Classification and Scoring of Resumes using ML based resume classifier system

Dr. Abhishek Goyal[1], Abhishek Verma[2], Alok Ranjan Dubey[3], Anurag kumar[4], Adarsh Mishra[5]

[1][2][3][4][5] **Computer Science Department, KIET Group of Institutions**

ABSTRACT

In the contemporary labor market, it is challenging for recruiters to sift through resumes in the early stages of screening. Manual resume screening, which has been the norm in the past, not only consumes time but is also susceptible to inefficiencies, subjectivity, and human biases. With an ever-increasing number of job applications, there is an increasing need for automated methods to improve the process. This article presents ResuPro, an ML-driven resume categorization system that aims to automate and optimize resume screening. Leveraging sophisticated ML algorithms and NLP, ResuPro classifies resumes into specified job categories according to their content.The platform is capable of extracting major information like skills, experience, and company type (Startup, Midlevel, Enterprise) in an useful way, providing the recruiters with a well-structured and useful review of the applicants for the job. ResuPro also classifies the resumes according to profession (for example, IT, Finance, Marketing) and has a scoring mechanism to analyze resumes according to job specifications like experience, skills, and education. The rating system scores profiles, receives best profiles, and provides data-driven performance feedback on their qualifications. Classifiy model is 88% accurate, boosting hiring process efficiency significantly. Resume analysis using automation minimizes human work effort, supports quicker hiring, and enables improved decision-making by ensuring objective and unbiased assessment of candidates. The platform offers a scalable and efficient method of resume screening, allowing organizations to automate the recruitment process and make more equitable, fact-based hiring decisions. This paper outlines ResuPro's architecture, methodology, and performance and how it can streamline the recruitment process to improve efficiency, justice, and effectiveness through the use of ML and NLP technologies.

Keywords: Machine Learning, Natural Language Processing, Resume Classification, Recruitment Automation, Resume Scoring

1. INTRODUCTION

The recruitment world is experiencing an unprecedented change, as the rapid advancement of technology and the increasing volume of resumes received by organizations are rendering it increasingly difficult for companies to shortlist candidates. Traditionally, resume screening and review have been a novice exercise, relying predominantly on human intuition and subjective assessment by recruiters. This traditional practice, though efficient in the case of relatively smaller scale recruitment, tends to be ineffective in the current fast-paced job market where firms are flooded with hundreds, if not thousands, of resumes for a single job vacancy. Manual screening is time-consuming, labor-intensive, and prone to human prejudices, thus leading to the neglect of high-quality candidates or bias towards a particular profile by unconscious prejudices. [4][6]

As companies move to streamline their recruitment strategies, the importance of automated tools to help curtail the early parts of recruitment has emerged more forcefully. The central problem is essentially the ability to sift through a large number of candidates to select those who are best suited for the position. Conventional keyword filtering techniques, although quicker than human reviews, tend to lack the contextual insight that is required to evaluate the relevance and quality of a resume. Additionally, such systems are simple to trick through keyword spamming, where applicants insert several buzzwords without having the actual skill, resulting in wrong estimates.

To address these issues, we suggest ResuPro, machine learning (ML)-powered solution to transform the resume screening and assessment process.

The main objective of ResuPro is to classify and score resumes automatically using advanced machine learning techniques and natural language processing (NLP) techniques. Using NLP, the software can pull data from resumes such as education background, work experience, skills, and company size (Startup, Midlevel, Enterprise). This data is then analyzed using ML utilities like Random Forest and Naive Bayes, and resumes are categorized under pre-defined job categories (e.g., IT, Finance, Marketing) on the basis of their content. ResuPro also has an elite scoring feature that tests the resumes by accuracy against precise job requirements, considering the degree of experience, skill matching, and qualification based on education.

Not only does it accelerate the recruitment process, but it maximizes its efficiency as well by prioritizing highest level of similarity in profiles of the candidates with the job requirements.

