RESUME PARSER USING NATURAL LANGUAGE PROCESSING

EC19603 - PROBLEM SOLVING USING AI AND ML TECHNIQUES

Submitted by

AMRITHAA R S	**(2116210801009)**
ANANDHA KRISHNAN S	**(2116210801010)**
ANUBAMA J	**(2116210801013)**

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

CHENNAI – 602105

ANNA UNIVERSITY, CHENNAI 600 025

MAY 2024

ANNA UNIVERSITY : CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report titled **"RESUME PARSER USING NATURAL LANGUAGE PROCESSING"** is the bonafide work of **AMRITHAA R S (2116210801009), ANANDHA KRISHNAN S (2116210801010), ANUBAMA J (2116210801013)** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE
Ms. SUSHMA S. JAGTAP, M.E., (Ph.D.),
SUPERVISOR
ASSISTANT PROFESSOR,
Department of Electronics and Communication Engineering,
Rajalakshmi Engineering College,
Thandalam, Chennai – 602 105.

Submitted to Mini Project Viva-Voce Examination held on ______________

Internal Examiner **External Examiner**

ACKNOWLEDGEMENT

We thank God Almighty for enabling us to complete our project work. Our sincere thanks to our Chairperson **Dr. (Mrs.) Thangam Meganathan**, for her sincere endeavor in educating us in her premier institution. We like to express our deep gratitude to our beloved Vice Chairman **Mr. Abhay Shankar Meganathan, B.E., M.S.** for his enthusiastic motivation which is a lot in completing this project.

We would like to thank our Principal **Dr. S.N. Murugesan, M.E., Ph.D.,** for his kind support and the facilities provided to carry out our work. We would also like to thank our Head of the Department **Dr. L. Bhagyalakshmi, M.E., Ph.D.,** and Academic Head **Dr. S. Chitra, M.E., Ph.D.,** for giving us the opportunity and facilities to complete our work in time.

We take this opportunity to express our profound gratitude and deep regards to our Supervisor **Ms. Sushma S. Jagtap, M.E., (Ph.D.),** Department of Electronics and Communication Engineering for his/her exemplary guidance, monitoring and constant encouragement throughout the course of this project.

We express our gratitude to the project coordinators **Dr. R. Gayathri, Professor/ECE, Dr. A. Asha, Professor/ECE, Ms. T. Helan Vidya, AP(SG)/ECE Ms. Indhumathi, AP/ECE** for their valuable suggestions in doing our project. We also would like to thank our parents and classmates for their moral during the tenure of the project.

ABSTRACT

Job Requirement is considered one of the major activities for humans which is a very strenuous job to find a fruitful talent. Corporate companies and recruitment agencies process numerous resumes daily. Going through one by one resumes is an hectic task to solve this problem we got solution. Our proposed model is basically to extract the details and statistics from the resume and ranking the resume based on the preference of the company associated and its requirements using the Natural Language Processing (NLP) techniques. Parsing and ranking the resume makes the hiring process easy and efficient. A resume contains various minute data within it and any respectable parser needs to extract out these data such as education, experience, project , address etc. So, basically we are going to build a job portal where the employees and applicants would upload their resume for any particular job and using the NLP technique, the necessary information will be parsed and a structured resume with information will be generated and also the resumes of employee will be ranked according to the requirement of the company skill set and employees skills in the provided resume.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVATION

NLP	Natural Language Processing
PDF	Portable Document Format
CV	Curriculum Vitae

CHAPTER 1

INTRODUCTION

Corporate companies and recruitment agencies process numerous resumes daily. This is no task for humans. An automated intelligent system is required which can take out all the vital information from the unstructured resumes and transform all of them to a common structured format which can then be ranked for a specific job position. Parsed information include name, email address, social profiles, personal websites, years of work experience, work experiences, years of education, education experiences, publications, certifications, volunteer experiences, keywords and finally the cluster of the resume (ex: computer science, human resource, etc.). The parsed information is then stored in a database for later use. Unlike other unstructured data (ex: email body, web page contents, etc.), resumes are a bit structured. Information is stored in discrete sets. Each set contains data about the person's contact, work experience or education details. In spite of this, resumes are difficult to parse. This is because they vary in types of information, their order, writing style, etc. Moreover, they can be written in various formats. Some of the common ones include '.txt', '.pdf', '.doc', '.docx', '.odt', '.rtf' etc. To parse the data from different kinds of resumes effectively and efficiently, the model must not rely on the order or type of data.

1.1 NLP ALGORITHM

Natural Language Processing is a field of artificial intelligence in which computers analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

“Apart from common word processor operations that treat text like a mere sequence of symbols, NLP considers the hierarchical structure of language: several words make a phrase, several phrases make a sentence and, ultimately, sentences convey ideas,” John Rehling, an NLP expert at Meltwater Group, says in How Natural Language Processing Helps Uncover Social Media Sentiment. “By analyzing language for its meaning, NLP systems have long filled useful roles, such as correcting grammar, converting speech to text and automatically translating between languages.”

NLP is used to analyze text, allowing machines to Understand how humans speak. This human-computer interaction enables real-world applications like automatic text summarization,sentiment analysis,topic extraction,named entity recognition,parts-of-speech tagging,relationship extraction,stemming, and more. NLP is commonly used for text mining,machine translation and automated questing answering.

1.1.1 Reason behind Implementing NLP

Resumes are full of text data, but that data isn't organized in a neat and easy to search way. This is where Natural Language Processing (NLP) comes in. NLP is a field of computer science that lets computers understand and process human language.

Here's how NLP helps resume parsers:

- **Structured data extraction:** Resumes can be formatted in many different ways. NLP helps parsers understand the meaning behind the text, not just the formatting, so they can find important information like skills, experience, and education regardless of where it's on the page or how it's phrased .
- **Keyword identification:** NLP can find relevant keywords and phrases in a resume, even if they're not worded exactly the way they are in a job description . This helps match qualified candidates to the right jobs.

- **Understanding synonyms and context:** NLP can understand that "managed" and "led" mean basically the same thing, so it can pick up on relevant skills and experience even if they're described in different words .
- **Data analysis:** By turning resumes into structured data, NLP lets recruiters analyze trends in their applicant pool. This can be helpful for things like understanding the skills gap or diversity in applicants .

In short, NLP helps resume parsers make sense of the messy world of resumes and turn them into a valuable source of information for recruiters. This can save recruiters time and effort, and help them find the best candidates for the job.

