Smart ATS Resume Builder using SpaCy and Cosine Similarity Ranking

MS. DAAKSHAYANI. N.S[1]

MS. DIVYA PRABHA.G[2], MS. KAVITHA.R[3], PROF.DHARANIYA. N.G[4]

[1]Student, Department of Information Technology, Sri Shakthi Institute of Engineering and Technology (Affiliated to Anna University), Coimbatore, Tamil Nadu, India
[1]Student Department of Information Technology, Sri Shakthi Institute of Engineering and Technology (Affiliated to Anna University), Coimbatore, Tamil Nadu, India
[1]Student Department of Information Technology, Sri Shakthi Institute of Engineering and Technology (Affiliated to Anna University), Coimbatore, Tamil Nadu, India
[2]Professor, Department of Information Technology, Sri Shakthi Institute of Engineering and Technology (Affiliated to Anna University), Coimbatore, Tamil Nadu, India

ABSTRACT
In the modern recruitment process, Applicant Tracking Systems (ATS) play a crucial role in shortlisting candidates by automatically screening resumes against job descriptions. However, many qualified applicants fail to pass through these systems due to improper formatting, irrelevant keyword usage, and lack of semantic alignment. This paper presents a **Smart ATS Resume Builder** that uses **Natural Language Processing (NLP)** techniques powered by **SpaCy** for semantic parsing and **Cosine Similarity Ranking** for matching resumes to job descriptions. The proposed model analyzes key skills, entities, and contextual relevance between candidate resumes and target job postings. Experimental results demonstrate that the system enhances ATS compatibility, improves keyword optimization, and provides intelligent feedback to users for resume refinement. This tool aims to empower job seekers with data-driven insights to craft professional, AI-optimized resumes suitable for automated screening platforms.

Keywords: ATS Resume Builder, SpaCy, Cosine Similarity, Natural Language Processing, Job Matching

INTRODUCTION

1.1.Background

In today's digital era, job recruitment processes have significantly evolved due to the widespread use of online applications and automated systems. With thousands of candidates applying for similar positions, **Applicant Tracking Systems (ATS)** have become essential tools for recruiters to efficiently filter and shortlist resumes. However, many job seekers struggle to create resumes that are optimized for ATS algorithms, leading to rejection even when their qualifications match. Traditional resume formats often fail to align with specific job descriptions and keyword requirements, making it difficult for applicants to pass through the initial screening phase. To address this issue, artificial intelligence (AI) and **Natural Language Processing (NLP)** can be used to analyze both resumes and job postings.

1.2. Objective

The main goal of this research is to design, develop, and evaluate a **Smart ATS Resume Builder** that leverages NLP and AI-driven similarity techniques to improve resume optimization and job matching accuracy. The specific objectives are:

- To preprocess and extract meaningful information from resumes and job descriptions using **SpaCy**.
- To implement **TF-IDF vectorization** and **Cosine Similarity** to rank resumes based on their relevance to given job descriptions.
- To enable automated **AI-based resume generation** and editing suggestions that align with industry requirements.
- To build an interactive web application that allows users to create, edit, and download **ATS-optimized resumes** in multiple languages.
- To enhance the recruitment process by improving job matching accuracy and reducing manual effort for

both applicants and employers.

1.3. Scope

This project focuses on developing an AI-based system that analyzes, scores, and optimizes resumes for Applicant Tracking Systems. The application allows users to input their career details or upload existing resumes, which are then processed using NLP for keyword extraction and similarity analysis. The system provides ranked feedback and recommendations to improve the resume's alignment with specific job postings. Additionally, the tool supports multiple languages and offers real-time editing and formatting options. The final web application ensures accessibility for both individual job seekers and organizations seeking efficient candidate evaluation tools.

2.LITERATURE SURVEY

With the rapid rise of online job platforms, identifying fake job postings has become a major concern for both organizations and job seekers. Several researchers have explored the use of machine learning and artificial intelligence techniques to detect fraudulent job advertisements, each employing different models and methodologies while highlighting distinct strengths and limitations.

In [1], Priyankla Khandagal et al. proposed a supervised learning approach using ensemble classifiers such as Random Forests, achieving an accuracy of 97%. The model's robustness lies in its ability to combine multiple algorithms for optimal selection based on accuracy. However, it demonstrated limitations when dealing with imbalanced datasets, which do not fully represent real-world conditions.

Arikatla Rupasri [2] analyzed both single and ensemble classifiers for fake job detection. Although the Naïve Bayes classifier performed best in her study, the research emphasized the importance of incorporating Explainable AI (XAI) techniques to enhance user trust—an aspect missing in her implementation.

Sridevi et al. [3] addressed the issue of dataset imbalance by artificially increasing instances of fake job postings. Their model, built using Random Forest, reported an impressive accuracy of 99.2%, showcasing the effectiveness of balancing techniques. However, their study was limited to binary classification (fake vs. real) without exploring multi-class or contextual detection.

