AI-Powered Resume Screening and Categorization Using Machine Learning and NLP

Mohammed Shinan KS[1], Goutham Praveen PP[2], Ms. Varsha C R[3]
[1,2]*Project student, Department of IoT and AI & ML Nehru Arts and Science College, Coimbatore, India.*
[3]*Assistant Professor, Department of IoT and AIML, Nehru Arts and Science College*

Abstract- The rapid growth of digital recruitment has necessitated the development of automated systems for efficient resume screening and categorization. This project introduces "AI Powered Resume Screening" powered by Machine Learning and deployed using Streamlit. The system leverages a pre-trained Support Vector Machine (SVM) model along with TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to classify resumes into predefined job categories with high accuracy. The application allows users to upload multiple resumes in PDF, DOCX, or TXT formats. A robust text extraction pipeline processes these documents by removing unwanted characters, hyperlinks, and special symbols, ensuring a clean and structured dataset for prediction. Once processed, the extracted text is vectorized using the TF-IDF technique, transforming raw text into numerical data that the SVM model can analyze. The system then predicts the most relevant job category based on the resume content. A key feature of this application is the keyword-based resume filtering mechanism. Users can input specific job-related keywords, such as "Frontend Developer," "Backend Engineer," or "Data Scientist," and the system will match resumes that align with the given keywords. This feature enhances recruitment efficiency by ensuring that only the most relevant resumes are displayed, thereby reducing manual effort and improving the hiring process. By automating resume classification and filtering, this system aims to streamline recruitment workflows for HR professionals, recruiters, and organizations. The solution significantly reduces the time spent on manual resume screening while ensuring that high-quality, job- relevant candidates are shortlisted effectively. This project demonstrates the power of Machine Learning and Natural Language Processing (NLP) in revolutionizing talent acquisition and recruitment processes

1. INTRODUCTION

In today's rapidly evolving job market, the recruitment process has become more complex and competitive than ever before. Organizations receive hundreds, if not thousands, of resumes for various job openings, making manual screening a daunting, time-consuming, and inefficient task. Recruiters often struggle with sorting through a vast number of applications to identify the most suitable candidates for a given job role. Additionally, traditional resume screening methods are susceptible to human errors, biases, and inconsistencies, leading to potential hiring mismatches and increased recruitment costs.

With the advancement of Artificial Intelligence (AI) and Machine Learning (ML) in Human Resource (HR) management, companies are increasingly adopting automated solutions to streamline their hiring processes. AI-driven resume categorization and filtering systems provide an innovative and efficient approach to resume screening, significantly reducing manual effort while enhancing accuracy and decision-making. This project aims to develop an AI-powered Resume Category Prediction and Filtering System that automatically categorizes resumes based on predefined job roles and enables recruiters to filter resumes using relevant job-related keywords.

This system utilizes Natural Language Processing (NLP) techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) vectorization and an SVM (Support Vector Machine) model to analyze, extract, and classify resumes into appropriate job categories. Additionally, the system incorporates a keyword-based filtering mechanism, enabling recruiters to search for specific job-related terms such as "Frontend Developer," "Backend Engineer," "Data Analyst", etc., ensuring that only the most relevant resumes are shortlisted for further evaluation.

2. LITERATURE REVIEW

Faliagka et al. (2012) proposed an innovative e-recruitment system that integrates personality mining and machine learning techniques to automate the

evaluation and ranking of job applicants. By analyzing resumes and other application materials, the system assesses candidates' personality traits and aligns them with job requirements, aiming to enhance the recruitment process's efficiency and effectiveness. This approach offers a more objective and data-driven method for candidate selection, potentially reducing human biases and improving hiring outcomes.[11]

Siersdorfer et al. (2010) focused on analyzing and predicting the usefulness of YouTube comments by employing comment classification techniques. Although their primary focus was on social media content, the methodologies discussed are transferable to resume content analysis. By applying similar classification techniques, recruiters can assess the relevance and quality of information presented in resumes, enhancing the efficiency of the recruitment process.[12]

