BMS COLLEGE OF ENGINEERING BENGALURU
Autonomous Institute, Affiliated to VTU

A Technical Seminar Report based on Technical Activity

Leveraging AI Chatbots for Resume Parsing and Rating using LLMs and NLP Techniques

Submitted in partial fulfillment for the award of degree of

Bachelor of Engineering
in
Computer Science and Engineering

Submitted by:
NEHA BHASKAR KAMATH (1BM21CS113)

Work carried out at

Internal Guide

Namratha M
Assistant Professor
B.M.S. College of Engineering

Department of Computer Science and Engineering
BMS College of Engineering
Bull Temple Road, Basavanagudi, Bangalore 560 019

2022-2023

BMS COLLEGE OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

I, NEHA BHAKAR KAMATH (1BM21CS113), student of 4thSemester, B.E, Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, hereby declare that, this technical seminar entitled "Leveraging AI Chatbots for Resume Parsing and Rating using LLMs and NLP Techniques" has been carried out under the guidance of Namratha M, Assistant Professor, Department of CSE, BMS College of Engineering, Bangalore during the academic semester March-July 2023. I also declare that to the best of my knowledge and belief, the technical seminar report is not from part of any other report by any other students.

Signature of the Candidate

NEHA BHASKAR KAMATH (1BM21CS113)

BMS COLLEGE OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Technical Seminar titled "**Leveraging AI Chatbots for Resume Parsing and Rating using LLMs and NLP Techniques"** has been carried out by NEHA BHAKAR KAMATH (1BM21CS113) during the academic year 2022-2023.

Signature of the Guide

Signature of the Head of Department

Signature of Examiners with date

1. Internal Examiner _____

2. External Examiner _____

ABSTRACT

This project presents a comprehensive system leveraging Natural Language Processing (NLP) techniques to revolutionize the recruitment process. The primary focus is on automating the extraction of critical information from candidate resumes in PDF format, significantly expediting the evaluation of qualifications. NLP algorithms play a pivotal role in identifying and categorizing elements such as candidate names, contact details, work history, educational background, skills, and more. This automated process not only accelerates candidate evaluation but also enables efficient storage of the extracted data for subsequent analysis.

Moreover, the project features a chatbot interface that harnesses NLP for recruiter interactions. Recruiters can effortlessly select job categories, with NLP algorithms adeptly deciphering user intent by recognizing keywords, phrases, and context. The chatbot then interfaces with a backend API to initiate the retrieval of pertinent resumes. NLP technology ensures precise communication between recruiters and the system, streamlining the process of accessing the most fitting candidate profiles.

This project not only demonstrates the potential of NLP in optimizing recruitment workflows but also underscores its ability to enhance the efficiency and effectiveness of the hiring process. The integration of resume parsing and chatbot-driven interactions showcases the power of NLP in modernizing talent acquisition, resulting in substantial time savings and improved candidate selection.

TABLE OF CONTENTS

Chapter 1: Introduction

1.1 Overview

In the current landscape of talent acquisition, the process of identifying, evaluating, and selecting the right candidates for job openings is marked by challenges that demand a modernized approach. Our project aims to introduce a cutting-edge system that harnesses the capabilities of Natural Language Processing (NLP) to revolutionize the recruitment process.

Our model utilizes NLP methods to analyze resumes uploaded by job seekers through a dedicated job portal. These resumes are categorized and stored in databases based on specific job titles. Recruiters access a recruiting portal where they can interact with a chatbot to specify their desired qualifications for candidates. The chatbot then leverages NLP to examine the parsed resumes in the database and provides the most suitable matches for the open positions.

1.2 Motivation

The motivation behind this project stems from the recognition of significant challenges in the traditional recruitment process. In the ever-evolving job market, both job seekers and employers face a myriad of hurdles that demand innovative solutions. These challenges include the sheer volume of resumes, the need for swift and accurate candidate evaluation, the imperative to reduce bias in hiring, and the desire for more efficient and data-driven recruitment practices.

