Grenze International Journal of Engineering and Technology, Jan Issue

Revolutionizing Recruitment Strategy: A Multi-faceted Approach to Resume Screening using NLP and Information Retrieval Techniques

Dr. Velvadivu P[1], Dr. Sujithra M[2], Preethi P[3], Dharshini M[4] and Priyadharshini P[5]
[1-2]Assistant Professor, Department of Computing, Coimbatore Institute of Technology, Tamil Nadu, India.
Email: {velvadivu, sujithra}@cit.edu.in
[3-5]M.Sc Data Science, Department of Computing, Coimbatore Institute of Technology, Tamil Nadu, India.
Email: {preethi8abi, shivani180301, devidharshini001@gmail.com

***Abstract*— In today's competitive job market, effective candidate evaluation is paramount for an organization to identify the most suitable candidate. This research study examines automated resume parsing and evaluation, offering a thorough methodology to speed up the hiring process. The research focuses on the automatic extraction of essential resume components, such as contact details (phone numbers, Gmail IDs), professional networking accounts (LinkedIn, GitHub), academic standing (CGPA), and abilities. Carefully extract and organize these crucial data points using a combination of regular expressions, natural language processing libraries (Spacy, NLTK), and PDF parsing tools (PDF Miner). Used mathematical methods for resume scoring to evaluate candidate fitness and to facilitate effective candidate ranking. In an era of digital transformation and fierce competition, this research provides practical, scalable ways to automate and enhance candidate evaluation, empowering recruiters to make better hiring decisions. This paper is an important step towards modernizing recruitment processes, improving candidate assessment, and fostering superior talent acquisition in an ever-changing job market.**

***Index Terms*— Regex, Natural Language Processing, Resume Scoring, Sentiment Analysis, Information Retrieval.**

I. INTRODUCTION

In today's competitive job market, where companies are constantly looking for the best candidates to add to their teams, candidate evaluation has become a critical and time-consuming process. Recruiters and hiring managers must sift through hundreds of resumes every day, each with its own unique set of skills and experience. That is why the project dives deep into the world of automated resume sorting and evaluation, providing a sophisticated and streamlined solution to help streamline the recruitment process.

This project is all about automating the process of extracting and analyzing key resume elements. These include important information like contact information (phone number and Gmail ID), professional profile information (LinkedIn and GitHub), academic achievements (CGPA), and the skills that need to prove their qualifications. Hence used a mix of regular expression, natural language processing (NLT) libraries like Spacy, and PDF parsing (PDF Miner) tools to extract and organize these important data points with precision.

In addition, this project adds a new element to resume analysis– sentiment analysis. This new approach looks for

Grenze ID: 01.GIJET.10.1.518

extracurricular activity mentioned within resumes to get a better understanding of candidates' personalities and interests, expanding the scope of the assessment beyond just formal qualifications. Used a powerful set of math calculations to get resume scores, which allows to rank candidates quickly and easily without the need for complex machine learning models. In today's highly competitive job market, where digital transformation is rapidly changing the landscape, this project provides practical and scalable solutions to automate and improve candidate evaluation. By giving recruiters a more complete toolkit for evaluating job applicants, helps to modernize recruitment processes and empower hiring professionals to make informed decisions in today's ever-changing job markets.

II. Literature Review

Resume screening is a critical process in the recruitment and selection of candidates for job positions in organizations. It plays a pivotal role in identifying qualified candidates and streamlining the hiring process. This literature review explores key research findings and trends in the field of resume screening, shedding light on the various methods and challenges associated with this HR function.

The practice of resume screening has evolved significantly over the years. Early approaches involved manual review of printed resumes, while modern techniques leverage technology and automation. Studies by Breaugh (2008) and Marler and Boudreau (2017) highlight the historical development of resume screening and its shift towards data-driven decision-making.

The advent of Applicant Tracking Systems (ATS) has transformed the way organizations process resumes. Research by Van den Heuvel and Bondarouk (2017) underscores the increasing prevalence of ATS and their impact on resume screening efficiency and accuracy. The effectiveness and limitations of ATS algorithms in matching candidates to job requirements have been discussed extensively in the literature.

Advances in machine learning and natural language processing (NLP) have enabled more sophisticated resume screening techniques. Studies by Narayanan and McDonald (2017) and Sheng et al. (2019) discuss the application of these technologies in improving the accuracy of candidate selection and the challenges associated with fine-tuning machine learning models for this purpose.