1.1.2 Work flow of NLP

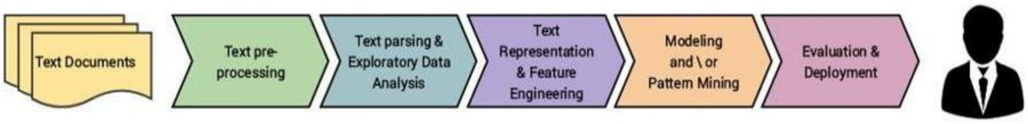

Text Documents
Text pre-processing
Text parsing & Exploratory Data Analysis
Text Representation & Feature Engineering
Modeling and \ or Pattern Mining
Evaluation & Deployment

Figure 1.1 :Generic Workflow of NLP

Natural language processing comprises a wide variety of methods for analyzing human language, based on machine learning techniques as well as rules-based and computational approaches. Tokenization, lemmatization and stemming, parsing, part-of-speech tagging, language identification are some basic NLP tasks. NLP tasks, in general, break downs the language into smaller, essential components, attempt to comprehend links between the pieces, and then examine how the components combine to form meaning .

The procedures given above represent the standard workflow for an NLP task.

The initial stage is generally text wrangling and pre-processing on the collection of documents. Then there's parsing and some basic exploratory data analysis. The next stage is to represent text using word embeddings and then do feature engineering. The final step of any ML workflow is model testing and deployment.

The NLP workflow can be broken down into several key stages:

1. **Text Preprocessing and Wrangling:** This is the first step, where the raw text data is cleaned and formatted for further processing. It involves things like removing punctuation, converting text to lowercase, and fixing any spelling errors.
2. **Tokenization:** Here, the text is broken down into smaller units that the NLP system can understand. These units are typically words, but they could also be phrases or characters depending on the task.
3. **Normalization:** This stage aims to ensure consistency in the data. For example, synonyms might be mapped to a single term, or numbers written out as text might be converted to numerical values.
4. **Part-of-Speech Tagging:** Each word in the text is assigned a label indicating its grammatical function (e.g., noun, verb, adjective). This helps the system understand the relationships between words in a sentence.
5. **Syntactic Parsing:** This stage involves analyzing the sentence structure to understand how the words relate to each other. It reveals the grammatical hierarchy of the sentence, like identifying subjects, verbs, and objects.
6. **Semantic Analysis:** Here, the system goes beyond the grammar and tries to understand the actual meaning of the text. This might involve tasks like identifying the sentiment of the text, recognizing named entities (people, places, organizations), or resolving ambiguities.
7. **Evaluation:** Depending on the specific NLP task, there might be a final evaluation stage where the performance of the system is measured. This helps ensure the system is working as expected and can be further refined if needed.

1.2 MOTIVATION

Imagine a world where sifting through hundreds of resumes becomes a breeze. No more tedious keyword searches or missed gems due to human bias. Here comes the motivation and power of Resume parser using Natural language processing.

1. **Revolutionizing Recruitment:** NLP breathes new life into the recruitment process. It automates the initial screening, extracting key skills and experiences from resumes with impressive accuracy. This translates to significant time savings for recruiters, allowing them to focus on more strategic tasks and candidate interactions.
2. **Objectivity Takes Center Stage:** NLP removes the subjectivity often present in resume reviews. By analyzing skills and experience based on factual information, NLP reduces unconscious bias and ensures qualified candidates aren't overlooked. This fosters a fairer and more diverse talent pool.
3. **Precision Matching:** Gone are the days of generic keyword searches. NLP parsers can understand the nuances of language, accurately matching a candidate's unique skillset to specific job requirements. This leads to a perfect fit between candidates and positions, ultimately boosting employee satisfaction and retention.
4. **Benefits Beyond Efficiency:** The impact of NLP extends beyond recruiters. Job seekers see a positive shift too. Resumes crafted with diverse terminology are now accurately parsed, increasing an applicant's visibility and chances of landing an interview. Additionally, NLP-powered platforms can recommend suitable opportunities based on skills, leading to a more focused and efficient job search.
5. **The Future is Now:** NLP is constantly evolving, and resume parsers powered by this technology will only become more accurate and sophisticated efficiency.

Problem Statement

The ever-growing number of applicants in today's job market creates a significant burden for recruiters. They are tasked with manually sifting through a massive volume of resumes, often in various formats like PDFs, Word documents, or even plain text emails. These resumes hold a wealth of information about a candidate's qualifications, including their skills, experience, and educational background. However, this valuable data is trapped within unstructured text, making it difficult and time-consuming to analyze efficiently.

- **Manual Sorting:** Recruiters would face a mountain of resumes, relying on manual keyword searches and skimming through text to identify qualified candidates. This would be incredibly time-consuming and error-prone.
- **Missed Talent:** Resumes with valuable skills described in non-standard terms or buried within text could be overlooked, leading to qualified candidates being passed over.

1.3 OBJECTIVE

The objective of a resume parser using Natural Language Processing (NLP) is to revolutionize the recruitment process by automating the extraction of key skills and experiences from resumes. This technology aims to achieve the following:

1. **Enhanced Efficiency:** NLP automates the initial screening of resumes, freeing up recruiters' time for more strategic tasks like candidate interviews and evaluations.
2. **Improved Accuracy:** By analyzing text with advanced language understanding, NLP parsers can extract information with high accuracy, minimizing the risk of missing qualified candidates due to keyword limitations or human bias.

3. **Objective Matching:** NLP focuses on factual information from resumes, reducing unconscious bias and ensuring a more objective selection process based on skills and experience alignment with job requirements.
4. **Targeted Recommendations:** For both recruiters and job seekers, NLP can facilitate targeted matching. Recruiters can find candidates with the precise skills needed, while job seekers can discover opportunities that genuinely align with their skillsets.
5. **Data-Driven Insights:** NLP parsers can analyze large volumes of resumes, extracting valuable data on applicant demographics, skill trends, and job requirements mentioned by candidates. This data can be used by recruiters to inform strategic decisions regarding talent acquisition, skills development programs, and job description optimization.
6. **Scalability and Flexibility:** NLP-based parsers can handle large volumes of resumes efficiently, making them ideal for organizations with high recruitment needs. Additionally, NLP parsers can be continuously updated and improved, adapting to evolving skills and job requirements in the market.

CHAPTER 2

LITERATURE REVIEW

Bhatia, et al., (2019), **“End-to-End Resume Parsing and Finding Candidates for a Job Description using BRET”** in this research, they explored the possibility of building a standard parser for all types of resumes and determined that it was impossible to do so without losing information in all situations, resulting in the unfair rejection of specific candidates' resumes. Instead, they used LinkedIn-style resumes to scan without losing any information. They wanted to investigate the vision-based site segmentation technique in the future in order to improve structural comprehension of resumes. In addition, the study also creates a firm foundation and a feasibility study that can lead to advancements in deep learning and language representation being used in the hiring process. The system diagram below represents the data flow and the completed task.