The conventional approach to reviewing resumes is time-consuming, prone to human error, and often influenced by unconscious biases. Recruiters are inundated with an avalanche of resumes, making it arduous to identify the most qualified candidates efficiently. Additionally, the risk of bias in the selection process, consciously or unconsciously, can hinder the pursuit of diversity and inclusivity in the workplace.

In summary, the motivation behind this project is to address the pressing challenges in modern recruitment by leveraging NLP techniques to enhance efficiency, accuracy, and fairness. We aim to empower both job seekers and employers with a state-of-the-art system that streamlines the hiring process, enabling faster, more informed, and unbiased decision-making, ultimately contributing to the advancement of talent acquisition practices in today's dynamic job market.

1.3 Objective

- Automated Resume Parsing: Develop a robust system that can automatically extract essential information from resumes uploaded by job seekers in PDF format. The objective is to accurately capture data such as candidate names, contact details, work history, education, and skills without manual intervention.

- Database Organization: Create a well-structured and categorized database to efficiently store parsed resumes based on specific job titles or positions. The goal is to facilitate quick and precise retrieval of resumes matching the requirements of open positions.

- Chatbot Interface: Implement an intelligent chatbot interface within the recruiting portal to enable seamless interactions between recruiters and the system. The chatbot should understand natural language input from recruiters and respond accordingly.

- Recruiter Communication: Ensure that the chatbot effectively interprets the intent of recruiters when they specify desired qualifications for candidates. This involves recognizing keywords, phrases, and context to initiate relevant searches.

- NLP-Powered Candidate Matching: Leverage NLP techniques to develop algorithms that can efficiently query the resume database based on the recruiter's input and retrieve the most suitable candidate profiles.

- Result Delivery: Present the matched candidate resumes to recruiters in a user-friendly format for easy review and consideration. Ensure that the system provides relevant and accurate results.

- Accuracy and Consistency: Strive for high accuracy and consistency in extracting information from resumes. Minimize errors and variations in data extraction to enhance the quality of candidate profiles.

- Bias Reduction: Implement NLP algorithms to minimize bias in candidate selection, focusing on qualifications and skills rather than personal characteristics. Promote fairness and diversity in the recruitment process.

- Efficiency Improvement: Reduce the time and effort required for manual resume screening by automating the process. Aim for significant time savings in candidate evaluation and selection.

- Scalability: Design the system to be scalable, allowing it to handle a large volume of resumes and job categories without compromising performance or responsiveness.

Chapter 2: Literature Survey

Sunhao, Ninglu, et al. 2023 [1] leverage prompts to harness ChatGPT's ranking abilities for recommendation tasks. Prompts consist of three components: (i) Task description, helping the model understand the domain, (ii) Demonstration examples, illustrating task performance, and (iii) Target item, the item to rank. This approach aims to unlock ChatGPT's capabilities in scenarios with limited training data.The study explores ChatGPT's potential in personalized recommendations through "prompt tuning." Experiments on real-world datasets demonstrate ChatGPT's superior performance compared to state-of-the-art recommendation systems in terms of accuracy and efficiency. The paper also acknowledges ChatGPT's limitations and suggests future research directions. This work illuminates the potential of large language models in personalized recommendations, emphasizing the significance of prompt tuning in challenging scenarios.

Bhavya, Kavya, etc. using NLP. (2022) [2] proposed a model for CV analysis. The algorithm developed for skill and CV comparison in this project provides incredibly accurate results. Skills required for a particular job are first provided as input and then preprocessed or sanitized by removing redundant information, large spaces, and punctuation. It then uses the Word2Vec algorithm from the Genism library to find word embeddings within bigrams of common words in the skill corpus. Candidate resumes are stored in a folder and extracted one by one using the PyPDF library and returned as a string sequence. The strings are first pre-processed and then further processed to create the candidate's profile. The array (taken from word2vec) is matched against the extracted text using Spacy's phrase matcher to generate a candidate profile and display it as a graph.