While technology plays a pivotal role, human involvement in resume screening remains crucial. Research by Cascio and Aguinis (2008) and Tippins and Bauer (2019) emphasizes the importance of a hybrid approach that combines the strengths of both automated systems and human judgment in achieving optimal hiring outcomes.

Resume screening is a dynamic field that continues to evolve with advancements in technology and changing societal expectations. This literature review highlights the historical context, challenges, and opportunities associated with resume screening in contemporary recruitment. As organizations strive to make informed hiring decisions while addressing bias concerns, a holistic approach that combines automation, human judgment, and ethical considerations emerges as a key theme in the literature.

III. Information Retrieval

Information retrieval is a fundamental field of study that lies at the crossroads of computer science and informatics. It is responsible for the systematic retrieval and presentation of relevant information from large and often unmanaged data sets. This multi-faceted area includes a wide range of techniques and approaches designed to maximize the effectiveness and efficiency of finding the needed information within large and diverse data sets.

At its most basic, information retrieval is concerned with meeting specific information requirements. It accomplishes this by matching user inquiries or inquiries with relevant documents or information, aiding in the acquisition of knowledge and providing answers to queries. Information retrieval is used in various applications, such as web search engines and digital libraries, as well as content recommendation systems and database queries.

There are several key elements that go into the process of information retrieval. It all starts with assembling a large document collection. This serves as the basis for subsequent retrieval operations, storing text, multimedia, and structured data. User inquiries are at the heart of information retrieval. They come in a variety of forms, from basic keyword searches to complex Boolean statements or natural language queries. These queries dictate the direction in which the information is to be sought.

An essential part of information retrieval is indexing. This is the process of creating a structured, searchable database that links terms, phrases, or metadata in documents to their respective locations. Indexing improves search speed and accuracy by making it easier for users to match their queries with indexed documents. A variety of search algorithms and models are used to match user queries to indexed documents. These include TF-IDF (Term Frequency Inverse Document Frequency), vector space models and machine learning based approaches. Once

indexed, documents are ranked according to how relevant they are to the user's query. The top results are presented in an easy-to-read format, allowing users to quickly access the most important information. The information retrieval process starts with assembling a large collection of documents. This serves as the basis for subsequent retrieval operations, storing text, multimedia content and structured datasets. User queries are a key part of information retrieval. They come in many forms, such as simple keyword searches, complex Boolean expressions, natural language questions, and more.

Information retrieval technologies provide recruiters more control by enabling effective search and filtering. Employing sophisticated search algorithms and keyword-based criteria, recruiters can easily query the database to find people that fulfil the job specifications. Recruiters can prioritize candidates who satisfy the prerequisites by implementing automated filtering methods to pre-screen resumes. Additionally, these systems' analytics and reporting capabilities provide insightful data on candidate demographics, emerging skill patterns, and recruitment process effectiveness. Organizations may improve their talent acquisition strategies, make better decisions, and ultimately find the greatest individuals for their teams by using a data-driven strategy, which also increases recruiting efficiency and effectiveness.

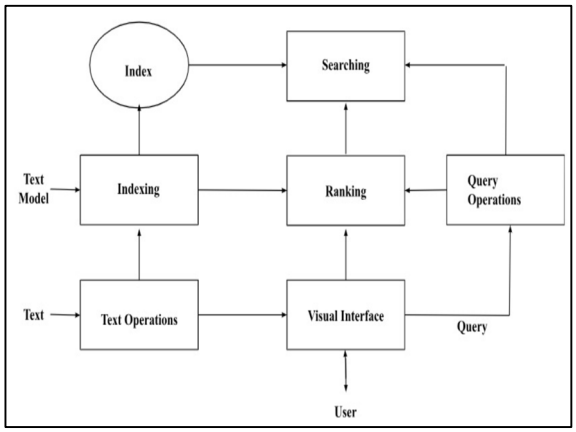

Figure 1. Information Retrieval Workflow

IV. METHODOLOGY

A. Techniques used in Information Retrieval:

1. Indexing

Indexing is a fundamental method in information retrieval systems, serving as a structured cataloging process for efficient data access. It involves the systematic organization and storage of keywords, phrases, or metadata from documents, making it possible to swiftly locate and retrieve specific information. Indexing plays a pivotal role in enhancing the speed and effectiveness of search operations across various domains, including databases and search engines.