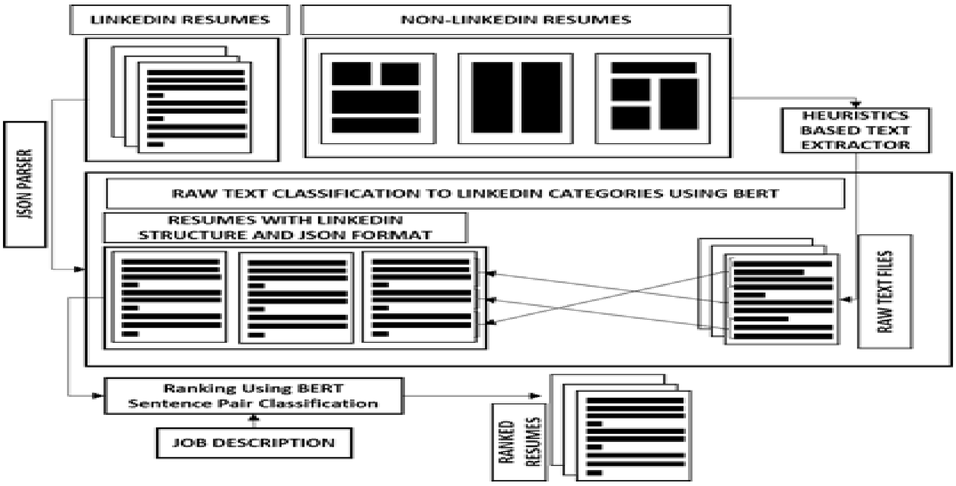

LINKEDIN RESUMES
NON-LINKEDIN RESUMES
HEURISTICS BASED TEXT EXTRACTOR
JSON PARSER
RAW TEXT CLASSIFICATION TO LINKEDIN CATEGORIES USING BERT
RESUMES WITH LINKEDIN STRUCTURE AND JSON FORMAT
RAW TEXT FILES
Ranking Using BERT Sentence Pair Classification
JOB DESCRIPTION
RANKED RESUMES

Figure 2.1: A system diagram showing data transmission and task completion.

Chen, et al., (2016) **“Information Extraction from Resume Documents in PDF Formats”**,this paper focuses on the problem of extracting data from PDF-format resumes and proposes a hierarchical extraction technique. Resume papers break a page into blocks using heuristic criteria, categorize each block using a Conditional Random Field (CRF) model, and approach the detailed information extraction problem as a sequence labeling issue. In comparison to HTML resumes, PDF resumes usually include more detailed information. They want to experiment with various page segmentation techniques in the future to improve their understanding of the document's layout and content. The report's method is explained below.

Nguyen, et al., (2018) **“ Study of Information Extraction in Resume”,**this paper proposes Text Segmentation, Rule-based Named Entity Recognition, Deep Neural Network Find Name Entities, and Text Normalization approaches that automatically retrieve and process multiple resumes formats. To label the sequence in the resume and then extract Name Entities from the labeled line, the author defined the Deep Learning model as a mix of Convolutional Neural Networks, Bidirectional Long Short-Term Memory, and Conditional Random Field. The developer collected promising findings with over 81 percent F1 for NER when experimenting on a medium-sized collection of CV information and compared this model to other systems. However, there are some flaws in this system that might be addressed. The quantity of data to train the model is insufficient. Thus, the more data, the more accurate the model will be. In addition, the model's calculations necessitate the use of sophisticated computers. The author wishes to improve the hardware systems to improve performance in the future.

Kopparapu, (2015) **“Automatic Extraction of Usable Information from Unstructured Resumes to Aid Search”,** this paper proposes a natural language processing (NLP) system that focuses on automated information extraction from resume to facilitate speedy resume search and management for structured and unstructured resumes. According,to the results of experiments conducted on many resumes, the suggested system can grip a wide range of resumes in various document formats with an accuaracy.

VUKADIN, et al.,(2021)**“Information Extraction from Free-Form CV Documents in Multiple Languages”** in this paper they explained that the use of two natural language processing algorithms to extract important data from an unstructured multilingual CV has provided a solution for selecting relevant document parts and the similar particular information at the low hierarchy In their practice, authors used the transformer architecture and its application of the encoder part in the BERT language model. A dual model was developed to extract both section and item level information from a CV document. A self-assessment model of skill proficiency categorizes the retrieved Skills section from the dual model. The authors claim that they have solved the CV parsing challenge by building an NLP system.

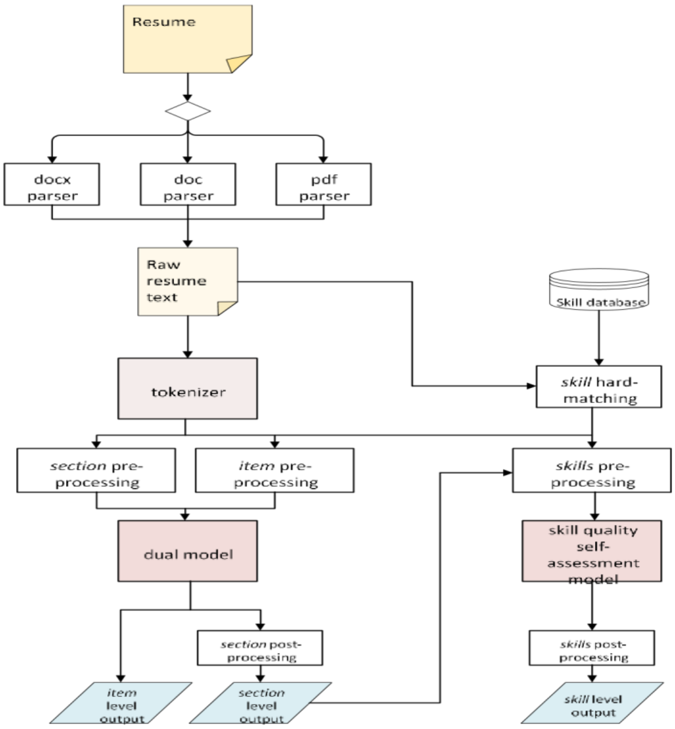


Figure 2.2: The Information Extraction System from Free-form CVs: A High-Level Overview

Wang & Zu(2019) **“Resume Information Extraction with a Novel Text Block Segmentation Algorithm”,**The proposed system combines text block segmentation with resume fact identification using position-wise line information and integrated word representations and named entity recognition using multiple sequence labeling classifiers inside labeled text blocks. The ablation experiment was conducted by eliminating the CNN layer from BLSTMCNN's-CRF and comparing various word embeddings testifying. Personal information, job experience, education, project experience, publications, and professional abilities are the six main areas they have focused on to normalize the resume parsing process. Finally, the authors created an online resume parser based on the proposed resume information extraction method, and it turned out that the system works effectively in practice.

CHAPTER 3

EXISTING SYSTEM

The ever-growing volume of resumes in the recruitment process has spurred the development of NLP-powered resume parsers. These systems aim to automate the tedious task of extracting key information from resumes, streamlining the workflow for recruiters and job seekers alike. However, parsing resumes effectively presents a unique challenge due to their unstructured nature. Layouts and formats can vary widely, with information presented in unpredictable sections and using inconsistent language. Existing NLP-based resume parsers typically follow a multi-stage approach to address these complexities. The first stage involves text processing, where the system cleans and prepares the resume text. This might involve converting it from PDF or DOCX format to plain text, removing irrelevant sections like headers or footers, and correcting any formatting errors.

