2024 3rd International Conference for Innovation in Technology (INOCON)
Karnataka, India. Mar 1-3, 2024

NLP based Resume Analysis and Adaptive Skill Assessment System

Anitha Julian
Department of Computer science and engineering
Saveetha engineering college
Chennai, India
cse.anithajulian@gmail.com

Haripriya K
Department of Computer Science and Engineering
Saveetha engineering college
Chennai, India
text2.haripriya@gmail.com

2024 3rd International Conference for Innovation in Technology (INOCON) | 979-8-3503-8193-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/INOCON60754.2024.10511967

***Abstract*— The proposed work introduces an innovative recruitment approach integrating LinkedIn web scraping to compile a comprehensive dataset of job descriptions and essential skills. Utilizing a pioneering Word2Vec-driven machine learning model, it efficiently categorizes job descriptions, transforming hiring practices. Job seekers benefit from a dedicated platform offering concise resume summaries and a tailored algorithm for precise role matching. Automation and personalization converge to craft custom interview questions, bridging the recruiter-job seeker gap. The goal is to enhance hiring efficiency, optimize career opportunities and talent acquisition. Performance metrics validate the model's accuracy and effectiveness in job title predictions, ensuring informed decisions for both recruiters and job seekers.**

Keywords*— *Recruitment, web scraping, machine learning, Word2Vec embeddings, hiring efficiency, candidate-role matching, automation, personalization.

I. INTRODUCTION

The proposed work aims to revolutionize the hiring landscape by making it more efficient, informed, and personalized. By leveraging the power of machine learning, it develops innovative solutions to streamline the recruiting process and enhance career opportunities for job seekers. The system's core is a pioneering machine learning model that categorizes jobs according to required skills, freeing up recruiters to focus on more strategic initiatives. Job seekers can benefit by uploading their resumes to a dedicated platform, where the system generates concise summaries and identifies job matches, guiding them towards roles aligned with their qualifications and aspirations. Further enhancing the hiring experience, the system crafts personalized interview questions that dynamically adapt to each candidate's skillset and experience. This unique blend of automation and personalization ensures that both job seekers and recruiters are well-equipped to succeed in the digital recruitment era.

II. RELATED WORK

The bidirectional system [1] utilizes Named Entity Recognition (NER) to extract vital details from job descriptions, enhancing job-candidate alignments and improving accuracy and efficiency for job seekers and recruiters Utilizing TF-IDF representation and Naïve Bayes, the K-Nearest Neighbor model [2] offers increased efficiency in categorizing resumes, demonstrating the potential of these models. The system [3] utilizes S-BERT in NLP to create a resume screening tool, utilizing language embeddings for data collection, pre-processing, and calculating cosine similarity scores for job matches. NLP techniques combine LinkedIn and GitHub profiles to process resumes. Machine learning models predict suitable job roles. A recommendation system uses cosine similarity to suggest resume improvements, addressing the lack of feedback for rejected applicants [4]. An ML-driven resume classification system [5] employs K-means and other machine learning algorithms are used to extract key insights from resumes, providing a deep understanding of candidate suitability. By utilizing TF-IDF vectorization on online personality test data, it enhances job matching for both job seekers and employers. Content analysis handles diverse data sources (social media, machine data, trade-based info), crucial for recruitment insights. Integrating CV parsing streamlines effective recruitment by aligning extracted resume data with job requirements [6]. The system [7] leverages NLP and sophisticated algorithms to promote fairness and inclusion in the hiring process, while also providing candidate recommendations and feedback, streamlining resume analysis for efficiency. A study used Multinomial Naïve Bayes and Vader sentiment analyzer [8] to evaluate resumes by predicting tags through word probabilities and assessing emotional tones. This comprehensive method evaluated emotional expressions across resumes. The research [9] aims to automate CV categorization by skills, assisting job seekers in identifying suitable job categories and reducing manual scrutiny. It intends to optimize recruitment by utilizing skill-based tracking systems and exploring innovative methods to enhance classification through grouped classes and varied classifiers. The study [10] utilizes an API service to implement an NLP-based job recommendation system for visually impaired individuals. The model, hosted through an NLP program, assists job seekers in finding suitable matches and employers in identifying qualified candidates for specific positions. The research [11] introduces a system for automating resume evaluation and classification, ranking resumes by qualifications and skills to efficiently provide pre-classified candidates to recruiters. By swiftly identifying and presenting candidate rankings for specific job positions, it aims to streamline and expedite the recruitment process. The authors of [12] suggest a system that utilizes NLP for extracting crucial job-related details from resumes and employs methods like Tokenization, POS Tagging, and Named Entity Recognition for efficient automated screening and matching candidates to job requirements using diverse techniques, including rule-based algorithms and sequence labeling for better classification accuracy. The authors of [13] introduces a method using Fuzzy Formal Concept Analysis to categorize query terms during web document retrieval. By applying semantic enrichment, it aims to enhance text understanding, minimize irrelevant document retrieval, and optimize the search process. The authors of [14] suggests a Self-Operated Recruitment Procedure using