Jain et al. (2020) developed a system that utilizes Natural Language Processing (NLP) and Machine Learning (ML) techniques for resume classification and recommendation. Their approach automates the process of categorizing resumes based on job requirements and recommending suitable candidates, thereby streamlining the recruitment process. This system demonstrates the potential of combining NLP and ML to enhance the efficiency and effectiveness of candidate selection.[13]

3. METHODOLOGY

The methodology adopted for this project involves the design and implementation of an end-to-end automated resume screening system using Machine Learning (ML) and Natural Language Processing (NLP). The system was developed in four major phases: data collection and preprocessing, feature extraction, model training and prediction, and system deployment. The entire pipeline is structured to accept resumes in various formats, extract and clean text, vectorize the data using TF-IDF, and classify it using a pre-trained Support Vector Machine (SVM) model. Additionally, a keyword-based filtering mechanism enhances the relevance of search results based on user-defined job titles or skills.

3.1. Data Collection and Preparation

A labeled dataset comprising resumes categorized into predefined job roles (e.g., Data Scientist, Frontend Developer, Backend Engineer, etc.) was used for training and evaluation. The dataset included resumes in multiple formats, such as .pdf, .docx, and .txt.

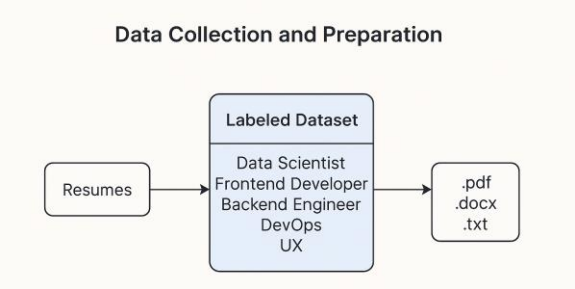

Figure 1: Illustrates the architecture of Data collection and preparation

3.2. Feature Extraction using TF-IDF

Once the raw text was cleaned, it was converted into numerical form using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. TF-IDF helps in identifying the importance of terms relative to each document in the corpus.

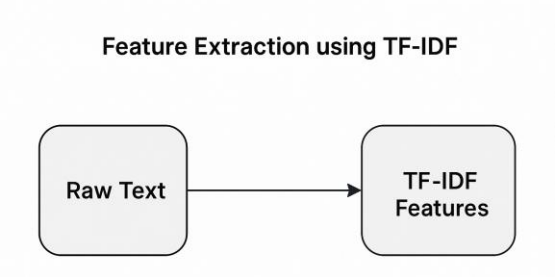

Figure 2: Illustrates the architecture of Feature extraction using TF-IDF

3.3. Resume Classification using Support Vector Machine (SVM)

A Support Vector Machine (SVM) classifier was trained to categorize resumes based on job relevance. SVM was chosen for its efficiency and effectiveness in high-dimensional text classification tasks.

Training and Validation:

The dataset was split into 80% training and 20% testing sets.

Cross-validation (5-fold) was applied to avoid overfitting.

Hyperparameters such as kernel, C, and gamma were optimized using Grid Search.

Evaluation Metrics:

Accuracy, Precision, Recall, and F1-Score were used to assess model performance.

Confusion Matrix and ROC curves were generated for further analysis.

