NLP-DRIVEN AUTOMATED RESUME SCREENING FOR EFFICIENT CANDIDATE PROFILING.

Mrunali Bhagyawant
dept. Electronics & Telecommunication
BRACT's Vishwakarma Institute of Information Technology
Pune, India.

Archana Ratnaparkhi
dept. Electronics & Telecommunication
BRACT's Vishwakarma Institute of Information Techology

***Abstract*— The challenge of screening countless resumes typically hampers the efficiency of hiring, resulting in delays, inconsistency, and potential bias in the selection process. Indeed, the process of manually screening resumes can take significant time and introduce human error, creating the risk of overlooking qualified candidates. To mitigate such inefficiencies, we have created an automated resume classification system based on NLP and machine learning. Our system takes any data it receives on a resume, clean the data, and classify applicants into existing job categories defined by the recruiting firm (Software Developer, Data Scientist, Electrical Engineer, etc.). The model saves screening time and reliably classifies applicants accurately and consistently by implementing the TF-IDF vectorization scheme and One-vs-Rest classifier. What is most invigorating about this project is it achieves high accuracy, which means a classifier can be reliably used for classifying resumes, thus eliminating some of the workload for the recruiter and making the selection process more systematic. The model is also complemented with an interface the recruiter can log into and simply upload resumes, thus immediately being classified. This system has the potential to further interface with Applicant Tracking Systems (ATS) and provide valuable skill gap analysis and upskilling recommendations. Overall, this project will go a long way to streamline the hiring process and create a more efficient, scalable, and fair solution.**

Keywords— Machine Learning, Bias Mitigation, Algorithmic Decision-Making, Resume Screening, Human Bias, Non-Traditional Candidates

I. INTRODUCTION

Organizations, even with advertisement efforts to fill job vacancies, receive high volumes of resumes for every job opening, making it more and more difficult to perform manual resume screening to short list the candidates with the best qualifications for the role recruited. HR professionals offering a screening step in the resume screening process are frequently subject to cognitive bias, thus impacting the efficiency of the recruitment process or missing out on potentially great applicants for the position. As such, automated resume parsing systems, driven from Natural Language Processing (NLP) and machine learning, are emerging as revolutionary tools in the modern resume screening process.

A resume parser is an automated system specifically designed to extract relevant attributes such as work experience, education, skills, and certifications from informal resumes. Typical resumes are unstructured and in multiple formats, but by leveraging NLP, the parser processes the information and standardizes it into structured data elements on a single readable format.

Nonetheless, it is not easy to create an effective and precise resume parsing system in a competitive marketplace, due to multiple variations of application types, resumes, writing styles, and organizational content structures, all factoring variability into the systems design process. Attending to these complexities includes many key design steps, including but not limited to; data pre-processing, data cleaning, vectorization, resume screening, and data splitting. Data pre-processing is the stage where the linguistic and general text data is prepared, normalized, and tokenized for analysis.

The pre-processing process would begin with data cleaning that ensures any required inverted & language, irrelevant wording, and typographical issues are corrected, reconstructed or modified. Vectorization treats a defined length of encoded or cleaned language into a series of numbers that can be processed through the normal machine learning algorithm. In addition, data splitting is the process of ensuring the machine learning model is properly designed to train and test for the desired methodology goals of general expansion of the model to a larger and a newer set of data information.

II. LITERATURE SURVEY

A. Automated Resume Screening Using Natural Language processing.

[1] present an automated resume screening system making use of NLP techniques such as Sentence-BERT (S-BERT) and cosine similarity to extract embedded information from resumes and job descriptions and to automatically compare them. The S-BERT model is then applied to resumes and job descriptions to obtain meaningful and relevant information. The higher similarity scores over a threshold will ultimately pass the relevant resumes on to the expert to cross-check and back to the best

applicant. This use of NLP technology gives reliability and efficiency in the working of the organization. Conventional manual resume screening processes have several problems such as:
(1) the need for an inordinate amount of time to shortlist applicants from a large pool of candidates.
(2) the identification of only 60-70% of eligible candidates.
(3) the difficulty in processing resumes in many formats and languages.

The use of simple approaches for the resume screening system along with the comparison with the automated course of action makes evident the increased efficiency in terms of precision.

B. Resume parser

[2] the authors describe a Resume Parser system based on Natural Language Processing (NLP) and Machine Learning functions for extracting specific details of candidates' resumes to assist with evaluations (e.g., personal information, work history, education, skills, and qualifications). The system accepts different document formats, such as PDF and Word, and can also review files that are scanned in different languages through Optical Character Recognition (OCR). The parsing process includes, but is not limited to, document ingestion, document preprocessing, parsing, storing documents, and output/output displays.