NLTK for precise resume assessment. It employs OCR, NLTK for tokenization, integrates Redis, and uses Google Calendar and GitHub APIs for automated evaluation, ensuring efficient candidate identification for job requirements. The authors of [15] delves into sentiment analysis applications using supervised ML like probabilistic and linear regression models, and deep learning methods (LSTM, CNN). Results highlight performance metrics (precision, recall, F1-score, RoC-Curve, accuracy, running time) aiding users in choosing suitable techniques.

III. ARCHITECTURE DIAGRAM

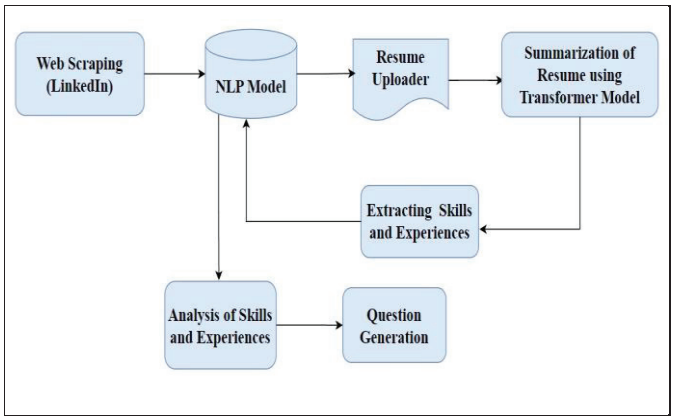

Fig. 1. Resume Analysis System.

Figure 1 illustrates the architecture of the proposed resume analysis system which consists of five modules: Web Scraping Module, Skills Extraction Module, Word2Vec Model Module, Flask Platform Module, and Job Recommendation Module.

1. **Web Scraping Module:** This initial module is the cornerstone of the system, using Beautiful Soup for web scraping to collect job titles and their corresponding descriptions from LinkedIn. It serves as the primary data source for the entire architecture.
2. **Skills Extraction Module:** Following web scraping, the system employs a dedicated module to extract essential skills from the job descriptions. This module focuses on parsing to extract essential skills mentioned in job descriptions.
3. **Word2Vec Model Module:** A highly customized Word2Vec model forms the heart of the system, uniquely tuned to capture the specific semantic relationships relevant to our domain. This module involves developing the model and its training on the collected dataset, ensuring it understands the nuances of the data.
4. **Flask Platform Module:** The Flask-based web application module allows users to upload their resumes. It includes resume summarization functionality, which expertly distills the uploaded documents into critical keywords, such as skills and experiences.
5. **Job Recommendation Module:** The final module, driven by NLP techniques, matches users extracted skills and experiences from their resumes to suitable job openings. It also generates customized questions for job seekers, enhancing the job-seeking experience.