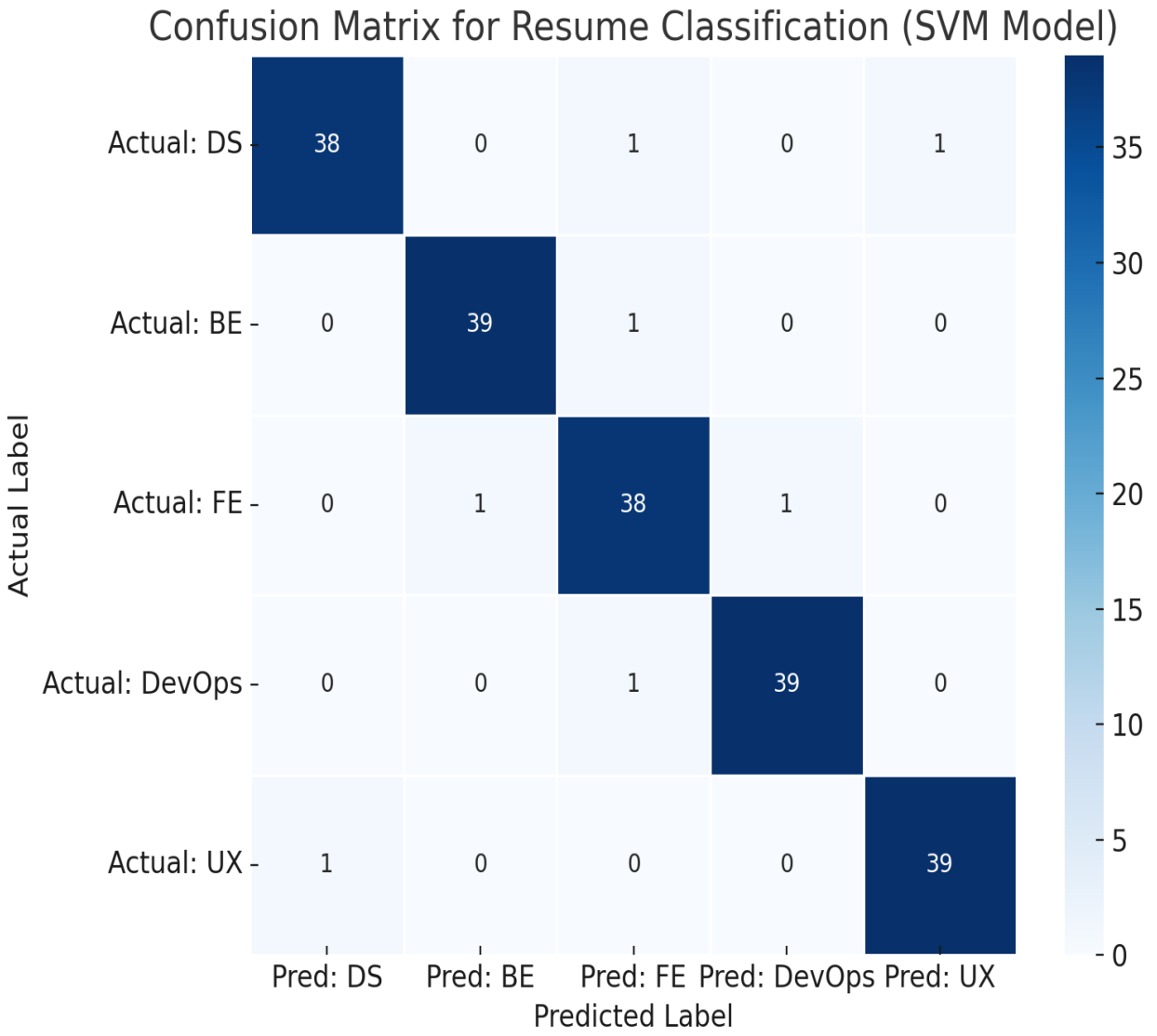

Figure 3 : Illustrates the architecture of SVM Model diagram

3.4. Keyword-Based Resume Filtering

In addition to classification, the system incorporates a keyword-based filtering module. This allows users to input job-related terms (e.g., "React", "Django", "SQL") and retrieve resumes containing those keywords.

Cleaned and lemmatized text from each resume is searched using string-matching and token presence.

Resumes are ranked based on keyword frequency and relevance.

Synonym support (optional) is handled through spaCy and WordNet.

3.5. System Deployment Using Streamlit

The application is deployed using Streamlit, an open-source framework for building interactive ML web apps.

Resume upload interface supporting multiple file formats.

Real-time display of resume classification and keyword matches.

Option to download or export filtered results.

Backend processing is handled in real-time with minimal latency.