The potential for the system to quickly and accurately conduct candidate evaluations can significantly streamline the resume extraction and parsing processes while reducing the time and manual efforts required to review candidates' resumes. Additionally, it improves candidate-job matching through AI-based recommendation systems, while also improving candidate evaluation through context-based and dynamic NLP processes.

C. Maintaining the Integrity of the Specifications

The advancement of resume parsing technologies has come a long way with the introduction of Natural Language Processing (NLP) and Machine Learning approaches (Pawar et al., 2024). Numerous studies point to the significance of extracting relevant information about job seekers by analyzing constructs such as personal information, work history, educational background, skills and qualifications from various types/styles of resumes (PDF, Word) and languages. When working with scanned documents, Optical Character Recognition (OCR) can also be used to ingest and preprocess data into the system properly. Studies suggest that automating the parsing of documents is more time efficient and improves the accuracy of candidate assessment, reduced manual labor. In addition, AI recommendation systems and improved comprehension of context enabled via NLP have also improved the accuracy in matching candidates/job postings as the demand for resume parsing systems continues to grow in response to the new formats and languages. This body of literature shows that resume parsing and advanced systems to organize and administer applicant selection, has provided more efficient ways to manage recruitment and selection processes and possibly improve overall hiring decisions.

D. An Automated Resume Screening System Using Natural Language Processing and Similarity.

Daryani et al. (n.d.) suggest an automated system to screen candidate resumes using Natural Language Processing and similarity measures to facilitate candidate selection. The proposed system consists of two phases. In the first phase, candidates' resumes are parsed and summarized through information extraction, often utilizing various NLP techniques, including Tokenization, Stemming, Part-of-Speech (POS) Tagging, and Named Entity Recognition (NER). In the second phase, a content-based recommender system is employed using the Vector Space Model and Cosine Similarity to compare candidate resumes with job descriptions and rank candidates based on their relevant similarities.

The authors concluded the study with recommendations for future work, including social networking data from job-related platforms such as LinkedIn and GitHub to enhance the screening process, collaborative filtering to filter candidates based on ratings of candidate profiles, and recommendations to implement Latent Semantic Analysis (LSA) to improve semantic similarity between resumes and job descriptions.

E. Resume screening with Natural Language Processing and Similarity

Daryani et al. describe a resume screening system that automates candidate evaluation using Natural Language Processing (NLP) techniques, demonstrating the rapid progress of automated candidate assessment. This resume screening system consists of two phases of operations: the first phase involves information extraction (including Tokenization, Stemming, Part-of-Speech (POS) tagging, and Named Entity Recognition) to parse and summarize the resumes, and the second phase recommends candidates' profiles and job descriptions according to a content-based recommendation system (i.e., Vector Space Model and Cosine Similarity) and its rank of relevance. Also, the authors documented the application of this model to specific job description summaries at Amazon. Candidate 2 received the highest similarity score of 0.680 to the job description and Candidate 4 scored second with a similarity score of 0.651. The authors suggested possible changes/improvements to the system such as utilizing a social networking site (e.g., LinkedIn and GitHub data) for more refined candidate recommendations, implementing collaborative filtering ratings of candidates' and job identifier profile similarities, and Latent Semantic Analysis (LSA) for processing and improving semantic similarity of resumes and job descriptions. This literature advances the understanding of how an automated system using NLP approaches may change the game of recruiting for both prospective candidates and employer advocates (better known as recruiters) in achieving higher accuracy and efficiency evaluating candidates' qualifications and matches (specific fit).

F. Resume Classification System using Natural Language Processing & Machine Learning Techniques

[3] describes their Resume Classification System that utilizes Natural Language Processing (NLP) and Machine Learning techniques to classify resumes into predetermined job categories. Their approach consists of five steps: first, they collected and labeled 962 resumes across 25 job categories; next, they preprocessed the data by removing stop words and stemming and lemmatizing. For feature engineering, they vectorized their text data with Term Frequency-Inverse Document Frequency (TF-IDF). In the model construction step, they applied nine classifiers for comparative performance analysis. Then, they assessed classifiers through precision, recall, and F-score. Support Vector Machine (SVM) classifiers scored well, most notably Linear SVC and Stochastic Gradient Descent (SGD), with their classification accuracies exceeding 96%, with as high as 99.6% and an F-score of 1.00. The authors recommend future enhancements related to ethics in automatic resume matching and underscore the importance of increasing the dataset size to help achieve better generalization and performance for their system.