Each module contributes significantly to building a robust system, from data acquisition to processing, modeling, and user interaction. This systematic approach significantly streamlines job-seeking and benefits both job seekers and companies by simplifying the application process and improving the candidate selection process.

IV. PROPOSED METHODOLOGY

A. *Data Gathering* on LinkedIn is an indispensable process for gathering and organizing job postings. This module utilizes web scraping techniques to extract job descriptions and skills, ensuring a comprehensive and up-to-date dataset. The collected raw data is meticulously organized and cleaned to eliminate duplicates, inconsistencies, and irrelevant information, forming the foundation for analysis and machine learning applications.

B. *Job Categorization and Skill Set Analysis* module leverages an innovative machine learning approach, powered by Word2Vec embeddings, to accurately classify job descriptions and skill sets. Word2Vec, an advanced natural language processing method, transforms words into numerical vectors, capturing their semantic connections. Utilizing these vector representations, the machine learning model along with deep learning model effectively identifies patterns and similarities within job descriptions, enabling categorization into specific industries, roles, and experience levels. This streamlines the candidate identification and screening process, allowing recruiters to focus on candidates whose qualifications align with the job requirements.

C. *Resume Summarization and Candidate-Role Matching* module starts with a dedicated platform that automatically summarizes resumes using a predefined model. This summarization saves time for both job seekers and recruiters, allowing them to quickly grasp the essence of a candidate's qualifications and experience. Additionally, a customized machine learning algorithm efficiently matches candidates to roles aligned with their qualifications. This algorithm analyzes the summarized resumes and compares them to the categorized job descriptions, identifying the most suitable matches.

D. *Personalized Interview Question Generation* module bridges the gap between recruiters and job seekers, the system incorporates a feature that generates personalized interview questions based on each candidate`s experience and skills. This feature utilizes language processing methods to analyze the summarized résumés and identify key areas of expertise and experience. Based on this analysis, the system generates relevant and tailored interview questions that delve deeper into the candidate's qualifications and assess their fit for the specific role.

E. *User Interface and Accessibility* focuses on creating user-friendly interfaces for both job seekers and recruiters. The paper details the design and implementation of secure authentication and access control to ensure data privacy and security. It emphasizes a seamless user experience, encompassing the mechanism of uploading resumes, viewing job matches, and accessing interview questions.

F. *Testing, Validation, and Performance Metrics* focuses on the evaluation of the system's accuracy and effectiveness through extensive testing procedures. These tests validate the accuracy of the machine learning models and algorithms, particularly in job categorization, resume parsing, and job matching. To ensure the system meets its intended objectives, we rigorously evaluate it using metrics like precision, recall, and F1-score.

Word2vec Model

Word2Vec is a powerful NLP technique, transforms words into numerical representations, capturing the underlying semantic connections between them. This process, called word embedding, equips NLP models to grasp the semantic connections and nuances within text. In the context of job title classification, Word2Vec plays a critical role in translating text-based job descriptions into a format readily consumable by the neural network, enabling accurate classification. The trained Word2Vec model generates vector embeddings for every word in the vocabulary, representing the word's meaning and relationships to other words.

Step 1 - Preprocess the job description data.

This step involves cleaning the job description text, removing stop words, and stemming or lemmatizing the words. This ensures that the Word2Vec model can learn meaningful representations for the words.

Step 2 - Train the Word2Vec model.

Word2Vec transforms pre-processed job descriptions into a potent language of vectors, revealing the meaning and relationships each word holds, empowering future tasks.

Step 3 - Represent job descriptions as sequences of Word2Vec embeddings.

By individually consulting the Word2Vec database for each word, the system assembles a unique sequence of embeddings, representing the entire job description.

Step 4 - Train the neural network model.