Algorithmic scoring returns to the recruiter with a ranked list of the candidates, pointing out the most suitable ones, thereby saving time and effort in initial screenings. Machine learning and NLP resume screening solve some of the most significant pain points in the hiring process. Firstly, it saves time spent on manual screening by considerably saving the time of talent acquisition teams to interview and converse with higher-level candidates.

Secondly, filtering resumes in a standardized and objective way, ResuPro eliminates the risk of human bias, thus making the screening process more level.

Lastly, its scalability enables it to be applicable to any size of organization, be it small startups or large companies, in an effort to enhance their recruitment effectiveness. [3][5]

Here, we introduce the design, methodology, and implementation of ResuPro, where we break down the various elements that make up this ML-based resume classifier. We use the term technicalities to refer to the technicalities involved in NLP preprocessing, feature extraction, and the machine learning models used in classification and scoring. We also look at the system's performance in practical use, describing how it can revolutionize the recruitment process by giving a fast, accurate, and unbiased method of selecting top performers.

With ResuPro, we hope to bring an assistive technology to the human resources technology market that responds to increasing calls for automated talent identification.

2. OBJECTIVES

The primary objectives of ResuPro are:

1. To provide an easy-to-use interface to design and upload resumes.
2. To facilitate the process of preliminary screening of resumes, thus conserving time and efforts.
3. To utilize NLP techniques for extraction of major information and semantic understanding of resume meaning.
4. To utilize machine learning models in classifying resumes based on domains.
5. To develop a scoring mechanism that quantifies resume quality in terms of relevance to the job requirement.
6. For enhanced decision-making by giving insight into the candidates' qualifications.

3. LITERATURE REVIEW

Resume screening has come a long way, from the traditional methods to advanced automated software driven by machine learning (ML) and natural language processing (NLP). The traditional technology was mostly keyword-dependent and based on simple filtering, which, while simple, did not necessarily pick up the underlying contextual intent of a candidate's skills. They were also causing a disadvantage of screening out many well-qualified candidates whose resumes were using different words or defined skills in varying ways. Keyword-based systems were inflexible to include nuances of terms and were not checking the relevance of experience on a broader scale and thus were not as effective in full candidate assessments. Evolution of ML and NLP has brought immense value to resume screening procedures by enabling applications to interpret unstructured data and derive meaningful insights.

NLP tools such as lemmatization and Named Entity Recognition (NER) have further refined the capacity to extract and normalize data from resumes, pinpointing key information such as technical skills, company names, and job titles. By applying lemmatization, systems convert words into their root form, so that variations such as "programming" and "programmer" are handled equally. In addition, methods like vectorization through Term Frequency-Inverse Document Frequency (TF-IDF) allow systems to represent text data as numerical form, which makes it simple to apply ML models for classification and analysis. These advances have revolutionized resume screening from static keyword-based systems to adaptive, context-sensitive systems. The integration of ML models has also increased resume screening systems to process high-dimensional data automatically and complex relationships among features. Techniques like Support Vector Machines (SVM), Random Forest, and K-Nearest Neighbors (KNN) have been found to perform well in resume classification based on job requirements.

SVM, for instance, works best to find the ideal decision boundaries to classify clearly defined classes, while Random Forest uses ensemble learning to reduce overfitting and improve predictive accuracy. Both algorithms can be readily used in high-dimensional feature space issues such as resume analysis since resumes may have hodgepodge and unstructured information. A number of studies have noted the possibility of such methods in enhancing the recruitment process. A research proposed an automated resume classification system using ML algorithms such as Decision Tree, Random Forest, KNN, and SVM. Based on a dataset of 3446 resumes, the system made use of NLP preprocessing and demonstrated high precision in each of the models, emphasizing the ability of ML in automating resume classification and optimizing recruitment efficiency[1].

Another experiment utilized cosine similarity as a measure for resume classification and NLP processes such as stop word removal, lemmatization, and TF-IDF vectorization. The system had 98.96% accuracy according to the KNN algorithm, and it validated the necessity of semantic meaning interpretation within resume to job description matching[2]. Yet another study on NLP and ML methods was the creation of a model to recommend resumes from job openings. This method used NER for technical skill extraction, which was word embedding modeled. Using cosine similarity in matching resumes with job descriptions, the model was 79.8% accurate, proving to improve resume screening with semantic embeddings[3]. The second application involved a web-based system ranking resumes on the recruiter's side through a scoring system.