2. Lemmatization

Lemmatization is a text normalization method that differs from stemming in its approach. It utilizes dictionaries or vocabularies to map words to their canonical or dictionary form, referred to as the lemma. This meticulous process ensures that the resulting words are valid words in the language, which is particularly critical for maintaining linguistic accuracy and coherence in various text analysis tasks.

3. Spelling Correction

Spelling correction is a vital process for automatically identifying and rectifying typographical errors or misspelled words in textual data. Employing algorithms and language dictionaries, this methodology suggests or applies corrections to words that deviate from recognized language standards. It substantially enhances text quality, readability, and the overall precision of language processing tasks.

4. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is an advanced natural language processing technique for assessing and characterizing the emotional tone or sentiment conveyed within text, such as reviews, comments, or social media posts. It entails the classification of sentiment as positive, negative, or neutral, offering valuable

insights into public sentiment, customer feedback, and brand reputation. Sentiment analysis methodologies often leverage machine learning models and lexicon-based approaches to systematically categorize and quantify sentiments present in textual data.

5. Inverted Indexing

Inverted indexing is a fundamental data structuring technique widely used in information retrieval systems and search engines. It revolves around the creation of an index for a collection of documents, such as web pages, books, or articles. The key idea is to organize terms or words within these documents and then associate them with lists of document IDs or positions where they occur. This inverted structure not only accelerates search operations but also allows for advanced functionalities like ranking and relevance scoring, making it a cornerstone in modern information retrieval. It enables users to find relevant information quickly and efficiently by looking up terms in the index and retrieving a list of documents containing those terms, significantly improving search performance and user experience.

B. Packages used for Information Retrieval:

1. Regex

The 'regex' package, short for regular expressions, is a powerful tool for pattern matching and text manipulation. It provides a concise and flexible syntax for searching, extracting, and manipulating strings based on specific patterns or rules. Regular expressions are indispensable for tasks like data validation, text parsing, and pattern matching in various programming languages.

2. PDF-Miner

'PDF-Miner' is a Python library used for extracting text and metadata from PDF documents. It offers a programmatic way to access and analyze content within PDF files, making it valuable for tasks like data extraction, text mining, and content indexing from PDF documents.

3. spaCy

'spaCy' is a leading natural language processing library that offers efficient and accurate text processing capabilities. It provides pre-trained models for various languages and supports tasks such as tokenization, part-of-speech tagging, named entity recognition, and dependency parsing. `spaCy` is widely used in NLP research and applications due to its speed and accuracy.

4. Whoosh

'Whoosh' is a fast and feature-rich search engine library for Python. It is used for creating custom search engines and full-text search functionality in applications. `Whoosh` is known for its simplicity, scalability, and flexibility, making it a popular choice for building search solutions.

5. TextBlob

'TextBlob' is a simple and user-friendly Python library for processing textual data. It offers a range of text analysis functionalities, including part-of-speech tagging, sentiment analysis, translation, and text classification. It's particularly valuable for quick and easy text processing tasks, especially for those new to NLP.

6. NLTK (Natural Language Toolkit)

The `NLTK` library is a comprehensive platform for natural language processing and text analysis in Python. It includes a wide array of resources, including corpora, lexicons, and pre-trained models, along with various tools and algorithms for tasks like tokenization, stemming, and machine learning-based NLP.

7. Pandas

'Pandas' is a powerful data manipulation and analysis library for Python. It provides data structures like Data Frames and Series, along with a wealth of functions for data cleaning, transformation, and analysis. `Pandas` is widely used in data science.

C. Content Extraction using Regex

1. Phone Number

Regex pattern: "\+91\s?-?\d{2,5}\s?\d{2,5}\s?\d{4,}"

This Regex pattern extracts Indian phone numbers with the country code '+91,' accommodating variations in formatting by allowing optional spaces and hyphens between digits. It captures numbers with 2 to 5 digits for the area code and 2 to 5 digits for the local number, ensuring a minimum of 4 digits in the phone number for accurate extraction.

2. CGPA

Regex pattern: ": \d+\.\d+"

To extract Cumulative Grade Point Averages (CGPA) represented as decimal numbers, this simple Regex pattern is employed. It searches for numeric values with at least one digit before and after a decimal point. This pattern ensures that it accurately captures CGPAs presented in the standard numeric format, such as "3.75" or "8.5," by matching any sequence of digits followed by a decimal point and another sequence of digits.