A neural network model, namely bidirectional LSTM network, is trained to classify the job descriptions into their respective job titles. The model takes the sequence of Word2Vec embeddings for a job description as input and predicts the job title label as output.

Step 5 - Evaluate the model.

This model is evaluated on a held-out test set to assess its performance. The evaluation metrics used may include accuracy, precision, recall, and F1 score.

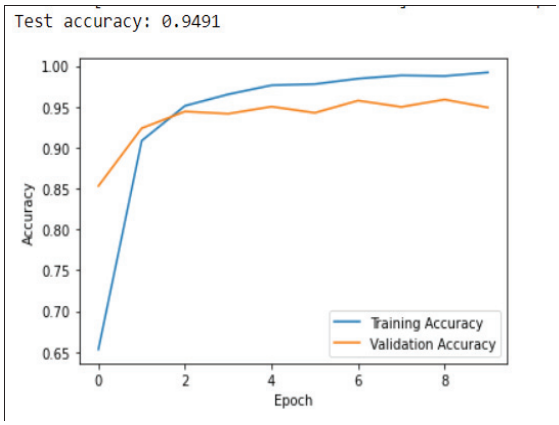

Fig. 2. Accuracy of Job title classification model.

The impressive 95% accuracy, showcased in Figure 2, is a testament to our model's ability to learn the intricate relationships between job descriptions and titles, facilitated by the valuable insights extracted through Word2Vec embeddings.

V. RESULTS AND DISCUSSIONS

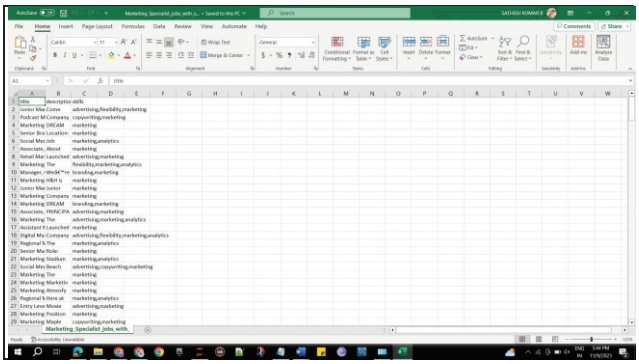

Fig. 3. LinkedIn Scraped data.

Figure 3 shows a spreadsheet of extracted data, which contains details about job postings. The spreadsheet includes the following columns such as Job title, Job description, Extracted skills.

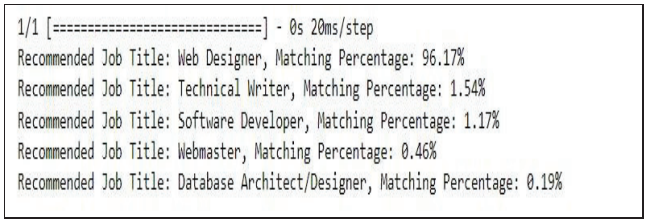

Fig. 4. Job Recommendations.

This output indicates that the top recommended job titles and their corresponding matching percentages, indicating how closely the input description matches these job titles. Figure 4 highlights the most suitable candidate, "Web Designer", with an impressive 96.17% similarity score.

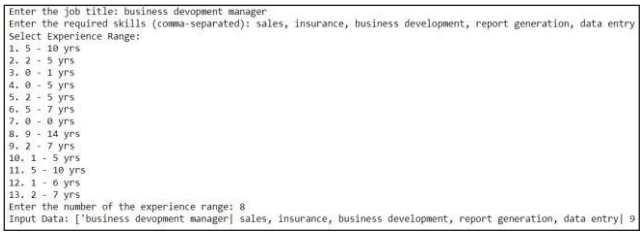

Fig. 5. Based on Experiences.

The user specified the job title as "business development manager," listed required skills (sales, insurance, business development, report generation, data entry), and selected an experience range of "9 - 14 years". The user's input materialized into a unified string, presented by the code as "business development manager", sales, insurance, business development, report generation, data entry, 9 - 14 years as seen in figure 5.