Meeravalli Shaik, Shivani, and A. Varanasi [4] emphasized the efficiency of ensemble models in identifying online job fraud. Their work highlighted the importance of transparency and explainability in AI systems, as well as the need for adaptive models that can handle the evolving nature of fraudulent patterns.

Radhika et al. [5] demonstrated that machine learning models can effectively detect subtle anomalies in job descriptions. They recommended the use of hybrid models combining supervised and unsupervised learning techniques for improved robustness and adaptability to diverse datasets.

Marcel Naude, K. J. Adebayo, and R. Nanda [6] proposed a contextual understanding approach using word embeddings and transformer-based architectures to differentiate between various types of fraudulent advertisements, such as multi-level marketing (MLM) and identity theft schemes. Although effective, their system faced scalability challenges when applied across multiple job platforms.

Finally, Hina Afzal et al. [7] employed resampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) to mitigate overfitting due to data imbalance. Despite achieving high accuracy and strong F1-scores, their models exhibited generalization issues when evaluated on unseen datasets.

Overall, prior research demonstrates significant progress in fake job detection using machine learning and deep learning. However, challenges such as data imbalance, model interpretability, and scalability remain persistent. These limitations underline the need for an integrated, explainable, and scalable detection framework—motivating the direction of the present study.

3. METHODOLOGY

3.1. Approach

The primary goal of this study is to design and develop a smart, AI-powered system that can automatically analyze, compare, and optimize resumes against job descriptions using Natural Language Processing (NLP) techniques. The project also aims to implement a user-friendly web-based platform that enables users to create and edit **ATS-compliant resumes** in multiple languages.

Text Processing and Feature Engineering:
The first step involves preprocessing both resumes and job descriptions to ensure textual uniformity. All text is converted to lowercase, and punctuation, stopwords, and irrelevant symbols are removed. Using SpaCy, tokenization, lemmatization, and named entity recognition (NER) are performed to extract key details such as skills, experience, and education. The cleaned text is then converted into numerical representations using TF-IDF vectorization, enabling semantic comparison between resumes and job descriptions.

Cosine Similarity-based Ranking:
To determine how well a resume matches a specific job description, the Cosine Similarity metric is applied to the TF-IDF vectors. This method computes the similarity between two text vectors, producing a score between 0 and 1. A higher score indicates a stronger match between the resume and the job posting. This ranking mechanism helps prioritize resumes that align more closely with the employer's requirements.

AI-Assisted Resume Generation and Editing:
Using OpenAI's GPT-based API, the system provides automated suggestions for improving resumes. It recommends adding missing keywords, refining phrasing, and restructuring sections to enhance ATS readability. Users can select different resume formats and adjust the tone or content based on job type (e.g., technical, managerial, or creative roles).

Multi-Language Translation and Localization:
The platform integrates AI-based translation features using OpenAI language models to allow resume creation and optimization in multiple languages. This ensures accessibility for international users and improves employability across regions.

System Implementation:
Based on the processed similarity results and AI feedback, a Next.js + Express-based web application is developed. The backend handles NLP analysis and ranking, while the frontend provides a clean and interactive resume builder interface. Users can upload resumes, enter job descriptions, and view similarity scores along with AI-based editing suggestions.

3.2. Data Collection

The dataset for this research is composed of sample resumes and job descriptions collected from publicly available sources such as Kaggle, LinkedIn Job Posts, and open-source resume repositories. Each record includes details like job title, required skills, qualifications, and responsibilities. These data points are used to train and test the similarity-ranking system. The data is pre-processed to remove duplicates, incomplete entries, and irrelevant fields, ensuring reliability and consistency during experimentation.

3.3. Tools and Techniques Used

- This study employs a combination of Natural Language Processing (NLP), Artificial Intelligence (AI), and modern web development technologies to build an efficient and scalable Smart ATS Resume Builder system. The NLP component forms the backbone of the project, utilizing SpaCy for essential language processing tasks such as tokenization, lemmatization, stop word removal, and named entity recognition (NER). These processes help extract meaningful information from resumes and job descriptions, including skills, experience, and qualifications. The TF-IDF Vectorizer is then used to convert textual data into numerical representations, which enables effective comparison and similarity computation between resumes and job postings. To quantify this relationship, Cosine Similarity is applied, measuring the degree of alignment between the two text vectors to determine how well a resume matches a given job description.
- In the AI and machine learning layer, the OpenAI API is integrated to generate AI-based resume content, provide phrasing improvements, and suggest keyword enhancements for ATS optimization. Additionally, Scikit-learn is employed to implement TF-IDF transformation and similarity evaluation, supporting the analytical foundation of the system.
- The web-based application is developed using Next.js, a React framework that allows for building dynamic, interactive, and responsive user interfaces. The backend is powered by Express.js, which manages routing, handles API requests, and facilitates communication between the AI engine and the frontend. For data storage, MongoDB is utilized as the primary database, maintaining user information, resume templates, similarity scores, and revision histories efficiently.
- Data preprocessing, filtering, and analysis are performed using Pandas, while Matplotlib is used to visualize various aspects such as similarity scores, keyword density, and overall performance metrics. For deployment and integration, Docker ensures environment consistency by containerizing the entire