G. Resume Screening using Machine Learning.

[4] introduce an automated resume screening system that employs Natural Language Processing (NLP) techniques for candidate evaluation. The system has two phases: information extraction using methods like Tokenization and Named Entity Recognition, and a content-based recommendation system utilizing the Vector Space Model and Cosine Similarity to rank candidates. In their study of Amazon job descriptions, Candidate 2 scored the highest similarity at 0.680, followed by Candidate 4 at 0.651. Proposed improvements include integrating social networking data and using Latent Semantic Analysis (LSA). This research demonstrates how NLP can enhance the accuracy and efficiency of candidate evaluations for both job seekers and recruiters.

H. Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening.

[5] The research collected on discrimination in the hiring process is indicative of considerable notation from human decision-making and decision robots, concerning potential beliefs regarding gender, race, or years of school, which often lead to the exclusion of qualified candidates for job openings (Bertrand & Mullainathan, 2004; Moss-Racusin et al., 2012). Cowgill's article, "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening" (2020), claims that well-designed algorithms and automation can limit these procedures and increase productivity by identifying high-potential candidates who chance being overlooked by human article readers. The extent of the potential advantage of automating these decision-making processes is reliant on the design of a transparent model, positively biased trained datasets, and human decision oversight to ascertain fairness (Dastin, 2018; Binns, 2018). Alongside this, Cowgill acknowledges the concepts of hybrid models, which brings together human decision-making, along with the productivity of a decision robot, to facilitate cross-disciplinary teamwork to assess and improve these decision-making processes.

I. A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring.

[6] Algorithmic bias poses a serious difficulty for job hiring as companies further invest in AI. Algorithmic bias leads to practices that can be unfair to diverse candidates. This survey people's literature examines AI techniques to mitigate bias in hiring, especially in Natural Language Processing (NLP) and deep learning, two places NLP is used is in applying word embedding and - or sentiment analysis to examine job descriptions or candidate communications. Bolukbasi et al., (2016) examined biases in word embeddings, demonstrated that word embeddings reflect the societal bias, and produced algorithms to enhance the word embeddings to be fairer when hiring. In deep learning, methodologies like adversarial training initiated by Zhao et.al., (2017) aim to reduce the impacts of sensitive traits during decision-making. Data augmentation (Kim et al., 2020) generates a synthetic dataset to help balance datasets using underrepresented groups thus creating bias reduction in datasets through augmentation methods standing up against data biases. Human AI partnership is a vital component noted by Binns (2018) who states that human input builds up AI methodologies by improving AI systems with a contextual understanding. For the future, research fires up on more bias mitigation processes, craftsmanship methodologies from diverse technologies like AI-enhanced um uncapped quantum computing (Albaroudi et al. 2024).

J. Intelligent Hiring with Resume Parser and Ranking using Natural Language Processing and Machine Learning.

[7]Technological advancements such as Natural Language Processing (NLP) and Machine Learning (ML) enhance recruitment techniques. Several systems create resume parsing and ranking. Goecks et al. (2009) was the first to promote activities to extract structured data from resumes. Nadarajah et al. (2020) published a study indicating nominal recall and precision rates through the application of recurrent neural networks (RNNs). Bonus candidate profile resources add additional information to a candidate's profile. Zhang et al. (2021) confirm this using social media like LinkedIn or GitHub, giving candidates a better chance to appeal and contextualize clients. Beyer et al. (2018) engage the reading of recruiters and employer research college or university studies. Beyer et al. upholds the clarity about precision and recall to rate accuracy. Another engagement feature of both the recruited participants and the previous quantitative studies by Kaiser et al. (2020) and Mohd Sadiq et al. provides mutual symbiotic relationship representations of the information these systems yield after a candidate is selected for review and rejects blind, objective assessments of hiring.

III. PROPOSED SYSTEM

Below is our proposed system explained. We have explained the how we developed the model and trained it stepwise.

A. *Data Acquisition*

The dataset used for this study was sourced from a CSV file containing two columns:

1) Category - Represents the job title or domain (e.g., Python Developer, Electrical Engineer).
2) Resume - Contains the raw text of the applicant's resume.