Next comes the crucial step of information extraction. Techniques like part-of-speech tagging help the system understand the grammatical context of keywords, further improving accuracy.Section identification is another critical aspect. The parser employs various strategies to discern sections like "Experience," "Skills," and "Education" within the resume text. This might involve using regular expressions for specific keywords or leveraging machine learning models trained to recognize section headers based on layout and formatting cues.

Once relevant sections are identified, the parser extracts specific details like job titles, companies, dates of employment, and skills. Here, NLP techniques like rule-based approaches, statistical methods (like TF-IDF), and deep learning models can be employed. Rule-based systems rely on pre-defined patterns to identify specific skills or experiences, while statistical methods analyze the frequency of keywords and their correlation with job descriptions.

Finally, the parsed information gets structured and stored in a standardized format, often a database or a table. This allows recruiters to easily search and filter resumes based on specific criteria like skills, experience level, or keywords. Additionally, some parsers may generate reports that highlight key qualifications and potential matches for job openings.With continued development, NLP-powered resume parsers have the potential to revolutionize the recruitment process, saving time, reducing bias, and facilitating a more efficient and effective talent acquisition process.

3.1 DRAWBACK OF EXISTING SYSTEM

- Low System Value with Limited Data**:** NLP parsers thrive on large datasets to train their models effectively. With a small number of resumes uploaded, the parser might not have enough data to learn the nuances of language used by candidates relevant to your industry or position. This can lead to inaccurate parsing and missed opportunities to identify qualified candidates from a smaller pool.
- Difficulty Adapting to Specific Needs**:** When there aren't many resumes uploaded, it becomes harder to tailor the NLP parser to the specific needs of your company or the particular job opening. The parser might struggle to identify the most relevant skills and experience for the role, leading to poor candidate matching.

In essence, with a low volume of resumes, NLP parsers might not be able to reach their full potential. They might require additional human intervention to define specific criteria or identify relevant skills, reducing the automation benefit.

CHAPTER 4

PROPOSED METHODOLOGY

In today's competitive job market, sifting through a mountain of resumes can be a time-consuming and laborious task for recruiters. Traditional methods often rely on manual keyword searches and skimming through text, leading to inconsistencies and missed opportunities. This is where Natural Language Processing (NLP) steps in as a game-changer. This proposal outlines a comprehensive NLP-powered resume parsing system designed to revolutionize the recruitment process by automatically extracting and structuring key information from resumes.

The system begins by tackling the challenge of unstructured data.This proposed NLP-powered resume parser tackles traditional limitations by handling bulk uploads, offering in-depth skill comparison, and promoting fairness in recruitment. It streamlines the process by accepting multiple resumes (PDF, DOCX, etc.) for batch processing with an advanced text extraction engine. A multi-stage information extraction follows, utilizing state-of-the-art NER models to identify key details, particularly skills. To move beyond basic keyword matching, the system employs a three-pronged approach: 1) Skills are mapped to a customizable taxonomy, providing a structured representation. 2) WordNet and synonym dictionaries help identify related terms, capturing a wider range of skill descriptions. 3) Sentiment analysis and dependency parsing into skill descriptions, understanding expertise level and associated tools/technologies. Recruiters create detailed job descriptions within the system, specifying required skills mapped to the taxonomy for consistent comparison. The core lies in the advanced skill matching algorithm. It compares extracted skills against job requirements, considering synonyms, descriptions, and taxonomy mapping, to offer a nuanced assessment of candidate fit. A ranked list of candidates is generated based on skill match score, encompassing the number of matching skills, expertise level, and relevance to specific job requirements. To address bias, anonymized processing is used initially, ensuring skills are evaluated objectively. Additionally, comprehensive reports with detailed skill matching information and candidate summaries empower recruiters to visualize candidate skillsets and identify potential matches efficiently. This system offers a significant advantage by handling a large volume of resumes simultaneously with accurate skill extraction. The advanced skill matching ensures a deeper understanding of candidate capabilities and better alignment with job requirements. Anonymized processing and a focus on skills help mitigate bias, while providing a ranked list of qualified candidates with detailed skill profiles empowers recruiters to make informed hiring decisions. By leveraging NLP's power and addressing current limitations, this proposed system promises to revolutionize resume parsing and evaluation in the recruitment landscape

In conclusion, this NLP-powered resume parsing system goes beyond simply automating tasks. It empowers recruiters to make data-driven decisions, identify top talent faster, and ultimately build a more efficient and effective recruitment process. By transforming unstructured resumes into valuable, structured data, this system allows recruiters to focus on what truly matters - finding the best person for the job.

4.1 ADVANTAGES OF PROPOSED SYSTEM

Enhanced Efficiency and Scalability:

- **Automates Information Extraction:** Manual screening of hundreds or even thousands of resumes is a herculean task. NLP parsers automate the process of extracting key information like skills, experience, and education, freeing up recruiters' time to focus on higher-level tasks like candidate evaluation and interview scheduling.
- **Bulk Upload Capability:** Unlike traditional methods limited to single resumes, NLP parsers can handle large batches of resumes uploaded at once.

This is particularly beneficial for high-volume recruitment drives or companies receiving a constant stream of applications.

Data-Driven Insights and Talent Pool Analysis:

- **Skill Trends and Candidate Demographics:** With a large dataset of parsed resumes, recruiters can identify trends in skills and qualifications sought after by applicants. This allows them to adapt their recruitment strategies and job descriptions to attract the best talent pool and adjust hiring requirements based on market demands.

Reduced Bias and Streamlined Workflow:

- **Focus on Objective Criteria:** NLP parsers rely on skills and experience extracted from the text, minimizing the influence of subjective factors on resume screening. This helps reduce bias in the hiring process and ensures a fairer evaluation of candidates based on their qualifications.
- **Faster Shortlisting and Candidate Matching:** With accurate and structured data readily available, recruiters can shortlist qualified candidates quickly and efficiently. NLP parsers can even match resumes to specific job openings based on skill requirements, further streamlining the workflow.