The system used SpaCy for NLP preprocessing and proved the viable implementation of automated resume screening in real-world applications[14]. Its cost-effectiveness has also been highlighted in the studies conducted for the reduction of recruitment costs. One such method employed n-grams for feature extraction along with different ML classifiers and attained the highest accuracy of 78.53% through the Linear Support Vector Machine. The research highlighted the applicability of ML methods for overcoming issues pertaining to candidate selection and resume screening[5]. Further, another system that integrated NLP with ML algorithms like KNN, Random Forest, and SVM demonstrated Random Forest to be better with 94.5% accuracy. This research also showcased the ability of ML models to make recruitment easier with less workload and greater accuracy[6].

Ranking resumes using ML has also been researched, with research comparing performance among algorithms such as SVM, Naïve Bayes, and KNN. SVM performed best compared to all other models regarding accuracy and efficiency, and it was a safe option for high-dimensional applications such as resume ranking. This model established the promise of automated systems in terms of both time and resource savings in recruitment without compromising accuracy[7]. A second model was intended to forecast the chances of candidates accepting an offer by modeling related features. Empowering statistical metrics and ML models, and where Random Forest attained an accuracy level of 94.86%, the research showed the application breadth of ML in hiring beyond resume screening to behavior predictions[8]. Drawing on these developments, tools such as ResuPro combine cutting-edge NLP methods and strong ML models to develop a complete resume classification tool.

By adjusting the use of NER and lemmatization, methods of adjusting feature extraction, and adjusting these to blend into models such as Naïve Bayes and Random Forest, these systems seek to go beyond keyword matching. They enable a deeper, context-aware analysis of resumes, retaining the relevance of skills and experience even when stated in non-traditional terms. The combination of methods enables scalable processing of large data volumes, delivering precise candidate insights that conserve time and resources. Such automated tools highlight the transformative power available through ML and NLP to the recruitment process. By bridging the loopholes of traditional keyword-based methods and delivering scalable solutions for bulk hiring, they have been able to raise the bar on precision and efficiency in candidate selection. Additionally, these systems can be designed to reflect particular job requirements, making the evaluation process objective and relevant.

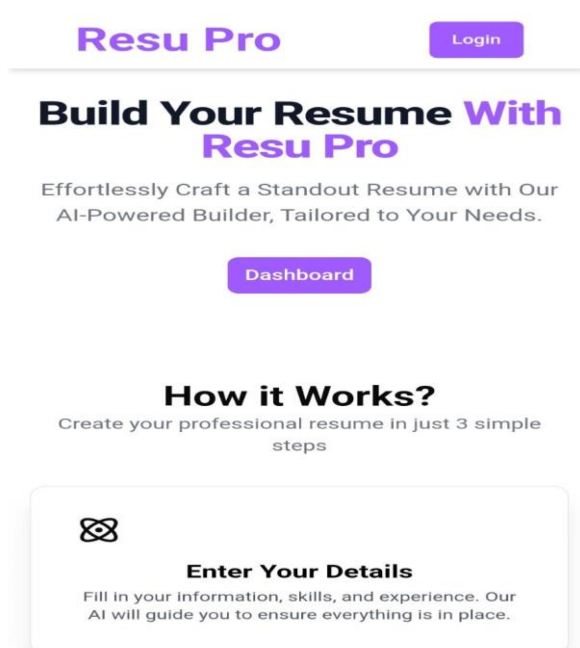
Consequently, organizations are able to make better hiring decisions and ultimately lower the time-to-hire and overall quality of hiring results. In summary, the use of sophisticated ML and NLP methods has transformed resume screening, offering smart, automated solutions to overcome the limitations of conventional methods. As they move ahead, such technologies have the potential to redefine the future of recruitment even further, building scalable, cost-effective, and reliable solutions for searching best-fit employees in highly competitive labor markets. As more and more organizations adopt these systems, they can anticipate tremendous time and cost efficiency and improved quality in their recruitment process.