The dataset was loaded using Pandas, and an initial exploration of the data was conducted to understand its shape and distribution. This is crucial for understanding the data balance and ensuring the training model receives an adequate variety of categories.

$$\text{df.shape} \Rightarrow (n_{\text{rows}}, n_{\text{columns}})$$

B. *Data Visualization and Distribution Analysis*

To visualize the distribution of resume categories, Seaborn was used to create a count plot, displaying the number of resumes per category. This helps identify category imbalances, which may need to be addressed.

$$\text{sns.countplot(df['Category'])}$$

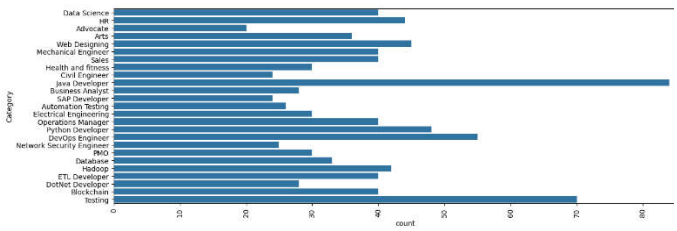

Fig.1 Output of countplot method

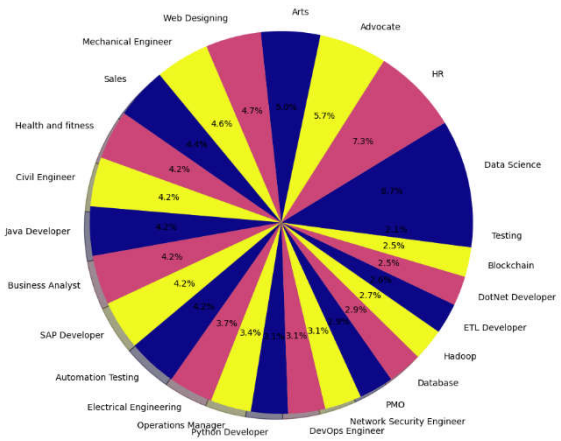

Fig.2 Distribution analysis

C. *Data preprocessing*

Resumes often contain noise in the form of URLs, hashtags, mentions, and special characters, which do not contribute to meaningful categorization.

The following steps were used to clean the text data:

1) Remove URLs, hashtags, mentions, and non-ASCII characters using regular expressions.
2) Replace punctuations with spaces to prevent tokenization errors.

The preprocessing function is mathematically represented as:

$$\text{clean\_text}(x) = \text{re.sub}(\text{pattern}, \text{replacement}, x)$$

Where x is the raw resume text, and pattern is the regular expression used to match undesired characters.

D. *Label Encoding*

The categorical variable, Category, was transformed into numerical form using label encoding. Each unique job title was assigned a unique integer ID.

Let C={c1,c2,...,cn} represent the set of categories, where nnn is the number of categories. The label encoder assigns each category a unique ID i:

$$\text{le.transform}(c_i) = i, \quad i \in [0, n-1]$$

E. *Text Vectorization Using TF-IDF*

To convert the resume text into numerical features suitable for model training, we applied Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. This approach measures the importance of a word in relation to a resume (document) and the entire dataset (corpus).

The TF-IDF formula is given by:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

Where:

a) t is a term (word)
b) d is a document (resume)
c) D is the corpus (collection of resumes)
d) TF (t,d) is the term frequency, i.e, the number of times t appears in d.
e) $\log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$ *is the inverse document frequency, which reduces the weight of commonly occurring terms.*

The result of this step is a sparse matrix where each resume is represented by a vector of TF-IDF values for each term in the dataset.

F. *Splitting the Dataset*

The dataset was split into training and testing sets using an 80-20 ratio to ensure that the model was trained on a large portion of the data while still having a reserved subset for validation.

Mathematically:

$$X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}} = \text{train\_test\_split}(X, y, \text{test\_size} = 0.2)$$

Where:

a) Xtrain and Xtest are the training and testing features (TF-IDF vectors).

b) Ytrain and Ytest are the training and testing labels (categories).

G. *Model Training Using K-Nearest Neighbors (KNN)*

We trained the classifier using the K-Nearest Neighbors (KNN) algorithm with the One-vs-Rest strategy. This multi-class classification approach trains binary classifiers for each class, with the "rest" representing all other categories.

For each resume vector x, KNN predicts the category by identifying the majority class among the k closest neighbors (in Euclidean distance):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Where:

a) $d(x,y)$ is the Euclidean distance between resume vectors x and y.

b) k is the number of nearest neighbors.

The One-vs-Rest classification involves training n classifiers, where each classifier distinguishes between one category and all others.