4.2 BLOCK DIAGRAM /FLOW DIAGRAM

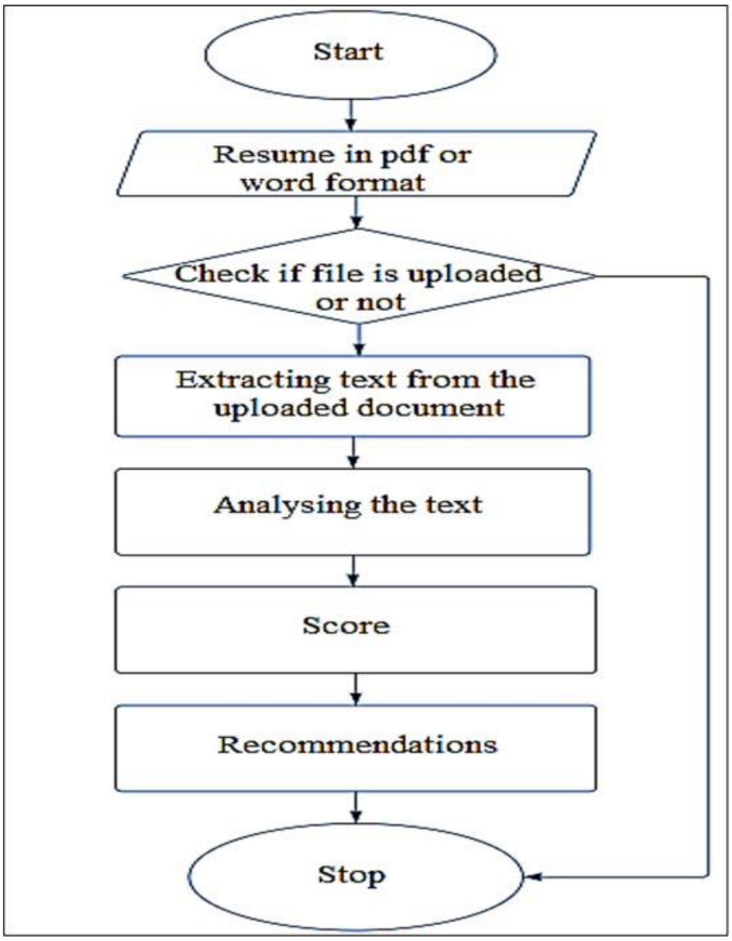

Start
Resume in pdf or word format
Check if file is uploaded or not
Extracting text from the uploaded document
Analysing the text
Score
Recommendations
Stop

Figure 4.1 : Workflow of the proposed system

1. Resume Ingestion and Preprocessing:

- **Bulk Upload:** The system allows uploading multiple resumes in various formats (PDF, Word doc, plain text) simultaneously.
- **Format Conversion (Optional):** If necessary, resumes are converted to a common format (e.g., plain text) for easier processing.

- **Text Cleaning:** Techniques like removing punctuation, converting text to lowercase, and correcting spelling errors are applied to ensure consistent data.

2. Skill and Experience Extraction:

- **Stop Word Removal:** Common words with little meaning (e.g., "the", "a", "an") are removed from the preprocessed text.
- **Keyword Extraction:**
  - **Custom Keyword List:** A comprehensive list of relevant skills and experience keywords specific to your industry and target roles is created. This can be built manually or through tools like job posting analysis.
- **Part-of-Speech Tagging:** NLP techniques are used to identify the grammatical role (noun, verb, adjective) of each word in the resume text.
- **Dependency Parsing (Optional):** Analyze the relationships between words in a sentence to understand the context of skills and experience mentioned, particularly helpful for identifying skills embedded within phrases.

3. **Pre-processing and Text Extraction:**

- The system utilizes a text extraction and conversion engine to convert the uploaded resumes into a machine-readable format (plain text) if necessary.
- Text cleaning techniques are applied to remove irrelevant sections like headers, footers, and watermarks.

4. **Company Requirement Integration:**

- Recruiters can create detailed job descriptions within the system, specifying required skills and desired experience levels. Similar to skills extracted from resumes, these can be mapped to the skills taxonomy for consistent comparison.

5. **Skill Matching and Candidate Ranking:**

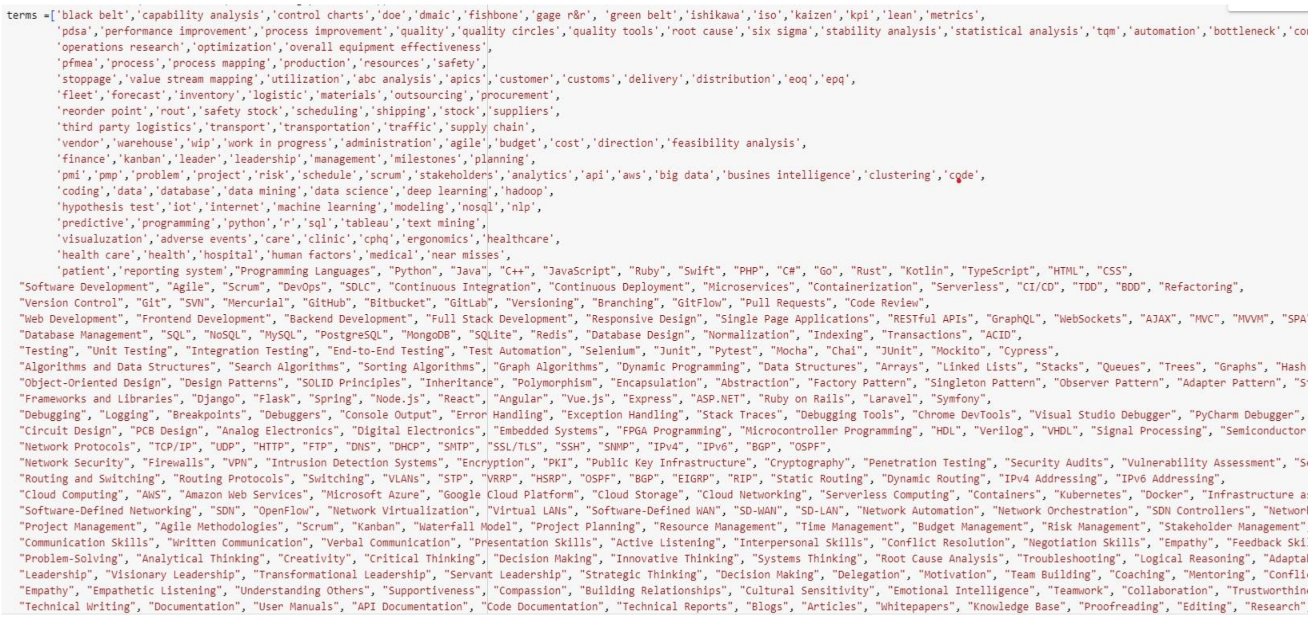
- The core of the system is the skill matching algorithm. It compares the extracted skills from resumes against the required skills in the job description. By considering factors like:
    - Exact matches
    - Synonyms and related terms (if normalization was applied)
    - Skill descriptions and context (e.g., level of expertise, tools used)
- The algorithm generates a ranked list of candidates based on their skill match score. This score reflects the number of matching skills, the level of expertise indicated in the resume, and the relevance to the specific job requirements.

6. **Output and Reporting:**

- The system generates reports with:
    - Ranked list of candidates with skill match scores
- Recruiters can use these reports to visualize candidate skill sets in pie chart or in any graphs where human can understand and identify potential matches efficiently.