3. Gmail ID

Regex pattern: "[\w\.-]+@[\w\.-]+"

This Regex pattern is designed to find and extract Gmail email addresses from text documents. It identifies Gmail IDs that consist of alphanumeric characters, dots, and hyphens before the "@" symbol, followed by the domain name. The pattern matches characters before "@" and in the domain part, allowing for accurate identification of Gmail addresses in various formats.

4. GitHub ID and LinkedIn ID:

Regex Pattern for GitHub ID: "https?://(?:www\.)?github\.com/[A-Za-z0-9_.-]+"

Regex Pattern for LinkedIn ID: "r'\S*linkedin\S*'"

These Regex patterns are designed for extracting GitHub profile links or IDs and LinkedIn IDs from text. The GitHub pattern accommodates various URL formats, including "http" and "https" protocols and optional "www" subdomains, while focusing on capturing alphanumeric characters, dots, underscores, or hyphens representing GitHub usernames. On the other hand, the LinkedIn ID pattern identifies the presence of the word "LinkedIn" within the text and captures non-whitespace characters before and after it, ensuring effective extraction in both scenarios.

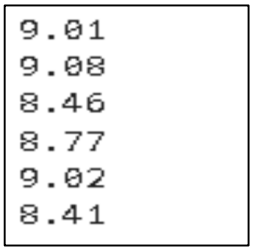

Figure 2. CGPA Extraction

Figure 3. Gmail ID Retrieval

Figure 4. Github Link Retrieval

D. Skills Extraction from Resume using spaCy

The process of extracting both soft skills and technical skills from resumes using the SpaCy package involves several key steps. The resume text is subjected to tokenization and analysis. For soft skills, the extraction process relies on contextual analysis and predefined patterns, identifying terms or phrases related to interpersonal abilities, communication, teamwork, and leadership. Technical skills are extracted using specific patterns that capture programming languages, tools, and domain-specific knowledge. These identified soft and technical skills are subsequently extracted separately, providing a comprehensive overview of a candidate's skillset, and aiding in effective talent assessment during the recruitment process.

E. Spelling Correction using TextBlob

Using the TextBlob package to correct spelling errors in the extracted skills and subsequently storing the corrected skills in a database is a practical and systematic approach to maintain a clean and reliable skills dataset. After extracting skills from resumes using SpaCy, Text Blob's spell-check and correction capabilities ensure that accuracy and the corrected skills are stored in database. This process enhances the quality and accuracy of the

skills data, which is essential for efficient talent assessment, resume screening, and matching candidates to job requirements during recruitment and talent management activities.

F. Document Inverted Indexing

It demonstrates its functionality by building an inverted index from 'all text' (assuming it contains text data) and searching for the term "HTML." Matching documents are displayed with their respective IDs, highlighting the core process of inverted indexing for efficient information retrieval.

G. Extracurricular activities using sentiment analysis

The process involves analyzing the sentiment of extracurricular activities in resumes. It begins by breaking down each activity description into sentences and words, lemmatizing words to their verb form for sentiment analysis. The sentiment polarity of each word is calculated and aggregated to determine the overall sentiment score for the activity. This score is then categorized as "Positive," "Negative," or "Neutral" based on predefined thresholds, providing insight into the emotional tone associated with each extracurricular activity.

H. Resume Scoring

Resume scores are determined by normalizing parameters like CGPA and creating binary indicators for factors like GitHub and LinkedIn presence and soft skills availability. Each parameter is then assigned a weight reflecting its importance, and individual scores are calculated by multiplying the parameter value by its weight. The overall resume score is obtained by summing up these individual scores, allowing for effective candidate ranking and selection based on academic performance, online presence, skills, and soft skills for recruitment or talent assessment purposes.

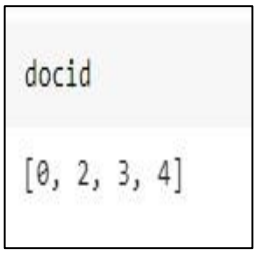

Fig 5. Document Indexing

Figure 6. Skills Extraction

V. RESULT

The extracted contents like contact number, email id, github link, linkedin link, skills, soft skills, CGPA and the scores. The overall score is calculated based on individual scores like github score, linkedin score, CGPA score, soft skills score and skills score.