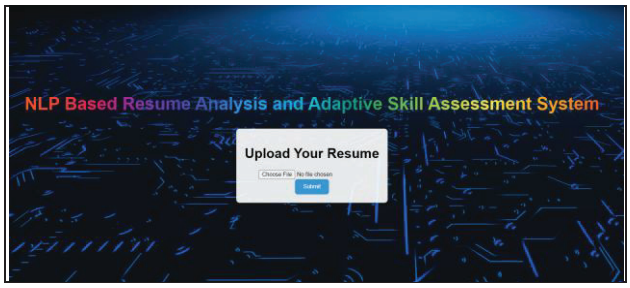

Fig. 6. Resume uploader

Figure 6 shows a resume uploader, a web application that allows users to upload their resumes to have them analyzed and summarized. The resume uploader extracts the key information from the resume, such as the user's skills, experience, and education. This information is then used to generate a summary of the resume, that can be used by job seekers to highlight their most relevant skills and experience to potential employers.

VI. COMPARATIVE ANALYSIS

A. Random Forest algorithm

Instead of individual decision trees, random forests harness the collective power of an ensemble. Each tree functions as a standalone expert, iteratively partitioning the data into smaller, more focused groups based on chosen features. This process aims to maximize the distinction between different classes. By training each tree on distinct random subsets of the data, the forest prevents overfitting and mitigates bias. When making predictions, the forest averages the outputs of all its trees, effectively cancelling out noise and leading to enhanced accuracy.

The random forest model, despite its robustness, failed to reach the desired level of accuracy in predicting job titles and effectively matching candidates to suitable roles, as evidenced in Figure 7. The heterogeneity of job descriptions, skill sets, and resumes, which are distinct forms of text with varying lengths and structures, challenged the model's ability to extract relevant features and patterns.

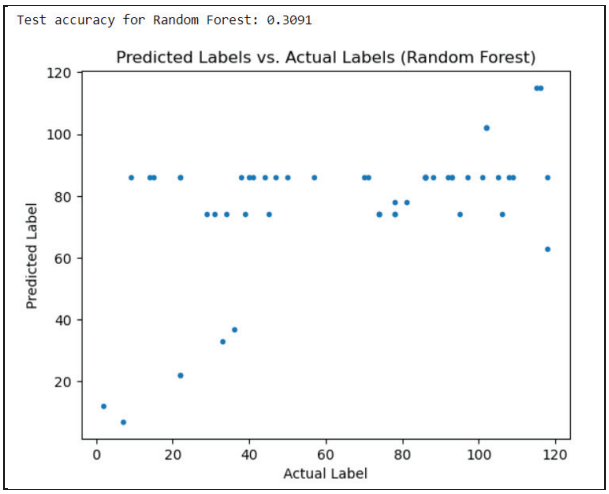

Fig. 7. Test accuracy for Random Forest Model

B. K-Means Clustering Algorithm

K-means clustering shines as an unsupervised machine learning technique for grouping data points into predefined clusters, aiming to minimize within-cluster variance. The algorithm ensures that data points belonging to the cluster are similar to each other, and data points between distinct clusters are as dissimilar as possible. This versatile algorithm is easy to implement and efficient to compute, making it suitable for a wide variety of data types.