4.3 DATASET COLLECTION

Our NLP-based resume parser relies on a meticulously curated keyword dataset to power its skill comparison and ranking capabilities. This dataset will be a comprehensive collection of keywords encompassing both technical and soft skills commonly found in resumes and job descriptions. The keywords will be organized into a hierarchical structure, potentially using a skills taxonomy, for improved understanding and comparison. Additionally, the dataset will incorporate synonyms and related terms for each skill, leveraging resources like WordNet or synonym dictionaries. This breadth of keywords ensures the system can capture various skill descriptions and variations present in resumes, leading to more accurate skill matching and candidate evaluation during bulk resume processing.



```
terms =['black belt','capability analysis','control charts','doe','dmaic','fishbone','gage r&r', 'green belt','ishikawa','iso','kaizen','kpi','lean','metrics',
        'pdsa','performance improvement','process improvement','quality','quality circles','quality tools','root cause','six sigma','stability analysis','statistical analysis','tqm','automation','bottleneck','co
        'operations research','optimization','overall equipment effectiveness',
        'pfmea','process','process mapping','production','resources','safety',
        'stoppage','value stream mapping','utilization','abc analysis','apics','customer','customs','delivery','distribution','eoq','epq',
        'fleet','forecast','inventory','logistic','materials','outsourcing','procurement',
        'reorder point','rout','safety stock','scheduling','shipping','stock','suppliers',
        'third party logistics','transport','transportation','traffic','supply chain',
        'vendor','warehouse','wip','work in progress','administration','agile','budget','cost','direction','feasibility analysis',
        'finance','kanban','leader','leadership','management','milestones','planning',
        'pmi','pmp','problem','project','risk','schedule','scrum','stakeholders','analytics','api','aws','big data','busines intelligence','clustering','code',
        'coding','data','database','data mining','data science','deep learning','hadoop',
        'hypothesis test','iot','internet','machine learning','modeling','nosql','nlp',
        'predictive','programming','python','r','sql','tableau','text mining',
        'visualuzation','adverse events','care','clinic','cphq','ergonomics','healthcare',
        'health care','health','hospital','human factors','medical','near misses',
        'patient','reporting system',"Programming Languages", "Python", "Java", "C++", "JavaScript", "Ruby", "Swift", "PHP", "C#", "Go", "Rust", "Kotlin", "TypeScript", "HTML", "CSS",
  "Software Development", "Agile", "Scrum", "DevOps", "SDLC", "Continuous Integration", "Continuous Deployment", "Microservices", "Containerization", "Serverless", "CI/CD", "TDD", "BDD", "Refactoring",
  "Version Control", "Git", "SVN", "Mercurial", "GitHub", "Bitbucket", "GitLab", "Versioning", "Branching", "GitFlow", "Pull Requests", "Code Review",
  "Web Development", "Frontend Development", "Backend Development", "Full Stack Development", "Responsive Design", "Single Page Applications", "RESTful APIs", "GraphQL", "WebSockets", "AJAX", "MVC", "MVVM", "SPA
  "Database Management", "SQL", "NoSQL", "MySQL", "PostgreSQL", "MongoDB", "SQLite", "Redis", "Database Design", "Normalization", "Indexing", "Transactions", "ACID",
  "Testing", "Unit Testing", "Integration Testing", "End-to-End Testing", "Test Automation", "Selenium", "Junit", "Pytest", "Mocha", "Chai", "JUnit", "Mockito", "Cypress",
  "Algorithms and Data Structures", "Search Algorithms", "Sorting Algorithms", "Graph Algorithms", "Dynamic Programming", "Data Structures", "Arrays", "Linked Lists", "Stacks", "Queues", "Trees", "Graphs", "Hash
  "Object-Oriented Design", "Design Patterns", "SOLID Principles", "Inheritance", "Polymorphism", "Encapsulation", "Abstraction", "Factory Pattern", "Singleton Pattern", "Observer Pattern", "Adapter Pattern", "S
  "Frameworks and Libraries", "Django", "Flask", "Spring", "Node.js", "React", "Angular", "Vue.js", "Express", "ASP.NET", "Ruby on Rails", "Laravel", "Symfony",
  "Debugging", "Logging", "Breakpoints", "Debuggers", "Console Output", "Error Handling", "Exception Handling", "Stack Traces", "Debugging Tools", "Chrome DevTools", "Visual Studio Debugger", "PyCharm Debugger",
  "Circuit Design", "PCB Design", "Analog Electronics", "Digital Electronics", "Embedded Systems", "FPGA Programming", "Microcontroller Programming", "HDL", "Verilog", "VHDL", "Signal Processing", "Semiconductor
  "Network Protocols", "TCP/IP", "UDP", "HTTP", "FTP", "DNS", "DHCP", "SMTP", "SSL/TLS", "SSH", "SNMP", "IPv4", "IPv6", "BGP", "OSPF",
  "Network Security", "Firewalls", "VPN", "Intrusion Detection Systems", "Encryption", "PKI", "Public Key Infrastructure", "Cryptography", "Penetration Testing", "Security Audits", "Vulnerability Assessment", "S
  "Routing and Switching", "Routing Protocols", "Switching", "VLANs", "STP", "VRRP", "HSRP", "OSPF", "BGP", "EIGRP", "RIP", "Static Routing", "Dynamic Routing", "IPv4 Addressing", "IPv6 Addressing",
  "Cloud Computing", "AWS", "Amazon Web Services", "Microsoft Azure", "Google Cloud Platform", "Cloud Storage", "Cloud Networking", "Serverless Computing", "Containers", "Kubernetes", "Docker", "Infrastructure a
  "Software-Defined Networking", "SDN", "OpenFlow", "Network Virtualization", "Virtual LANs", "Software-Defined WAN", "SD-WAN", "SD-LAN", "Network Automation", "Network Orchestration", "SDN Controllers", "Networ
  "Project Management", "Agile Methodologies", "Scrum", "Kanban", "Waterfall Model", "Project Planning", "Resource Management", "Time Management", "Budget Management", "Risk Management", "Stakeholder Management"
  "Communication Skills", "Written Communication", "Verbal Communication", "Presentation Skills", "Active Listening", "Interpersonal Skills", "Conflict Resolution", "Negotiation Skills", "Empathy", "Feedback Ski
  "Problem-Solving", "Analytical Thinking", "Creativity", "Critical Thinking", "Decision Making", "Innovative Thinking", "Systems Thinking", "Root Cause Analysis", "Troubleshooting", "Logical Reasoning", "Adapta
  "Leadership", "Visionary Leadership", "Transformational Leadership", "Servant Leadership", "Strategic Thinking", "Decision Making", "Delegation", "Motivation", "Team Building", "Coaching", "Mentoring", "Confli
  "Empathy", "Empathetic Listening", "Understanding Others", "Supportiveness", "Compassion", "Building Relationships", "Cultural Sensitivity", "Emotional Intelligence", "Teamwork", "Collaboration", "Trustworthin
  "Technical Writing", "Documentation", "User Manuals", "API Documentation", "Code Documentation", "Technical Reports", "Blogs", "Articles", "Whitepapers", "Knowledge Base", "Proofreading", "Editing", "Research"
```