Id	Contact	Email	CGPA	Github	LinkedIn	Skills
1	+919835754326	mdharshini1	9.01	https://githul	https://wv	Python
2	+918453578564	keerthana30	9.08	Not available	https://wv	Kubernete
3	+916754758465	nidhhi01@gr	8.46	Not available	https://wv	Python,SQ
4	+916478590321	preethi.p080	8.77	https://githul	https://wv	Statistics,P
5	+917563658485	priyadharshi	9.02	https://githul	https://wv	R,SQL,C++,
6	+919345678923	gssoundarya	8.41	Not available	https://wv	R,Python,S

Soft skills	CGPA_Score	GitHub_Score	LinkedIn_Score
Not available	0.2703	0.2	0.2
Leadership	0.2724	0	0.2
Not available	0.2538	0	0.2
Presentation	0.2631	0.2	0.2
Presentation	0.2706	0.2	0.2
Communicatio	0.2523	0	0.2

Figure 7. Result set stored in Excel File

Skills_Score	Soft_Skills_Score	Overall_Score
0.04	0	0.7103
0.2	0.1	0.7724
0.2	0	0.6538
0.12	0.1	0.8831
0.24	0.1	1.0106
0.16	0.1	0.7123

Figure 7. Result set stored in Excel File

VI. CONCLUSION

This study introduces an efficient approach to parsing and evaluating resumes in the ever-changing recruitment landscape. It utilizes a carefully chosen toolkit, including regular expressions for precise data extraction, spaCy for advanced natural language processing, PDF-Miner for accurate PDF analysis, and NLTK for effective text preparation. The framework systematically extracts critical resume components and integrates sentiment analysis, providing a comprehensive view of candidates beyond their academic and professional qualifications. Importantly, it prioritizes transparency with clear mathematical computations for resume scoring, ensuring fairness and accessibility in the recruitment process. This research bridges the gap between technology and human expertise, highlighting the role of technological innovation in informed talent acquisition decisions, benefiting both organizations and individuals in today's dynamic job market.

REFERENCES

[1] Sujithra, M., P. Velvadivu, J. Rathika, R. Priyadharshini, and P. Preethi. "A Study on Psychological Stress of Working Women in Educational Institution Using Machine Learning." In 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-7. IEEE, 2022.

[2] A Kmail, M Maree, M Belkhatir, and S Alhashmi "An Automatic Online Recruitment System based on Exploiting Multiple Semantic Resources and Concept-relatedness Measures," Proceedings of the EEE 27th International Conference on Tools with Artificial Intelligence (ICTAI) pp. 620-627, 2015.

[3] Sujithra, M., and G. Padmavathi. "Ensuring Security on mobile device data with two phase RSA algorithm over cloud storage." Journal of Theoretical and Applied Information Technology 80, no. 2 (2015): 221.

[4] C Hauff G Gousios, "Matching GitHub developer profiles to job advertisements" Proceedings of the 12th Working Conf. on Mining Software Repositories, pp. 362-366, 2015.

[5] T Schmitt, P Caillou, M Sebag, "Matching Jobs and Resumes: a Deep Collaborative Filtering Task." Proc. of the 2nd Global Conf. on Artificial Intelligence, pp.1-14, 2016.

[6] Kaviya, K., K. K. Shanthini, and M. Sujithra. "Micro-blogging sentimental analysis on Twitter data using Naïve Bayes machine learning algorithm in Python." Int J Math Comput Sci 4 (2018).

[7] S ALOtaibi and M Ykhlef, "Tob Recommendation Systems for Enhancing E-recruitment Process", in Proceedings of the International Conference on Information and Knowledge Engineering (IKE), Las Vegas Nevada, USA, pp. 433-439, 2012.

[8] The International Association of Employment Web Sites (IAEWS), available from: http://www icmaonline org international-association-of-employment-web-sites, Date Visited: June 20, 2017.

[9] A Kmail M Maree, and M Belkhatir, "Matching Sem: Online recruitment system based on multiple semantic resources," Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, pp. 2654-2659, 2015.

[10] Dhamodaran, B., Siddhesh, S., Suhas, P. S., Sujithra, M., & Velvadivu, P. (2020). Big Data Performance Comparison Over Pyspark Tensorflow and Scikit-Learn. no, 11, 239-242.