Gunawardhana, **one of Stefan and others. (2022) [3]** allows users to upload their resumes to the site, then analyzes the resumes and searches for keywords. Then use the job API to search for those values using those keywords.

According to **Vukadin, Davor, et al. (2021) [4]**, significant information from an unstructured multilingual CV can be recovered by using two NLP algorithms to choose important document elements and related specific information at the low hierarchy level. Their strategy takes advantage of the BERT language model's transformer architecture and the encoder component's multilingual implementation. The associated detailed information, such as names, dates, organisations, positions, university degrees, individual abilities, and their (self-assessed) competency levels, are extracted and categorised at a lower hierarchical level of the model. The model takes the pertinent bits of a document and extracts them (personal information, education, previous employment, and skills).

Bhor, Shubham, et al. (2021) [5] transform resumes uploaded in any format—.txt,.pdf,.doc,.docx,.odt, etc.—into a single text format using optical character recognition (OCR). After receiving these converted CVs, the Natural Language Processing (NLP) engine processes them using a number of NLP techniques, such as tokenization, lexical analysis, syntactic analysis, and named entity recognition (NER). The required keywords and extracted entities are then contrasted in order to rank the resumes using the ranking algorithm. The outcomes are then displayed as pie charts and bar charts.

The system proposed in **Satheesh, K., et al.(2020) [6]** automatically ranks resumes, saving recruiters time in carefully reviewing job descriptions. Using her non-traditional NER model,

it automatically extracts specified entities from resumes to simplify and streamline the hiring process. A graph is then created showing the scores for each resume. Based on scores, recruiters select desired candidates instead of sifting through piles of unqualified applicant resumes.

Kelkar, Shedbal et al. (2020) [7] introduces a useful business her recommendation system that uses text mining and machine learning tools to help recruiters find the best candidates for specific positions. When applicants upload their resumes, these resumes are ranked according to the hiring organization's requirements. Companies can use rankings to find the best candidates. One of the key steps in this process is keyword research. Preprocess the input text received using NLP techniques such as POS tagging, chunking, tokenization, lexical analysis, and sentiment analysis. MySQL connector is used to retrieve data from the keyword table and its associated values from the database. Keywords extracted from the resume are compared with these words from the database using keyword matching algorithms in Python. To rank resumes and rank candidates based on their scores, all keyword values are summed to generate score.

T(2018) Pham Van Long et al.[8] This paper describes a parsing application built for resumes (or CVs) received in various formats such as doc, docx, pdf, txt. An automated resume information extraction system can find and process these resumes. The extracted data is stored in a database as structured information and can be used in various areas such as names, phone numbers, email addresses, qualifications, experience and skills. The system is formed in four steps.

Soumya, Gayatri et al. (2018) [9] This paper introduces the concept of a personal chatbot assistant that also acts as a resume parser. Users can use this chatbot to schedule meetings and add them to their calendars. It's an automated chatbot resume. This chatbot helps recruiters know who you are in addition to all your credentials and personal information. They used a pattern matching technique that compares the text entered by the user against all the text stored in the database. After running the comparison detection, the corresponding output is generated.

Sanyal, Satyaki, et al. 2017 [10] developed a model that uses NLP techniques including tokenization, lemmatization, stemming, parts of speech tagging, lexical analysis, syntactic analysis, and semantic analysis to interpret information from unstructured resumes and convert it to a structured JSON format. The structured resumes are parsed, then saved in a NoSQL database where AI is used to score the resumes and forecast which applicant is most qualified for the position, giving the recruiting process credibility.