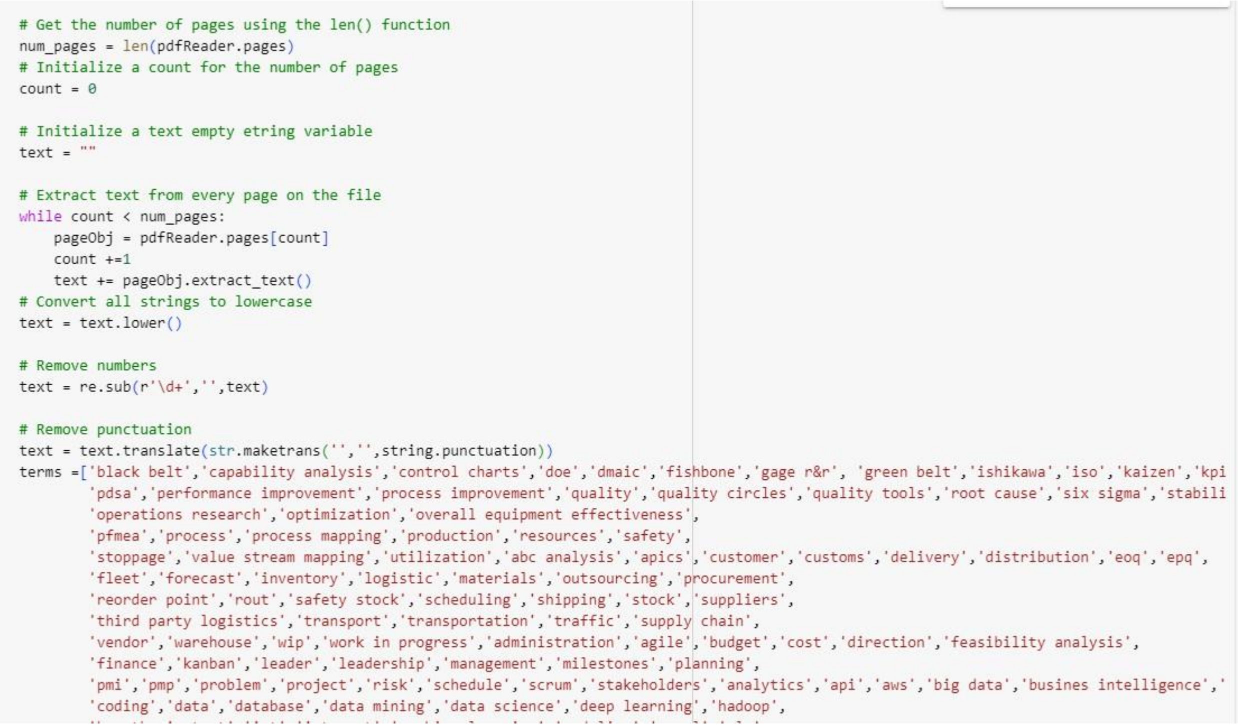
Figure 4.2 : Dataset Collection

4.4 PSEUDOCODE

```
# Initialize an empty array
array = []
# Loop through each resume
For i from 1 to 1:
    # Open pdf file
    pdfFileObj = open('/content/resume' + str(i) + '.pdf', 'rb')
    # Read file
    pdfReader = PyPDF2.PdfReader(pdfFileObj)


    # Get the number of pages
    num_pages = len(pdfReader.pages)
    # Initialize a text variable
    text = ""
    # Extract text from every page on the file
    For each page in pdfReader.pages:
        text += page.extract_text()
    # Convert text to lowercase
    text = text.lower()
    # Remove numbers
    text = re.sub(r'\d+', ", text)
    # Remove punctuation
    text = text.translate(str.maketrans(", ", string.punctuation))
    # Define a list of terms to search for
    terms = ["effectiveness", "intelligence", "clustering", ...]
    # Initialize a score variable
    scores = 0
    # Loop through each term
    For each word in terms:
        If word in text:
            scores += 1
    # Append score to the array
    array.append(scores)

# Create a list of skills
skills = []
For i from 1 to the length of array:
    skills.append("resume" + str(i))
# Plot pie chart
pie = plt.figure(figsize=(10, 10))
plt.pie(array, labels=skills, autopct='%.2f', startangle=90)
plt.title('Resume Decomposition by Areas')
plt.show()

# Save pie chart as a .png file
pie.savefig('resume_screening_results.png')
```

CHAPTER 5
RESULTS AND DISCUSSION

NLP can revolutionize resume screening by parsing documents for specific job requirements. Resumes are first converted into a format the computer understands. NLP then analyzes the text, identifying parts of speech and recognizing key entities like skills and experience. You provide a list of 8 keywords reflecting the essential qualities for the position. The system matches these keywords against the resume, but also goes beyond by considering context. It can find skills described within work experience sections, not just dedicated "Skills" areas. This structured analysis helps recruiters efficiently shortlist candidates with the most relevant qualifications. While challenges exist with resume format variations and keyword nuances, NLP offers a powerful tool to identify top talent based on your specific needs.



```
# Get the number of pages using the len() function
num_pages = len(pdfReader.pages)
# Initialize a count for the number of pages
count = 0

# Initialize a text empty etring variable
text = ""

# Extract text from every page on the file
while count < num_pages:
    pageObj = pdfReader.pages[count]
    count +=1
    text += pageObj.extract_text()
# Convert all strings to lowercase
text = text.lower()

# Remove numbers
text = re.sub(r'\d+','',text)

# Remove punctuation
text = text.translate(str.maketrans('','',string.punctuation))
terms =['black belt','capability analysis','control charts','doe','dmaic','fishbone','gage r&r', 'green belt','ishikawa','iso','kaizen','kpi
        'pdsa','performance improvement','process improvement','quality','quality circles','quality tools','root cause','six sigma','stabili
        'operations research','optimization','overall equipment effectiveness',
        'pfmea','process','process mapping','production','resources','safety',
        'stoppage','value stream mapping','utilization','abc analysis','apics','customer','customs','delivery','distribution','eoq','epq',
        'fleet','forecast','inventory','logistic','materials','outsourcing','procurement',
        'reorder point','rout','safety stock','scheduling','shipping','stock','suppliers',
        'third party logistics','transport','transportation','traffic','supply chain',
        'vendor','warehouse','wip','work in progress','administration','agile','budget','cost','direction','feasibility analysis',
        'finance','kanban','leader','leadership','management','milestones','planning',
        'pmi','pmp','problem','project','risk','schedule','scrum','stakeholders','analytics','api','aws','big data','busines intelligence','
        'coding','data','database','data mining','data science','deep learning','hadoop',
```

Figure 5.1: Extracting text from a PDF and comparing it with keywords

The output of a resume parser using NLP that compares skills in a resume with job requirements is typically a pie chart. This chart visually represents the percentage between the candidate's skills and the skills listed in the job description and also the rank between the candidates through pie chart.

