SPEECH & LANGUAGE PROCESSING DA1

**Optimizing Resume Analysis with SpaCy: Enhancing Candidate Selection Through NLP**

**VISHNUPPRIYAN PL – 21BAI1323**

**JEEVA SVITHRA S – 21BAI1659**

**PAPER 1:  
a. Authors:** Nimish Patil, Shubham Yadav, Vikas Biradar  
**b.** **Year:** April-2023  
**c.** **Title of the Article:** Resume Parser and Analyzer Using NLP  
**d.** **Objective:** The objective of the work is to create a system that can parse resumes using Natural Language Processing (NLP) to extract key information such as personal details, education, work experience, skills, and publications, and then analyze this information to provide scores and suggestions for improvement.  
**e.** **General Challenges in the Work:** The challenges faced in this work include the variability in resume formats, the need for accurate extraction of information, and the development of algorithms that can handle diverse data formats (like PDF and DOC files).  
**f.** **Methodology Used in the Proposed Work:** The methodology involves using Python libraries such as NLTK and SpaCy for NLP tasks, including text extraction, entity recognition, and data preprocessing steps like removing stop words, tokenization, and lemmatization.  
**g.** **Name of the Dataset Used in the Proposed Work:** The document does not explicitly mention the name of the dataset used. It also does not specify whether the dataset is publicly available or not.  
**h.** **Evaluation Metrics and Its Values:** The document does not provide specific evaluation metrics or their values for the proposed system.  
**i.** **Issues Faced in Developing the Proposed Work:** The document does not list specific issues faced during the development of the proposed work.  
**j.** **Future Work:** The document does not outline future work plans for the proposed system.

**PAPER 2:**

**a. Authors:** Chuan Qin, Hengshu Zhu, Tong Xu, Chen Zhu, Liang Jiang, Enhong Chen, Hui Xiong  
**b. Year:** 2018 **c. Title of the Article:** Enhancing Person-Job Fit for Talent Recruitment: An Ability-aware Neural Network Approach  
**d. Objective:** To propose an end-to-end Ability-aware Person-Job Fit Neural Network (APJFNN) model that reduces reliance on manual labor in recruitment and provides better interpretation of fitting results by leveraging historical job application data.  
**e. General Challenges in the work:** The challenges include the subjective, incomplete, and inefficient nature of human judgment in traditional Person-Job Fit methods, and the need for effective and comprehensive techniques to measure the matching degree between talent qualifications and job requirements.  
**f. Methodology used in the proposed work:** The APJFNN model uses a word-level semantic representation based on Recurrent Neural Network (RNN) for both job requirements and job seekers' experiences. It employs four hierarchical ability-aware attention strategies to measure the importance of job requirements and the relevance of candidates' experiences to specific ability requirements.  
**g. Name of the Dataset used in the proposed work:** A real-world dataset provided by a high-tech company in China, which includes 17,766 job postings and 898,914 resumes. The dataset is not mentioned as publicly available.  
**h. Evaluation Metrics and its values:** The evaluation metrics used include Accuracy, Precision, Recall, F1-measure, and AUC. The APJFNN model outperforms baselines with significant margins, achieving an Accuracy of 0.7559, Precision of 0.7545, Recall of 0.7603, F1-measure of 0.7574, and AUC of 0.8316.  
**i. Issues faced in developing the proposed work**: Not explicitly mentioned in the provided text.  
**j. Future work:** Not explicitly mentioned in the provided text.