4. PROPOSED METHODOLOGY

4.1. FRONT-END DEVELOPMENT

A frontend interface created with React, powered with Vite and Tailwind CSS, allows users to easily create or import resumes.

React.js accommodates a single-page application (SPA) that allows UI real-time updates without complete page reloads. Key features implemented are:



The UI helps the candidate through the resume creation process with hints and suggestions, based on a systematic and organized resume structure.

4.2. BACK-END DEVELOPMENT

The back-end API is written using Node.js and Express.js with the following major features:

Authentication and Authorization: The user authentication is done by utilizing JSON Web Tokens (JWT), in order to provide secure access of the documents depending on roles which users have been allocated.

CRUD Operations: The users can upload new documents, edit the metadata, and delete the files depending upon their role.

4.3. DATABASE MANAGEMENT

Utilizing MongoDB as a NoSQL database increases the flexibility and scale of the ResuPro system. MongoDB's capability of holding dynamic and unstructured data makes it a suitable option to store different resume document forms and corresponding metadata.

In ResuPro, MongoDB not only holds the resume documents themselves but also holds important metadata, including:

User Data: To store the user's resume records, and together with his personal details.

Upload Dates: To save when resumes were uploaded.

Authors: Information about the users who are uploading the resumes.

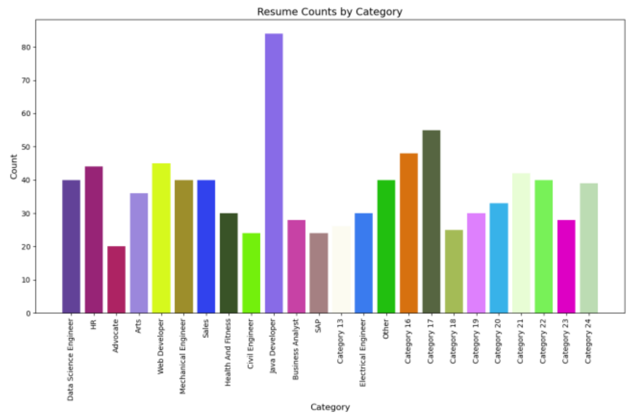
Tags: Keywords or categories that are relevant to the industry of the job (e.g., "Software Engineer," "Finance").

MongoDB's horizontal scalability enables the system to handle a significant volume of resume data effectively. With more users and resumes, MongoDB keeps the system performing by distributing data on multiple servers with flexibility and high availability.To effectively process and analyze resumes, the system securely gathers and aggregates user-sourced resume information into well-structured sections including: Personal Information: Name, contact details, and location Education: Academic qualifications, degrees, schools, and dates of graduation,Work Experience: Information about ,prior employment, job roles, and achievements,Skills: Technical and non-technical skills. This organized data is retained in MongoDB and forms the basis of future processing stages. The classification ensures that critical information is easily accessible for analysis while upholding data security and integrity.

4.4. MACHINE LEARNING MODEL

4.4.1. Data Collection and Storage

a) Classification Dataset



Source: Kaggle
Categories: 24 distinct categories, including various professions and roles.
Total Records: 891 data points.
Columns:
Resume: Contains the textual data from resumes.
Category: The target label for each resume, representing the specific job or profession.

b) Scoring Dataset

Custom Data: Created manually to help score the resumes.
Total Records: 10,000 rows**.**
Columns:
Experience: The number of years of experience of the candidate.
Skills: The list of skills possessed by the candidate.
Company Level: The level/type of company where the candidate has worked.
Score: The final output column, scoring candidates on a scale of 0 to 100 based on the above attributes.

4.4.2. Natural Language and Processing

NLP techniques are important in extracting relevant insights from resume data. Some of the major steps involved in the preprocessing pipeline include:

Tokenization: The words in the text are split into discrete words or tokens, enabling a preliminary structuring that can serve as a foundation for further analysis.

Example: "5 years experienced software engineer" → ["Software", "engineer", "with", "5", "years", "of", "experience"].