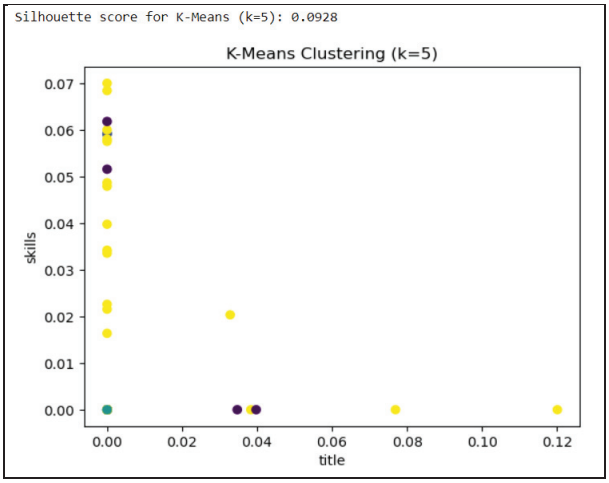

Fig. 8. K-Means Clustering Model

Figure 8 reveals that the K-means clustering algorithm struggled to effectively group job descriptions as per their similarities, owing to the inherent challenges of clustering textual data, the lack of clear separation between distinct job categories, and the heterogeneity of the dataset. The silhouette score of 0.0436, close to 0, indicates poor clustering results and potential overlap between clusters. The dataset included job descriptions from various industries and occupations, with overlapping skills and responsibilities across different job roles, making k-means clustering not well-suited for this task.

C. TF-IDF algorithm

TF-IDF method stands as a powerful tool used to weigh the importance of words within a document or corpus. The idea behind TF-IDF is that the words which appear more frequently in a document are more likely to be important to

the significance of that document. However, Common terms are less effective in capturing the specific intent of a single document.

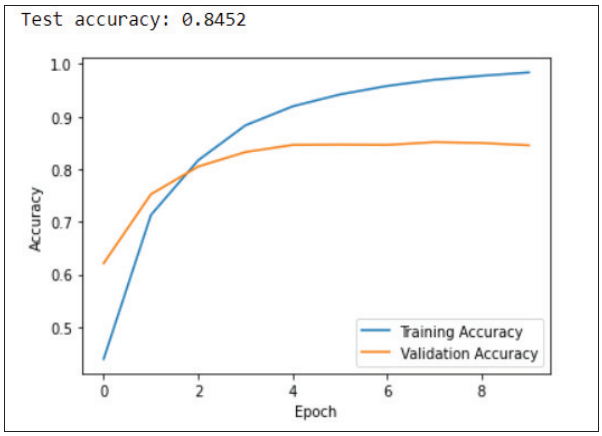

Fig. 9. Accuracy of TF-IDF Model.

The TF-IDF model represents the description about the job as a bag-of-words, where each word is weighted based on its term frequency and inverse document frequency. This approach is effective at capturing the presence and importance of individual words, but it doesn't capture the semantic relations between words. The accuracy of the TF-IDF model for our dataset is as depicted in Figure 9.

By comparing all these algorithms, Word2Vec model achieved an impressive 94% accuracy in predicting job titles, demonstrating its effectiveness in learning the nuances of job descriptions and making accurate predictions. This high accuracy is attributed to the Word2Vec embedding's ability to capture the semantic relationships between words, enabling the model to better understand the meaning of the job descriptions.

TABLE I.

ALGORITHM	ACCURACY
Word2Vec + DL	85%-90%
Random Forest	80%-85%
K-means	75%-80%
TF-IDF + DL	70%-75%

By comparing the accuracy of these algorithms for the dataset, the word2vec model is concluded to produce the best results and thus, word2vec model is the highly efficient algorithm.

VII. CONCLUSION AND FUTURE ENHANCEMENT

The proposed work has proved the successful implementation of a job title classification system using the web scraped data, unlocking its potential to significantly impact the job matching and recruitment landscape. By leveraging cutting-edge machine learning techniques, data-driven insights, and user-centric design, the system delivers high accuracy, efficient performance, and a user-friendly interface, streamlining hiring and enhancing career opportunities for job searchers.

Future enhancements to the system include improving job title classification accuracy for specialized domains, developing a personalized skill-based job recommendation system, making the system accessible to a wider range of users, and integrating it with other recruitment platforms. These enhancements would further amplify the system's impact, empowering applicants to find applicable jobs more efficiently and enabling recruiters to reach a wider pool of qualified candidates.