**PAPER 3:**

**a. Authors:** Chuan Qin, Hengshu Zhu, Tong Xu, Chen Zhu, Chao Ma, Enhong Chen, and Hui Xiong.  
**b. Year:** 2020 **c. Title of the Article:** An Enhanced Neural Network Approach to Person-Job Fit in Talent Recruitment.  
**d. Objective:** The objective of the article is to propose a novel end-to-end Topic-based Ability-aware Person-Job Fit Neural Network (TAPJFNN) model to improve the efficiency and interpretability of talent recruitment by reducing the reliance on manual labor. **e. General Challenges in the work**: The challenges include the need to handle a large amount of data from online recruitment platforms, the subjectivity and inefficiency of traditional manual Person-Job Fit (PJF) assessments, and the scarcity of negative instances in the data, which can bias the learning algorithm.  
**f. Methodology used in the proposed work:** The methodology involves using Recurrent Neural Networks (RNNs) for semantic representation of job requirements and candidate experiences, hierarchical topic-based ability-aware attention strategies, and a refinement strategy based on historical recruitment records.  
**g. Name of the Dataset used in the proposed work:** The dataset used is not explicitly named in the provided context, but it is described as a large-scale real-world dataset. There is no information on whether it is publicly available.  
**h. Evaluation Metrics and its values:** The evaluation metrics include AUC, Accuracy, Precision, Recall, F1-measure, and Hit Ratio (HR). Specific values for these metrics are not provided in the context.  
**i. Issues faced in developing the proposed work:** Issues include dealing with the imbalance in the data due to the scarcity of negative instances and the need to design a novel learning algorithm for the job recommendation task to address the natural scarcity of negative instances. **j. Future work:** The document does not explicitly mention future work in the provided context. However, based on the content, potential future work could involve refining the model further, exploring the impact of non-textual features on algorithm fairness, and applying the model to other recommendation applications.

**PAPER 4:**

**a. Author:** Bhushan Kinge, Shrinivas Mandhare, Pranali Chavan, S. M. Chaware  
**b. Year:** 2022 **c. Title of the Article**: Resume Screening Using Machine Learning and NLP: A Proposed System  
**d. Objective:** The objective is to create a system that can screen resumes using machine learning and natural language processing techniques. The system aims to rank resumes based on their relevance to job descriptions and provide feedback to applicants on how to improve their resumes. **e. General Challenges in the work:** The challenges include the volume of resumes received by talent acquisition companies, the need for efficient and accurate screening methods, and the lack of feedback for applicants on why their resumes are rejected.  
**f. Methodology used in the proposed work:** The proposed system uses machine learning models such as K-Nearest Neighbors (KNN) or Support Vector Machines (SVM) for predictions. It also employs web scraping to gather additional information from LinkedIn and GitHub profiles, and cosine similarity for recommendations on resume improvement.  
**g. Name of the Dataset used in the proposed work (whether it is publicly available or not):** The dataset used in the proposed work is collected from the open platform Kaggle, which implies it is publicly available.  
**h. Evaluation Metrics and its values**: The paper presents an accuracy table with various methods and their corresponding accuracy percentages, ranging from 78.53% to 98.96%.  
**i. Issues faced in developing the proposed work:** The paper does not explicitly mention the specific issues faced during the development of the proposed work, but it does discuss the limitations of existing systems such as the inability to improve over time and the tendency to plateau with large datasets.  
**j. Future work:** The paper does not provide explicit details on future work, but it implies that the proposed system will continuously learn and adapt, addressing the limitations of current systems.

**PAPER 5:**

**a. Author:** Shubham Bhor, Harish Shinde, Vivek Gupta, Vishak Nair, and Prof. Manasi Kulkarni

**b. Year**: Not explicitly mentioned in the text provided, but it is implied to be recent given the references to current technologies.

**c. Title of the Article:** "Resume Parser Using Natural Language Processing Techniques"

**d. Objective:** The objective is to develop an intelligent resume parsing system using Natural Language Processing (NLP) techniques to automate the extraction and ranking of information from resumes, making the recruitment process more efficient and less biased.

**e. General Challenges in the Work:** Handling unstructured data in various formats, converting this data into a structured format suitable for comparison and ranking, and overcoming biases and inefficiencies inherent in manual processing.

**f. Methodology Used in the Proposed Work:** The proposed work uses NLP techniques to parse resumes, Optical Character Recognition (OCR) to handle various file formats, and a ranking algorithm to order candidates based on company-specific requirements. Named Entity Recognition (NER) is also used to fine-tune the NLP model for domain-specific terminology.

**g. Name of the Dataset Used in the Proposed Work:** The dataset details are not provided in the text; it is not mentioned if it is publicly available.