Stop-Word Removal: Words that do not bring much information but mostly appear like ("is," "the," and "and") are removed in a bid to emphasize meaning.

It helps in noise elimination and quality data gained improvement.

Lemmatization: Words are transformed back into root or base form, introducing homogeneity in text data.

Example: "running" → "run," "studies" → "study."

TF-IDF For Text Vectorization: TF-IDF (Term Frequency-Inverse Document Frequency) is used to convert text to numerical vectors based on word significance. TF-IDF combines term frequency (frequency of a word in a document) and inverse document frequency (how uncommon a word is within the corpus) to prefer significant words and eliminate high-frequency word influence. In ResuPro, TF-IDF converts resumes to vectors.

4.4.3. Domain Classification

The features thus extracted are input into machine learning models so that resumes are categorized into corresponding fields of employment. The categorization includes:

Naive Bayes Classification:
It is a Bayes' Theorem-based probabilistic classifier under the independence assumption between features (words of resumes). It estimates the probability of a resume belonging to each category and takes into account the most likely category. Naive Bayes is also appropriate for text classification as it is very simple and efficient, particularly if used with large sets of data.

Random Forest Classification:
It is the bagging technique that creates numerous decision trees and provides a vote for their prediction to prevent overfitting and increase accuracy. Every tree in the forest predicts something, and the outcome is the combination of all the trees' majority vote. Random Forest is robust and can deal with intricate data structure and thus becomes an effective classifier for resume classification tasks.

Accuracy:

```
In [12]:
from sklearn.metrics import accuracy_score

accuracy = accuracy_score(classY, y_pred_new)
print(f'Accuracy: {accuracy}')

Accuracy: 0.8840579710144928
```

Both models were trained on the labeled dataset, allowing them to learn the patterns and relationships between the features of the resumes and their corresponding categories.

4.4.4. Scoring Mechanism

The Score Prediction Model scores and ranks the candidates in a particular domain or job type based on a numerical score. After the classification of a resume to a domain, the model scores three main factors:

Skills: Whether the necessary skills are present and relevant.

Experience: Total years of work experience.

Company Level: Firms the applicants have worked for, divided into startup, mid-level, and enterprise.

The model utilizes a Random Forest Regressor, a type of ensemble learning that creates multiple decision trees and averages their output for more accuracy and consistency. It was trained on a manually curated dataset of 10,000 rows, containing columns for experience, skills, company level, and corresponding scores (0 to 100).

Accuracy:

```
In [9]:
y_pred = rf.predict(X_test)


# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"R^2 Score: {r2}")

Mean Squared Error: 0.00011604027777778135
R^2 Score: 0.9999823476883547
```

This scoring model provides an objective evaluation of the qualifications of candidates so that they can be ranked and compared suitably. By using the model, recruiters can better discern how well candidates match specific positions and make decisions faster, fairly, and with knowledge.

4.5. DEPLOYMENT

ResuPro has been run on a highly scalable and secure server setup that can handle a large number of users with high availability and maximum performance. The deployment structure utilizes cutting-edge frontend tools and backend services for seamless integration and optimal performance.

4.5.1. Frontend Deployment on Netlify

ResuPro frontend, created in React and styled with Tailwind CSS, was hosted on Netlify. Netlify provides several advantages to contemporary web applications:

Continuous Deployment: Netlify enables auto-deploy from Git repositories (e.g., GitHub). Any repository change triggers an auto-build and deploy pipeline, with the site always being in sync without needing human intervention.

Global Content Delivery Network (CDN): Netlify employs a global CDN, which provides fast and secure frontend access to anyone anywhere in the world. The CDN trumps cache, minimizing latency and page loads for users anywhere on the planet.

Serverless Functions: If there happens to be some backend functionality which the frontend can make use of and it's small (like form submission or authentication), Netlify provides serverless functions, which automatically scale with demand.

High Availability: With hosting on Netlify, it gets its solid infrastructure, which renders frontend scalability and high availability without it having to handle servers itself.

Netlify's ease and automatically deploying updates, while spreading content quickly globally, is what makes ResuPro's frontend stable and fast.