H. *Model Evaluation*

After training, the model's performance was evaluated on the test set by calculating the accuracy score, which is the ratio of correct predictions to the total number of predictions made:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

The accuracy score helps assess how well the model categorizes resumes into their respective job categories.

The accuracy that our model achieved is 98%.

IV. Result and Discussion

TABLE I. Comparision

Comparison of Manual and Automated Résumé Screening Processes		
Aspect	***Manual Screening***	***Automated Screening***
Bias	Subject to human biases such as education, race, or gender preferences.	Reduces bias by using data-driven decisions based on resume content.
Consistency	Inconsistent due to noise and subjective variations among recruiters.	Provides consistent evaluations, applying the same criteria for all resumes.
Time Efficiency	Time-consuming, especially with large applicant volumes.	Fast and scalable, instantly categorizing resumes using machine learning.
Non-Traditional Candidates	Tends to overlook non-traditional candidates from unconventional backgrounds.	Considers all candidates equally, focusing on content rather than format.
Productivity	Improved productivity with algorithms, but risks reinforcing biases.	Enhances productivity by reducing errors and selecting the best-fit candidates.
Noise and Variability	Affected by external factors like mood or environment, leading to random noise.	Eliminates noise, ensuring decisions are based on content alone.
Feature Engineering	Relies on subjective judgment, often overlooking resume details.	Uses TF-IDF and NLP to systematically analyze relevant resume features.
Scalability	Does not scale well with large numbers of applicants.	Highly scalable, processing large volumes of résumés quickly and efficiently.
Accuracy	Limited accuracy with no feedback loop for improvements.	Provides high accuracy with a feedback loop to continuously improve the model.
Handling Unstructured Data	Difficult for humans to handle unstructured resume formats.	Uses text cleaning to handle and structure unstructured data.

V. Conclusion

The introduction of an NLP-based automatic resume screening system has the potential to lead the hiring process to be more efficient and precise. The system can effectively reduce the amount of time that is spent on manually reviewing resumes and eliminates the biases that are often exhibited through human driven processes. This is a consistent and scalable approach to automatic resume screening. The application of phrasing techniques such as TF-IDF vectorization and One-vs-Rest classifier can classify resumes in different employment positions accurately and reliably. The algorithms accuracy rates also provide a level of confidence to recruiters in terms of data-driven decision-making, so that the automatic resume screening system can save time by providing a list of candidates for candidates profiled with skills gaps based on the job description while also integrating applicants tracking system (ATS) that also have candidate profiling. This method provides a certain "fair" way of conducting automated resume screening which is efficient and scalable. Future work on this research area could focus on expanding the dataset for increased generalizability and to start utilizing social media data to improve recommendations.

References

[1] Resume Screening using Machine Learning. Dr. Sandeep Tayal[1], Taniya Sharma[2], Shivansh Singhal[3], Anurag Kumar Thakur[4]

[2] Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening. Bo Cowgill Columbia University.

[3] A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. Elham Albaroudi, Taha Mansouri and Ali Alameer.

[4] Intelligent Hiring with Resume Parser and Ranking using Natural Language Processing and Machine Learning. Sayed Zainul Abideen Mohd Sadiq, Juneja Afzal Ayub, Gunduka Rakesh Narsayya, Momin Adnan Ayyas, Prof. Khan Tabrez Mohd. Tahir Students, Dept. of Computer Engineering, AIKTC, Mumbai University, India.

[5] Resume Classification System using Natural Language Processing & Machine Learning Techniques Irfan Ali, Nimra , Ghulam Mujtaba, Zahid Hussain Khand, Zafar Ali, and Sajid Khan.

[6] Resume Parser Arpita Pawar, Sanika Kosabe, Akshay Warde, Prof. Kirti Mhamunkar. Students of Information Technology Saraswati College Of Engineering.

[7] Automated Resume Screening Using Natural Language Processing. Dr. D. Lakshmi Padmaja1 , Ch. Vishnuvardhan2 , G. Rajeev3 , K. Nitish Sanjeev Kumar4

[8] Resume Screening with Natural Language Processing in Python Shradha Pujari Student, Department of Computer Engineering, Vidyalankar Institute of Technology, Mumbai, India.

[9] RESUME SCREENING AND RECOMMENDATION SYSTEM USING MACHINE LEARNING APPROACHES Lokesh. S, Mano Balaje. S, Prathish. E and B. Bharathi Department of Computer Science and Engineering, SSN College of Engineering, Rajiv Gandhi Salai, Kalavakkam - 603110