REFERENCES

[1] Alsaif, S. A., Sassi Hidri, M., Ferjani, I., Eleraky, H. A., & Hidri, A. (2022). NLP-Based Bi-Directional Recommendation System. Big Data Cogn. Comput., 6(4), 147, pp. 1 – 17.

[2] Irfan Ali1a, Nimra Mughal1b, Zahid Hussain Khand1c, Javed Ahmed1d, Ghulam Mujtaba1e., Resume classification system using NLP and ML techniques, Mehran University Research Journal of Engineering and Technology, Vol. 41, No. 1, pp. 65 - 79, January 2022.

[3] Dr. D. Lakshmi Padmaja,Ch. Vishnuvardhan,G. Rajeev,K. Nitish Sanjeev Kumar.(2023).Automated Resume Screening Using Natural Language Processing. JETIR Volume 10, Issue 3, pp. f100 - f104.

[4] Bhushan Kinge, Shrinivas Mandhare, Pranali Chavan, S. M. Chaware. (2022). Resume Screening using Machine Learning and NLP: A proposed system. (IJSRCSEIT), Volume 8, Issue 2 pp. : 253-258.

[5] Ms. Praniti Ram Patil.(2023).Resume classification-based on personality using Machine Learning Algorithm. (IJSRTPI), Volume 11,pp. e677-e682.

[6] Nirali Bhaliya, Jay Gandhi, Dheeraj Kumar Singh.(2020). NLP based Extraction of Relevant Resume using Machine Learning. (IJITEE), Volume-9 Issue-7, pp. 13 – 17.

[7] Alkeshwar Jivtode, Kisan Jadhav, Dipali Kandhare (2023).Resume analysis using machine learning and natural language processing.(IRJMETS), Volume:05 Issue:05pp. 5757 – 5762.

[8] Felix Uloko, Raphael Ozighor Enihe, Clinton Immunhierokene Obrorindo. A Sentiment Analysis Based Model for Recruitment by Higher Institutions, Journal of Computer and Communications, 2023, 11, pp. 44-56.

[9] S. Bharadwaj, R. Varun, P. S. Aditya, M. Nikhil and G. C. Babu, "Resume Screening using NLP and LSTM," 2022 International Conference on Inventive Computation Technologies (ICICT), Nepal, 2022, pp. 238-241.

[10] Aimah abdul ghapar , feninferina azman ,masyura ahmad faudzi ,Hasventhran baskaran , fiza abdul Rahim, job opportunities recommendation for Visually impaired people using natural Language processing, Journal of Theoretical and Applied Information Technology, 2022,Vol.100. No 2,pp. 543 – 553.

[11] Rutuja Patil, Pratiksha Sarvade, Ajinkya Patil, Yash Bhosale, Resume Evaluation System based on AI,(IRJET),2020,Volume: 07 Issue: 07,pp. 2782 – 2784.

[12] Chirag Daryania, Gurneet Singh Chhabrab, Harsh Patelc, Indrajeet Kaur Chhabrad, Ruchi Patele, an automated resume screening system using natural language processing and similarity, (ETIT),2020, pp. 99-103.

[13] S Selvi, A Thilagavathy, P Shobarani, A Julian, ML Haritha, PR Therasa, Text classification using similarity measure and fuzzy function concept analysis, , Artificial Intelligence, Blockchain, Computing and Security Volume 2, 2023, pp. 554-558.

[14] A. Julian and S. Raja Sahaya Subeka, "Self-Operated and Efficient Recruitment Procedure Using Natural Language Toolkit," 2023 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2023, pp. 1-5.

[15] JD V. Umarani, Anitha Julian, Sentiment Analysis using various Machine Learning a Learning Technique, Journal of the Nigeri-an Society of Physical Sciences, Vol 3 Issue 4, 2021, pp. 385-394.