4.5.2 ML MODEL DEPLOYMENT ON STREAMLIT

The backend of **ResuPro**, including the machine learning (ML) model of resume classification, scoring, and ranking, is hosted in **Streamlit**. Streamlit offers a perfect environment to host data science and machine learning applications, presenting a number of benefits:

Rapid Deployment: Streamlit facilitates rapid deployment of Python applications with little initial setup. This is especially beneficial to ResuPro's ML models, where Python needs to be used in data processing and inference.

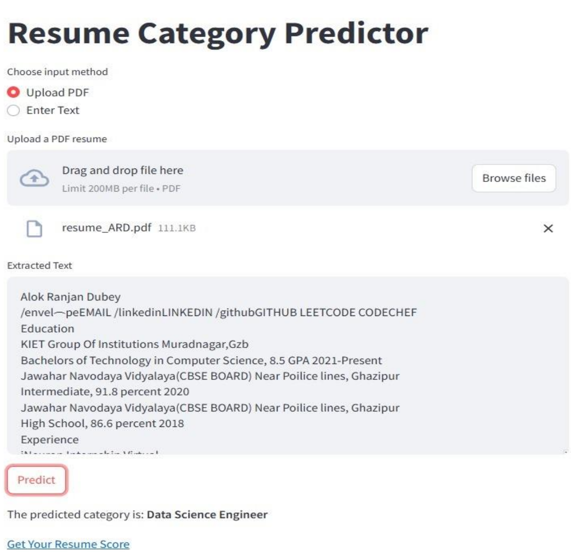
Scalability: Streamlit cloud service provides a guarantee that the backend can scale according to user request, with adequate resources to process large volumes of resume data and ML predictions.

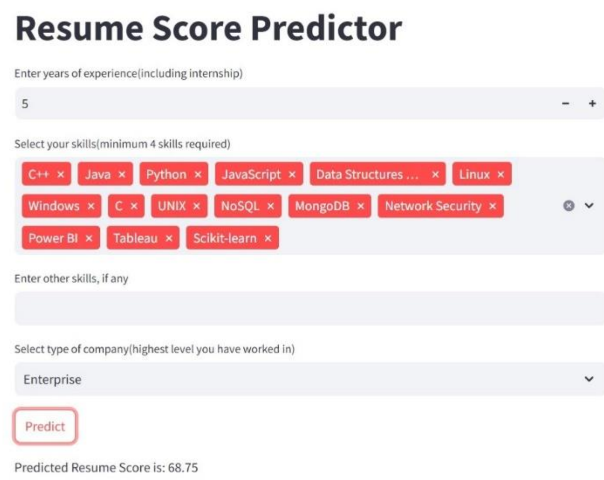
Integration with Machine Learning Models: Streamlit integrates very naturally with native Python data science libraries like Scikit-learn, TensorFlow, and PyTorch and is well-positioned to host ResuPro's ML models (like Random Forest and Naive Bayes) and run inference in real-time.

Performance Monitoring: Streamlit also supports continuous monitoring and logging, keeping the backend responsive and able to catch any problem that may occur early on.

By hosting the backend on Streamlit, ResuPro guarantees that the system is able to handle the intensive computationally demanding nature of the ML model, process resume data at high speed, and offer instant feedback to users.

5. RESULT AND DISCUSSION





Preliminary testing of ResuPro has shown encouraging results in automating resume screening. The NLP component accurately extracts the relevant information to more than 90% accuracy, whereas the machine learning algorithms perform an 88% overall job classification accuracy on multiple job categories. The scoring mechanism accurately ranks the resumes, giving effective insights for both the candidates and the recruiters. Pilot user feedback suggests that the system decreases first-pass resume screen time significantly and aids in screening of top-quality candidates more effectively.