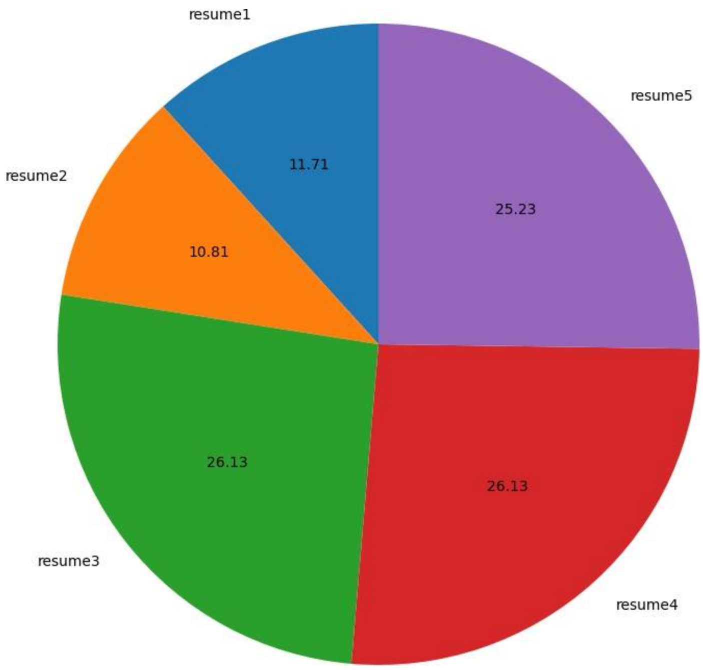


Figure 5.2: Pie chart representation of Resume Parser

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

CONCLUSION

This resume parser, powered by Natural Language Processing (NLP), streamlines the recruitment process by efficiently comparing candidate skills to specific job requirements. It can handle numerous resumes, allowing you to evaluate a larger pool of. It then matches these skills against a pre-defined list of job requirements. Scores are calculated based on skill matches, providing a quantitative assessment of each candidate's suitability. This approach helps prioritize resumes based on relevant skills, saving valuable time and resources during the hiring process. By automating the initial screening stage, you can focus your efforts on interviewing the most promising candidates.

FUTURE SCOPE

Skill Level Detection: Move beyond simply identifying skills. Explore techniques to categorize skills by proficiency level (beginner, intermediate, advanced) based on keywords and context within the resume. This can involve analyzing the specific duties and responsibilities mentioned for past experiences.Recommendation Systems: Build recommendation systems that suggest potential candidates based ontheir extracted skills and experience. This can be particularly valuable fidentifying passive candidates who might not be actively searching for new opportunities.

REFERENCES

1. Bhatia, V., Rawat, P., Kumar, A., & Shah, R. R. (2019), “End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT”, arXiv:1910.03089, https://doi.org/10.48550/arXiv.1910.03089.
2. Chen, J., Gao, L., & Tang, Z. (2016), “Information Extraction from Resume Documents in PDF”, IS&T International Symposium on Electronic Imaging 2016, Society for Imaging Science and Technology, DRR-064.1-8.
3. Dr. Parkavi A, Pooja Pandey, Poornima J, Vaibhavi G S, Kaveri BW, “E-Recruitment System Through Resume Parsing, Psychometric Test and Social Media Analysis”, 2019, International journal of advanced research in basic engineering sciences and technology(IJARBEST).
4. Nguyen, V. V., Pham, V. L., & Vu, N. S. (2018). Study of Information Extraction in Resume.
5. Nirali Bhaliya , Jay Gandhi, Dheeraj Kumar Singh, “NLP-based Extraction of Relevant Resume using Machine Learning”, 2020, International Journal of Innovative Technology and Exploring Engineering(IJITEE).
6. Resume Parser with Natural Language Processing. (IJESC) by Satyaki Sanyal, Neelanjan Ghosh, SouvikHazra, Soumyashree Adhikary (2007).
7. Study of Information Extraction in Resume. (semanticscholar) by Nguyen, V. V., Pham, V. L., & Vu, N. S. (2018).
8. Unstructured Text Analytics Approach for Qualitative Evaluation of Resumes, An (IJIRAE) by Vinaya R. Kudatarkar, ManjulaRamannavar, Dr.Nandini S. Sidnal (2015).
9. Vukadin, D., Kurdija, A. S., Delac, G., & Silic, M. (2021). Information Extraction From Free-Form Text Documents. (IEEE Explore)
10. Wang, X., & Zu, S. (2019). Resume Information Extraction with A Novel Text Block Segmentation Algorithm, International Journal on Natural Language Computing (IJNLC) Vol.8, No.5, October 2019.

RAJALAKSHMI ENGINEERING COLLEGE

DEPARTMENT OF ECE

PROGRAM OUTCOMES (POs)

Engineering Graduates will be able to:

PO1 Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

PO2 Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

PO3 Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

PO4 Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO5 Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

PO6 The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO7 Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO8 Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

PO9Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO10Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO11 Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO12 Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1: An ability to carry out research in different areas of Electronics and Communication Engineering fields resulting in journal publications and product development.

PSO2: To design and formulate solutions for industrial requirements using Electronics and Communication engineering

PSO3: To understand and develop solutions required in multidisciplinary engineering fields.

COURSE OUTCOMES (COs)

CO1	To acquire practical knowledge within the chosen area of technology for project development.
CO2	To identify, analyze, formulate and handle projects with a comprehensive and systematic approach.
CO3	To contribute as an individual or in a team in development of technical projects.
CO4	To develop effective communication skills for presentation of project related activities.
CO5	To extend the work and make it as a final year project.

EC19603 – PROBLEM SOLVING USING AI AND ML TECHNIQUES

(Mini Project)

Project Title: RESUME PARSER USING NATURAL LANGUAGE PROCESSING

Batch Members : AMRITHAA R.S (2116210801009)

ANANDHA KRISHNAN S (2116210801010)

ANUBAMA J (2116210801013)

Name of the Supervisor : MS.SUSHMA S.JAGTAP, M.E., (Ph.D.),

ASSISTANT PROFESSOR.

CO - PO – PSO matrices of course

PO/PSO CO	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PS O1	PS O2	PS O3
CO1	3	3	3	3	1	3	3	3	1	2	-	2	2	3	3
CO2	3	3	3	3	3	2	2	2	3	3	3	3	3	3	3
CO4	3	3	3	3	3	2	2	2	3	3	3	3	3	3	3
CO5	3	3	3	3	3	2	2	2	3	3	3	3	3	3	3
Average	3	3	3	3	2.6	2.2	2.2	2.2	2.6	2.8	3	2.8	2.8	3	3

Note: Enter correlation levels 1, 2 or 3 as defined below:

1: Slight (Low) 2: Moderate (Medium) 3: Substantial (High), If there is no correlation, put -“

Signature of the Supervisor