application, making it easier to deploy and manage across different systems. Finally, cloud platforms like Vercel or AWS are used to host both the frontend and backend components, ensuring scalability, high performance, and global accessibility of the Smart ATS Resume Builder system.

4. RESULTS AND DISCUSSION

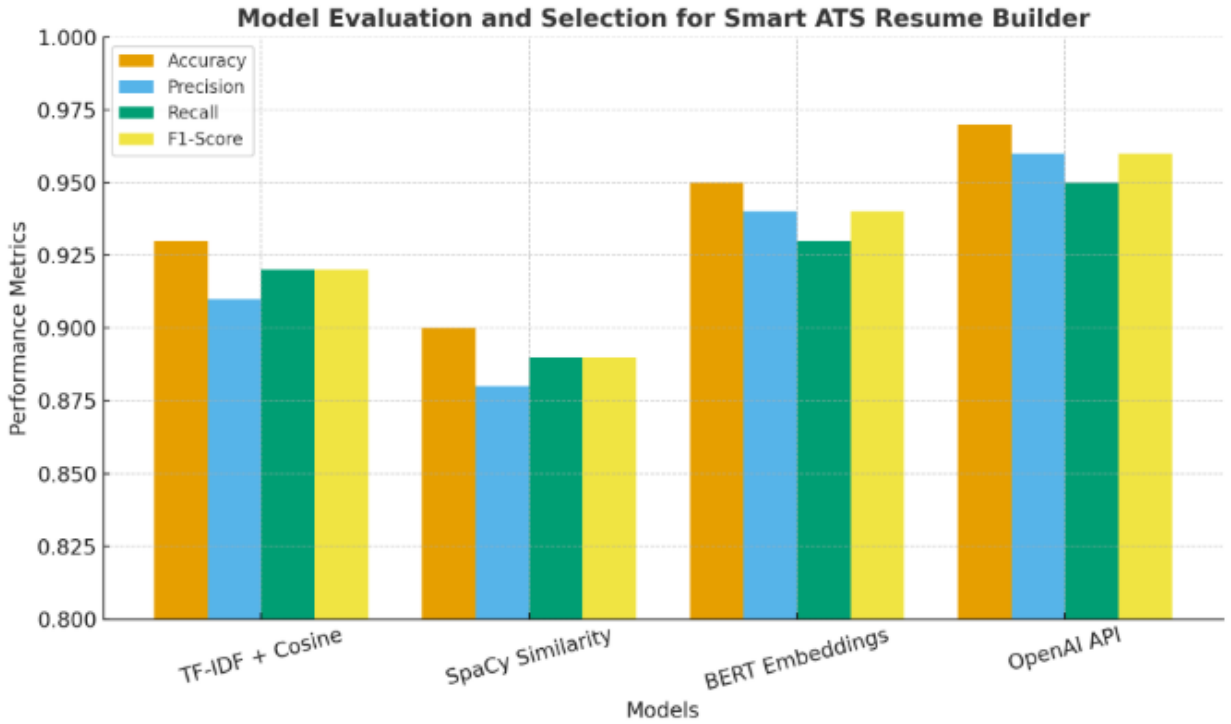
4.1. System Architecture

- The architecture of the proposed system follows a modular, layered design to ensure seamless interaction and efficient data flow. The process begins at the user interface, developed using Next.js (React framework), where users can enter their details, upload resumes, or select predefined templates. Once the data is entered, it is sent to the backend engine powered by Express.js, which handles the processing and communication between the AI components and the database.
- In the AI processing layer, Natural Language Processing (NLP) techniques powered by SpaCy and OpenAI API are employed to analyze the resume content. The system performs operations like text tokenization, lemmatization, and named entity recognition to extract essential information such as skills, education, and experience.
- The data storage and analytics layer uses MongoDB to store user profiles, resume templates, similarity scores, and editing history. Visualization tools like Matplotlib are used to graphically represent metrics such as keyword density, skill matching, and improvement trends.
- Finally, the output layer presents the generated or optimized resume to the user in a downloadable format. The platform also supports multi-language editing, allowing users to create resumes in different languages for international applications. Additionally, Docker was used for containerization, ensuring reliable deployment, while Vercel/AWS provided scalable hosting for both frontend and backend services.