An online resume analysis system developed by **Chandola, Divyanshu et al. (2015) [11]** proposes the use of text analysis. The authors propose a text analysis that evaluates resumes based on their content. Using this technique can help employers identify the best candidate for the position's requirements. The Python-based keyword matching algorithm is simplified by a MySQL connector that retrieves the data required for matching from the keyword table and its associated values to create the knowledge base. Text segmentation, name entity recognition with rule-based algorithms, name lookup with deep neural networks, and text normalization. There are over 81% F1 NERs, making encouraging progress.

Chapter 3: Methodology/Techniques Used

Tokenization: In the English language, a sentence is made up of several significant words. These words are made up of tokens that combine to form a sentence. Tokenization is the process of dividing the complete raw text into manageable chunks that can be utilised for processing.
The proposed system uses several instances of tokenization:

- Sentence tokenization: SpaCy's built-in sentence tokenizer is used to break the resume text into a list of individual sentences. This is crucial because many parsing and analysis tasks require working with sentences as separate units.
- Skill extraction: Tokenization is used to split the resume text into individual tokens. These tokens are then processed to extract both individual skills and bigrams (pairs of adjacent words that represent skills, like Machine learning) from the resume.
- Email address extraction: Tokenization is used to split the resume text into individual tokens, which are then searched for email patterns to extract email addresses.

Test extraction using OCR: In the proposed system**, pdf2image** library is used to convert the pages of the PDF document into images after which the images after which **easyocr** library is used for Optical Character Recognition (OCR) to extract text from these images.

Named Entity Recognition (NER): Named Entity Recognition (NER) is a natural language processing technique that involves identifying and classifying named entities (real-world objects such as names of persons, organizations, locations, dates, numerical values, etc.) in text. The goal of NER is to locate and categorize these entities to provide context and extract structured information from unstructured text.

Lemmatization: Lemmatization is a linguistic and natural language processing (NLP) technique used to reduce words to their base or dictionary form, known as the "lemma." The lemma represents the canonical or normalized form of a word, regardless of its inflections or variations, which helps in normalizing the text and improving the accuracy of various NLP tasks.In the proposed system, lemmatization is mainly used in skill extraction. The resume text is processed using spacy’s NLP pipeline. Each token is iterated over to apply lemmatization on each of them, to get the base form of the word. This is important mainly for skill extraction because skills can be mentioned in different forms (e.g., "programming," "programmer," "programmed"). Lemmatizing the tokens helps match them to the skills listed in the skills CSV file, increasing the chances of accurate skill extraction.

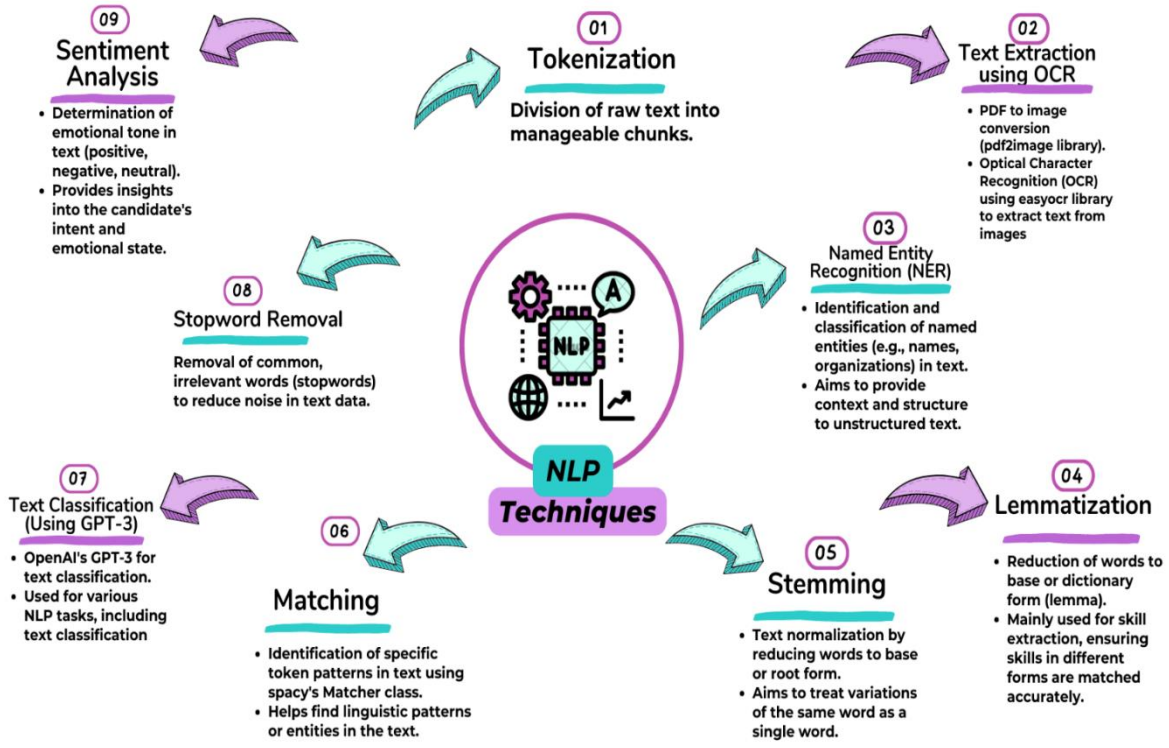
Stemming: Stemming is a text normalization technique in Natural Language Processing (NLP) that reduces words to their base or root form by removing suffixes or prefixes. The purpose of stemming is to simplify words so that variations of the same word can be treated as a single word, even if they have different forms.