**h. Evaluation Metrics and Its Values:** The specific evaluation metrics and their values are not detailed in the provided text.

**i. Issues Faced in Developing the Proposed Work:** Challenges include dealing with various file formats, the need for domain-specific adjustments to NLP models, and ensuring fairness and accuracy in resume ranking.

**j. Future Work:** Future work includes expanding the system's capability to parse resumes from different platforms like LinkedIn, GitHub, and Naukri.com, integrating a wider range of psychometric tests, and enhancing the system's performance with a larger resume dataset.

**PAPER 6:**

**a. Author:** Sujit Amin, Nikita Jayakar, Sonia Sunny, Pheba Babu, M. Kiruthika, Ambarish Gurjar

**b. Year:** 2019

**c. Title of the Article:** "Web Application for Screening Resume"

**d. Objective:** The objective is to design a web application that screens resumes for job postings using machine learning and Natural Language Processing (NLP) to ease the recruitment process for recruiters by filtering and ranking resumes based on job requirements.

**e. General Challenges in the Work:** High time complexity when matching every candidate resume with every job posting on online recruitment sites. Manual viewing and analysis of a large number of resumes by recruiters, which is time-consuming and burdensome​.

**f. Methodology Used in the Proposed Work:** The web application uses machine learning to train the dataset for a specific job position. Natural Language Processing (NLP) with Named Entity Recognition (NER) and section-based segmentation is used for data extraction from resumes​.

**g. Name of the Dataset Used in the Proposed Work (Whether It Is Publicly Available or Not):** The dataset is not specifically named but consists of 220 resumes, with 200 used for training and 20 for testing. There is no indication that the dataset is publicly available.

**h. Evaluation Metrics and Its Values:** Accuracy was used as an evaluation metric. The model was trained and tested with a dataset of resumes, yielding the expected accuracy for the current implementation.

**i. Issues Faced in Developing the Proposed Work:** One of the major challenges was reducing the time complexity of matching candidate resumes with job postings​.

**j. Future Work:** Future work includes parsing resumes from various applications and websites (e.g., LinkedIn, GitHub, Naukri.com), adding psychometric tests, and enlarging the resume dataset to improve the system's performance.

**PAPER 7:**

**a. Author:** Chirag Daryani, Gurneet Singh Chhabra, Harsh Patel, Indrajeet Kaur Chhabra, Ruchi Patel.

**b. Year:** 2020​(7)

**c. Title of the Article:** "An Automated Resume Screening System Using Natural Language Processing and Similarity"​.

**d. Objective:** To automate the resume screening process using Natural Language Processing (NLP) and cosine similarity to extract relevant information from resumes and match them with job descriptions, making the recruitment process more efficient and reducing the burden on recruiters​.

**e. General Challenges in the Work:** Handling a large number of applications, dealing with unstructured resume formats, and ensuring fairness and accuracy in the screening process​.

**f. Methodology Used in the Proposed Work:** The proposed system works in two phases. The first phase involves information extraction from unstructured resumes using NLP techniques like Tokenization, Stemming, Part-of-Speech (POS) Tagging, and Named Entity Recognition (NER). The second phase ranks resume based on their content similarity with the job description using a Vector Space Model and cosine similarity​.

**g. Name of the Dataset Used in the Proposed Work (Whether It Is Publicly Available or Not):** The dataset details are not explicitly mentioned, and it is not clear if it is publicly available​.

**h. Evaluation Metrics and Its Values:** The evaluation metric used is cosine similarity, with specific similarity scores provided for different resumes compared to a job description. For example, similarity scores for candidates ranged from 0.4907 to 0.6803​.

**i. Issues Faced in Developing the Proposed Work:** One issue was the need to reduce the time and complexity of manually matching resumes with job descriptions​.

**j. Future Work:** Future work includes integrating social networking data to improve candidate recommendations, using collaborative filtering, and employing Latent Semantic Analysis (LSA) for better semantic similarity measures​.