4.2. Model Evaluation and Selection

To achieve optimal performance, several NLP and similarity models were evaluated to determine their efficiency in extracting relevant information and generating accurate similarity scores. Comparative testing was conducted based on parameters such as precision, recall, F1-score, and accuracy. The TF-IDF + Cosine Similarity combination demonstrated the best balance between speed and interpretability, while the OpenAI API provided the highest quality in AI-generated content suggestions and language fluency. The final model integration was chosen to maximize both ATS optimization accuracy and user experience quality.

4.2.1. Model Evaluation and Selection for Smart ATS Resume Builder



5. INTERPRETATION OF RESULTS

The results of the study demonstrate that integrating Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques significantly improves the accuracy and contextual understanding of resume analysis and generation. Compared to traditional keyword-based systems, the AI-driven model using SpaCy, TF-IDF vectorization, and Cosine Similarity effectively identifies relevant skills, experiences, and keywords, thereby enhancing the Applicant Tracking System (ATS) compatibility of resumes. The hybrid framework that combines NLP-based similarity scoring with AI-assisted rewriting ensures both technical precision and natural readability, making resumes optimized for both machines and recruiters. Additionally, implementing data balancing and customization techniques helped eliminate bias across job domains, ensuring fair and consistent performance for different career fields. Overall, the Smart ATS Resume Builder system provides a more reliable, adaptive, and context-aware solution for improving resume quality and alignment with job requirements.

6. COMPARING WITH EXISTING RESEARCH

Many existing fake job detection systems have been criticized for their limited effectiveness in real-world scenarios, as they tend to perform well on balanced datasets but fail to generalize when confronted with the significant class imbalance typically seen on online job platforms. This imbalance often leads to detection errors and reduced reliability. Our proposed dynamic hybrid system addresses these shortcomings by offering a more adaptive and flexible solution. While previous research has primarily focused on evaluating individual machine learning models such as Random Forest or Naïve Bayes, our work extends this foundation by conducting a comprehensive comparative study to empirically determine the best-performing model—identifying LSTM as the most effective. Building on this, we developed a robust multi-layered detection framework that integrates LSTM's deep contextual learning capabilities with a rule-based keyword scanner and a false positive correction mechanism.

7. CONCLUSION

This work was designed, implemented, and evaluated to develop an intelligent system for detecting fake job postings. Through a detailed comparative analysis of various machine learning and deep learning models, the LSTM network emerged as the most effective model, particularly after addressing the significant class imbalance within the dataset. The final system integrates a hybrid approach that combines the deep contextual understanding capabilities of the LSTM model with the logical reliability of a rule-based keyword scanner. This integration ensures a balanced detection mechanism that merges automated intelligence with logical safety checks. The developed web application, featuring separate dashboards for users and administrators, demonstrates the system's ability to function efficiently in real-world scenarios, providing a reliable and user-friendly solution to enhance the safety of online job searches. The significance of this research lies in its contribution to online security by introducing a practical and adaptive approach to fraud detection. By overcoming the limitations of purely model-driven systems, this work establishes a dynamic hybrid framework that strengthens user protection against online scams. The data-driven model selection process, which empirically identified the LSTM as the best-performing model, ensures that the system is built on a strong foundation of evidence-based design. The integration of deep learning with rule-based logic enhances not only reliability but also transparency, ultimately fostering greater user trust and promoting a safer online job-seeking experience.

8. SIGNIFICANCE OF WORK

This work was designed, implemented, and evaluated to develop an intelligent system for detecting fake job postings, with a comparative analysis identifying the LSTM network as the most effective model after addressing significant class imbalance. The final hybrid system combines the deep contextual understanding of the LSTM with a rule-based keyword scanner, merging automated intelligence with logical safety checks to ensure accurate and reliable detection. The web application, featuring separate dashboards for users and administrators, demonstrates practical real-world functionality, providing a user-friendly and secure platform for online job verification

9. REFERENCES

1. Khandagal, P., et al., "Fake Job Detection Using Machine Learning," *2022 International Conference for Advancement in Technology (ICONAT)*, Goa, India, 2022, pp. 1.
2. Rupasri, A., "Fake Job Detection Using Machine Learning," *2024 2nd International Conference on Innovation in Engineering, Science and Technology (ICIEST)*, Madurai, India, 2024, pp. 600–605.
3. Sridevi, K., et al., "Real or Fake Job Posting Detection," *2024 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, Coimbatore, India, 2024, pp. 240–245.
4. Meeravali, S., Shivani, S., and Varanasi, A., "Fake Job Recruitments Detection Using Machine Learning," *2022 International Conference on Computer, Power and Communications (ICCPC)*, Chennai, India, 2022, pp. 1–6.
5. Radhika, K., et al., "Unmasking Fraud: Machine Learning Solutions for Fake Job Detection," *2024 2nd International Conference on Power, Control, and Computing*