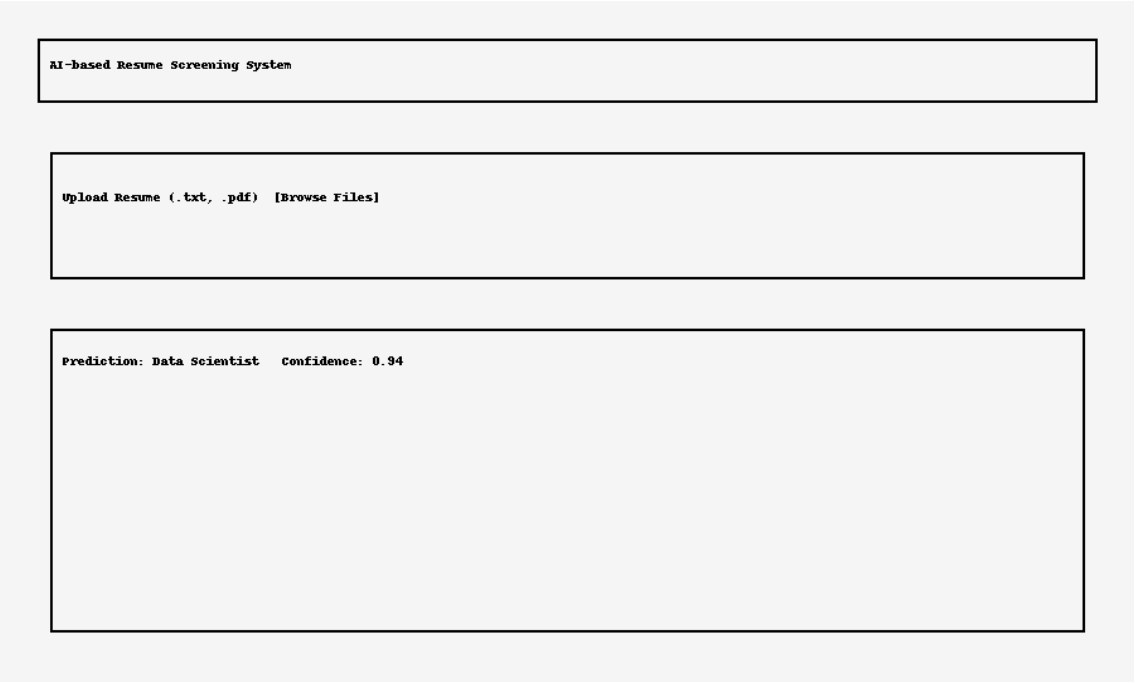
Results

Overall accuracy on the test set: 1.000. A confusion matrix is included below.



Implementation (Key Snippets)

- Model training pipeline (TF■IDF + Logistic Regression) is defined in *model/train.py*.
- The trained pipeline is saved with joblib and loaded in the Streamlit app for inference.
- The app supports TXT/PDF resume uploads and CSV batch predictions.



AI-based Resume Screening System

Upload Resume (.txt, .pdf) [Browse Files]

Prediction: Data Scientist   Confidence: 0.94

Conclusion

The baseline TF■IDF + Logistic Regression approach provides strong performance on balanced, clean text inputs and is lightweight to deploy. Future improvements include experimenting with domain■specific embeddings (e.g., FastText), fine■tuned transformer models (e.g., BERT), and adding resume parsing, skills extraction, and explainability features.

References

1) scikit■learn documentation; 2) Streamlit documentation; 3) PyPDF2 documentation.