**PAPER 8:**

**a. Author:** Suleiman Ali Alsaif, Minyar Sassi Hidri, Imen Ferjani, Hassan Ahmed Eleraky, Adel Hidri

**b. Year:** 2022

**c. Title:** NLP-Based Bi-Directional Recommendation System: Towards Recommending Jobs to Job Seekers and Resumes to Recruiters

**d. Objective:** Propose a system to match job seekers with jobs and help recruiters find candidates.

**e. Challenges:** Handling large data, matching candidates with jobs, and overcoming information gaps.

**f. Methodology:** NLP techniques, web scraping, data pre-processing, NER model training, and bi-directional matching.

**g. Dataset:** Resumes from DataTurks (public) and job listings from Indeed.

**h. Evaluation Metrics:** Accuracy, precision, recall, F1-score (e.g., accuracy: 99.885% for "Name").

**i. Issues:** Skill overlap and accurate differentiation between similar job roles.

**j. Future Work:** Improve experience relevance, build a skills ontology, and incorporate geo-location.

**PAPER 9:**

**a. Author:** Irfan Ali, Nimra Mughal, Zahid Hussain Khand, Javed Ahmed, Ghulam Mujtaba

**b. Year:** 2022

**c. Title:** Resume Classification System using Natural Language Processing and Machine Learning Techniques

**d. Objective:** Develop an automated system to classify resumes into job categories using NLP and ML techniques.

**e. Challenges:** Handling diverse resume formats, mapping resumes to job descriptions, and managing costs.

**f. Methodology:** NLP techniques for data preprocessing, TF-IDF for feature extraction, and various ML classifiers (e.g., SVM, Naïve Bayes).

**g. Dataset:** 962 labelled resumes across 25 job categories (publicly available).

**h. Evaluation Metrics:** Precision, Recall, F-Score, Accuracy (e.g., SVM achieved 99.6% accuracy).

**i. Issues:** Difficulty in finding standard datasets and dealing with low data size.

**j. Future Work:** Extend the model to match resume content with job descriptions for better recruitment processes.

**PAPER 10:**

**a. Author:** Thi-Thuy-Quynh Trinh, Thanh-Tuan Dang

**b. Year:** 2022

**c. Title:** Automatic Process Resume in Talent Pool by Applying Natural Language Processing

**d. Objective:** Propose a framework to enhance talent pool management by automating resume classification and ranking.

**e. Challenges:** Handling diverse resume formats, managing large volumes of data, and improving database efficiency.

**f. Methodology:** The framework uses NLP techniques (e.g., cosine similarity, word embedding with fastText) for resume and job description processing.

**g. Dataset:** 2,302 job descriptions and 8,529 resumes from the IT industry (collected online in 2020).

**h. Evaluation Metrics:** Similarity scores between resumes and job descriptions, ranking resumes based on these scores.

**i. Issues:** Need for dataset enhancement and reduction of manual tasks to improve efficiency.

**j. Future Work:** Implement and evaluate the proposed framework in real-world scenarios, enhancing datasets and automating more tasks.

**PAPER 11:**

**a. Author:** Dipti Suhas Chavare and Mrs. Archana Bhaskar Patil

**b. Year:** 2023

**c. Title of the Article:** Resume Parsing using Natural Language Processing

**d. Objective:** Automate resume screening using NLP and knowledge graphs to improve the hiring process.

**e. General Challenges:** Time-consuming manual screening, irrelevant resume content, poor job recommendations, and limited NLP applications in job search.

**f. Methodology:** Custom NER model using spaCy and BERT for entity extraction, with relation extraction to uncover hidden links.

**g. Dataset:** Not explicitly mentioned; involves manually labelled data.

**h. Evaluation Metrics:** Not specified.

**i. Issues Faced:** Training complexity, data labeling, and limited application scope.

j. Future Work: Expand techniques to other industries and improve job recommendation accuracy.

**PAPER 12:**

**a. Author:** Zainab Naveed, Bakhtawar Nisar, Dr. Muhammad Saifullah, and Junaid Iqbal Baig

**b. Year:** 2024

**c. Title of the Article: R**esume Ranking Using Natural Language Processing

**d. Objective:** Automate resume screening using NLP to rank candidates based on skills and experience, streamlining the hiring process.