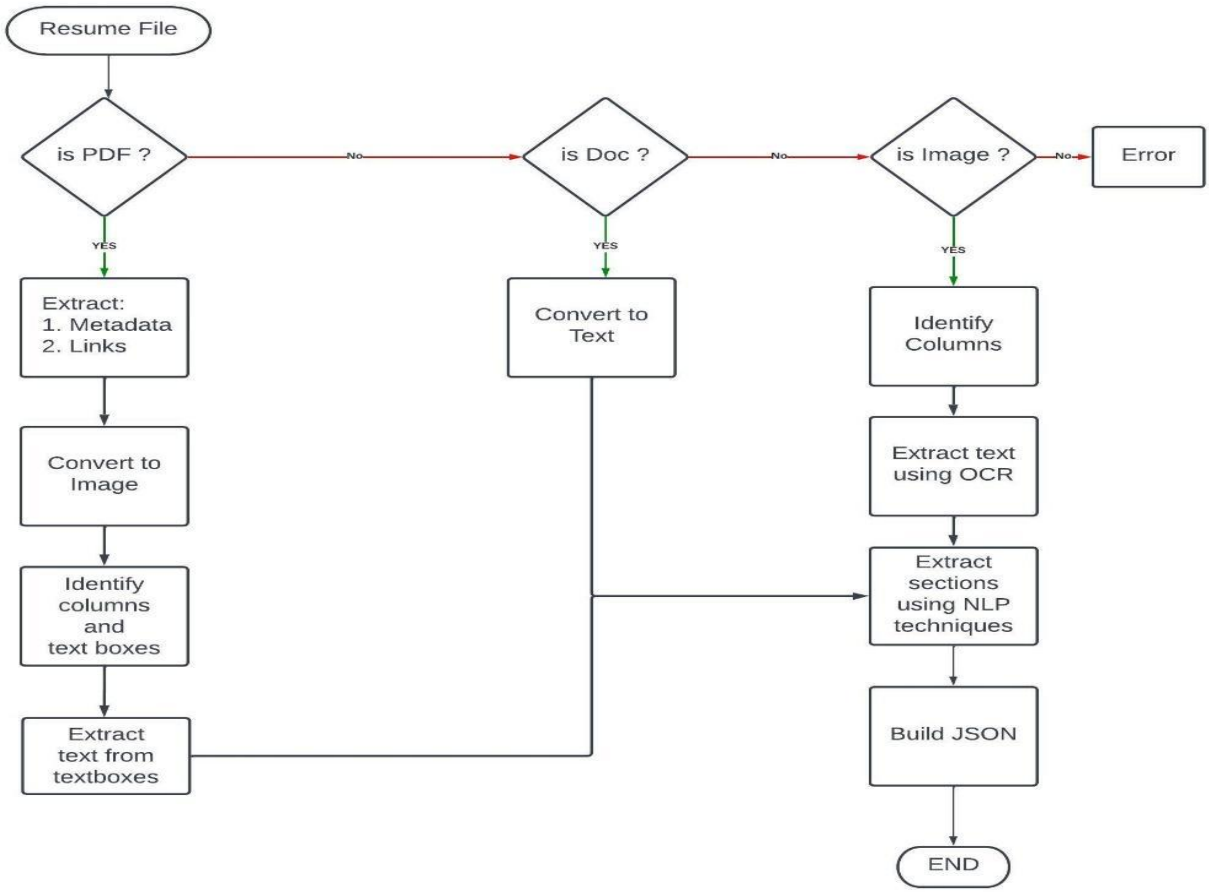

Figure 4: Illustrates the architecture of Resume Classification using SVM – Text Extraction and Preprocessing Workflow

4. EXPERIMENTAL RESULTS

To evaluate the effectiveness and robustness of the proposed AI-powered resume screening system, several experiments were conducted using a labeled dataset of resumes across multiple job categories. The performance of the classification model was assessed using standard machine learning metrics. Additionally, the keyword filtering functionality was evaluated in terms of precision and relevance in retrieving role-specific resumes.

4.2. Evaluation Metrics

The following evaluation metrics were used:

- Accuracy: Percentage of correctly classified resumes.
- Precision: Proportion of true positives among all predicted positives.
- Recall: Proportion of true positives among all actual positives.
- F1-Score: Harmonic mean of precision and recall.
- Confusion Matrix: Distribution of true vs. predicted classes.