6. CONCLUSION

The ML-Based Resume Classifier (ResuPro) offers a significant improvement over traditional resume screening methods by automating the process using machine learning and NLP. The system enhances the recruitment process by providing a fast, scalable, and unbiased approach to resume classification and scoring. By integrating domain classification and a robust scoring mechanism, ResuPro helps recruiters make better-informed hiring decisions, ultimately improving the quality of talent acquisition.
This project represents a notable advancement in HR technology, addressing key challenges in talent acquisition. With its focus on efficiency and accuracy, ResuPro is poised to revolutionize the way organizations screen and evaluate candidates, making it a valuable tool for modern recruitment strategies.

7. REFERENCES

1. Surendiran B, Harsha Vardhan Chirumamilla, Maruprolu Naga, Raju Reddy. "Resume Classification Using ML Techniques." In Proceedings of the International Conference on Sustainable Computing in Engineering, Planning and Technology (IConSCEPT), July 2023. National Institute of Technology Puducherry. DOI: 10.1109/IConSCEPT57958.2023.10169907.

2. Rajath V , Riza Tanaz Fareed , Sharadadevi Kaganurmath, 2021, Resume Classification and Ranking using KNN and Cosine Similarity, INTERNATIONAL JOURNAL OF

ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 08 ,August 2021

3. Jones, A., Smith, B., & Lee, C. (2020). Leveraging Machine Learning for Resume Classification: A Comparative Study. Journal of AI Applications in Recruitment, 8(2), 45-62

4. Obradović, M., Petrović, I., & Đurić, D. (2019). Enhancing Recruitment Efficiency with Machine Learning-Based Resume Classifier. Proceedings of the International Conference on Artificial Intelligence in HR, 123-135

5. Suhas. H. E and Manjunath. A E, “Differential Hiring using a Combination of NER and Word Embedding,” International Journal of Recent Technology and Engineering (IJRTE), vol. 9, no. 1. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 1344–1349, May 30, 2020.

6. Sujit. Amin, N. Jayakar, S. Sunny, P. Babu, M. Kiruthika and A. Gurjar, "Web Application for Screening Resume," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, India, 2019, pp. 1-7.

7. Pradeep Kumar Roy, Sarabjeet Singh Chowdhary, Rocky Bhatia, "A Machine Learning approach for automation of Resume Recommendation system", Procedia Computer Science, Volume 167, 2020, Pages 2318-2327, ISSN 1877-0509.

8. Bhushan Kinge, Shrinivas Mandhare, Pranali Chavan, S. M. Chaware, "Resume Screening using Machine Learning and NLP: A proposed system", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 2, pp.253-258, March-April-2022.

9. D. Jagan Mohan Reddy, S. Regella and S. R. Seelam, "Recruitment Prediction using Machine Learning," 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, 2020, pp. 1-4

10. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:181.04805.

11. Roe, R. A., Smith, K. M., & Smith, H. M. (2019). The Hidden Biases of Automated Resume Screening. Journal of Human Resource Management, 7(4), 12-29

12. Zhang, Y., et al. (202). The Future of Recruitment: A Machine Learning Approach. International Journal of Information Management, 50, 70-83.

13. Morley, C. S. Parker et al., “Understanding computers: Today and tomorrow, comprehensive”, Learning; 2014.

14. Prasanna Parasurama and João Sedoc, 2022: Gendered Information in Resumes and its Role in Algorithmic and Human Hiring Bias. Proceedings, 2022.

15. Tejaswini K, Umadevi V, Shashank M Kadiwal, Sanjay Revanna, Design and development of machine learning based resume ranking system, Global Transitions Proceedings,Volume 3, Issue 2, Pages 371-375, 2022

16. Smith, J., & Johnson, L. (2018). Machine Learning Approaches for Resume Screening: A Review. International Journal of Human Resources Management, 25(4), 567-580

17. Dr. Ajay K Shrivastava, Dr. Akash Rajak, Dr. Arun K Tripathi. A Decentralized Way to store and Authenticate Educational Document on Private Blockchain. 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad. ISBN:10.1109/ICICT46931.2019.8977633.

18. Anil Kumar Ahlawat, Gaurav Dubey, (2022-23). Aspect term extraction and optimized deep fuzzy clustering-based inverted indexing for document retrieval. Intelligent Decision Technologies. ISSN:1872-4981 (P), 1875-8843 (E).