**e. General Challenges:** Manual resume screening is time-consuming, often includes irrelevant data, and lacks a standardized format.

**f. Methodology:** Utilizes NLP techniques, including tokenization, syntactic and semantic analysis, and spaCy for Named Entity Recognition (NER). The model is trained and tested using Python libraries like NLTK, TextBlob, and spaCy.

**g. Dataset:** Resumes are manually labeled for training. The dataset specifics are not mentioned.

**h. Evaluation Metrics:** Accuracy, precision, recall, and F1-score were calculated for each resume's entity recognition.

**i. Issues Faced:** Challenges include the need for morphological, syntactic, and semantic analysis during preprocessing.

**j. Future Work:** Expand the application to other sectors such as Telecom, Healthcare, E-commerce, and public sector jobs.

**PAPER 13:**

**a. Author:** c

**b. Year:** 2023

**c. Title of the Article:** Resume Parser with Natural Language Processing

**d. Objective:** Develop an NLP-based system to assist HR departments in efficiently screening resumes by converting them to text, extracting relevant information, and comparing them to job descriptions.

**e. General Challenges:** Handling various resume formats, reducing human error, and fatigue during manual screening.

**f. Methodology:** Uses Named Entity Recognition (NER) and Regular Expressions to extract key details like name, designation, university, skills, and experience. Converts PDF/DOC files to text for processing.

**g. Dataset: Two datasets:** 200 resumes from GitHub, and additional datasets for global universities and skills.

**h. Evaluation Metrics:** Percentage similarity between resumes and job descriptions using cosine similarity.

**i. Issues Faced:** Limited dataset coverage, difficulty in processing certain data (e.g., dates), and ethical concerns related to bias in resume parsing.

**j. Future Work:** Expand the dataset, enhance functionality for web-based applications, and support more job positions by incorporating portfolio views and ranking resumes based on similarity to job descriptions.

**PAPER 14:**

**a. Author:** Rogelio Valdez-Almada, Oscar M. Rodriguez-Elias, César Enrique Rose-Gómez, María De Jesús Velázquez-Mendoza, Samuel González-López

**b. Year:** 2017

**c. Title:** "NLP and Text Mining to Identify Knowledge Profiles for Software Engineering Positions"

**d. Objective:** To develop a web application (KP GENERATOR) that identifies technical knowledge in software engineering resumes to aid in hiring.

**e. Challenges:** Processing unstructured resume text and integrating NLP and Text Mining effectively.

**f. Methodology:** Used Stanford CoreNLP for NLP and text categorization techniques to extract technical knowledge.

**g. Dataset:** Forty software engineering resumes (not publicly available).

**h. Evaluation Metrics:** Accuracy (98.1%) and F-measure (90%).

**i. Issues:** Variability in technical knowledge naming, processing time, and scalability.

**j. Future Work:** Enhancing expertise level detection and further testing for HR efficiency.

**PAPER 15:**

**a. Author:** Sai Tarun Boddu, Sujeeth Desu, Sreekanth Puli

**b. Year:** 2023

**c. Title:** "Resume Summarizer and Job Description Matcher Using Natural Language Processing and SpaCy"

**d. Objective:** To streamline the HR recruitment process by developing tools that summarize resumes and match them with job descriptions using NLP techniques.

**e. Challenges:** Variability in resume templates, missing/incomplete information, unstructured data, and diverse file formats.

**f. Methodology:** Used NLP techniques with SpaCy for resume screening and classification, employing regular expressions, OCR, and NLTK for data extraction and analysis.

g. Dataset: Not explicitly mentioned; tools were tested on various resume formats (PDF, DOCX, JPG, etc.).

**h. Evaluation Metrics:** Accuracy of 92% in entity extraction and 87% in summary generation.

**i. Issues:** Handling diverse resume formats and improving the extraction accuracy of contact details.

**j. Future Work:** Enhancing NLP techniques for better accuracy and extending the tool's application across various industries and job types.