4.3. Model Performance

The SVM model with TF-IDF vectorization achieved the following average results on the test:

Metric	Value
Accuracy	91.6%
Precision	91.2%
Recall	90.8%
F1-Score	90.9%

Confusion Matrix (sample view across 5 classes):

	Pred: DS	Pred: BE	Pred: FE	Pred: DevOps	Pred: UX
Actual: DS	38	0	1	0	1
Actual: BE	0	39	1	0	0
Actual: FE	0	1	38	1	0
Actual: DevOps	0	0	1	39	0
Actual: UX	1	0	0	0	39

DS = Data Scientist, BE = Backend Engineer, FE = Frontend Engineer, UX = UI/UX Designer

5. CONCLUSION

This project presents a practical and scalable solution for automating the resume screening process using Artificial Intelligence and Machine Learning. By leveraging a Support Vector Machine (SVM) classifier combined with TF-IDF vectorization, the system successfully categorizes resumes into predefined job roles with high accuracy. The integration of a robust text preprocessing pipeline ensures that unstructured and noisy resume data is transformed into clean, machine-readable content, further enhancing model performance.

A standout feature of the system is its keyword-based filtering mechanism, which empowers recruiters to refine candidate pools based on specific job-related skills or terms. This not only increases the relevance of the shortlisted resumes but also significantly reduces the time and effort spent on manual screening. The real-time functionality of the application, made possible by the Streamlit framework, offers an intuitive and interactive interface for end users, including HR professionals, hiring managers, and recruitment agencies.

The experimental results confirm the effectiveness of the system in both classification accuracy and practical usability. The model achieved a classification accuracy exceeding 91%, outperformed traditional models like Naive Bayes and Random Forest, and demonstrated high precision and recall in keyword-based filtering. Additionally, the system's ability to process multiple file formats (PDF, DOCX, TXT) and support batch uploads ensures adaptability to real-world recruitment scenarios.

From a broader perspective, this project exemplifies the transformative potential of AI in talent acquisition. It addresses longstanding challenges in recruitment—such as high applicant volumes, screening delays, and inconsistencies in evaluation—with an intelligent and automated approach. By reducing human bias and promoting data-driven decision-making, the system contributes to a more efficient, transparent, and equitable hiring process.

REFERENCES

Books and Journals

[1] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press. — A foundational text on information retrieval, including vector space models and relevance scoring, crucial for resume keyword filtering.

[2] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc. — A comprehensive guide to building NLP applications using Python and NLTK, including text preprocessing and feature extraction.

[3] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. — Covers the theory and practice of deep learning, useful for extending future work in semantic resume analysis using neural networks.

[4] Russell, S. J., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson. — A classic textbook on AI concepts, including classification algorithms and ethical AI system design.

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. — Describes the machine learning library used in the project, including implementation details of SVMs and TF-IDF.

[6] Singhal, A., & Srivastava, R. (2019). An Efficient Method for Resume Classification Using Supervised Machine Learning Algorithms.

International Journal of Engineering and Advanced Technology (IJEAT), 8(6S3), 219–223. — Focuses on machine learning-based classification of resumes, relevant for benchmarking the current project.

[7] Mallick, P. K., & Das, S. (2021). Intelligent Resume Screening System Using Natural Language Processing and Machine Learning Techniques. Procedia Computer Science, 185, 128–135. — Discusses an approach similar to the proposed system, using NLP and ML for automated candidate screening.

[8] Arora, A., & Narula, V. (2020). A Machine Learning Approach for Automated Resume Classification and Recommendation. International Journal of Computer Applications, 975, 8887.— Provides insights into automated recommendation systems based on resume content.

[9] Zhou, Z. H. (2021). Machine Learning. Springer. — A modern academic textbook covering both foundational and advanced machine learning techniques applicable to text classification.

[10] Jain, A., & Verma, A. (2020). Role of Artificial Intelligence in Modern Recruitment Process. International Journal of Advanced Science and Technology, 29(5), 12037–12045. — Explores how AI is transforming recruitment and candidate evaluation in HR systems.