Matching: Matching is used to identify specific patterns or sequences of tokens (words) in the input text that correspond to certain predefined patterns or criteria. It allows to find instances of linguistic patterns or entities within the text. Matcher class of spacy library is used to implement this.

Text Classification: OpenAI's GPT-3 is used for text classification. GPT-3 is a powerful language model that can generate human-like text and is often used for a wide range of NLP tasks, including text classification.

Stopword Removal: Stopwords are commonly occurring words like "and" "the," "is," etc., that are often removed to reduce noise in text data. The NLTK library is used to download stopwords, and they are used to filter out potentially irrelevant words.

Sentiment Analysis: This is a natural language processing technique that involves determining the emotional tone or sentiment expressed in a piece of text, whether it's positive, negative, or neutral. It aims to understand and quantify the emotions, opinions, attitudes, and feelings conveyed in the text. Sentiment analysis can provide an additional layer of information about the resumes that are being analysed. Sentiment analysis can help gain insights into the candidate's intent and emotional state. Positive sentiments might indicate confidence, enthusiasm, or achievements, while negative sentiments could indicate frustration, challenges, or areas for improvement.



Chapter 4: Tools Used

Libraries for Natural Language Processing (NLP):

- nltk (Natural Language Toolkit): Used for a wide range of text analysis tasks, including tokenization, part-of-speech tagging, named entity recognition, parsing, stemming, and sentiment analysis.
- spacy: An open-source NLP library known for its speed and efficiency, used for various text processing and analysis tasks. It also includes the Matcher class for pattern matching.
- easyocr: A Python library designed for Optical Character Recognition (OCR), enabling the extraction of text from images, including handwritten or printed text in multiple languages and fonts.
- openai: A Python library that provides access to OpenAI's language models and APIs, allowing you to leverage pre-trained models for various NLP tasks, including text classification

OpenAI's GPT-3:

- Used for text classification tasks, including text classification.

External Libraries:

- Snowball Stemmer (from nltk): Utilized for stemming, which is a text normalization technique.
- NLTK (Natural Language Toolkit): Used to download stopwords for stopword removal, a process to reduce noise in text data.
- pdf2image: Used to convert PDF pages into images, which are then processed by the EasyOCR library for text extraction using OCR.

Chapter 5: Modules Implemented and Output

Module 1: Resume Text Extraction using OCR (Optical Character Recognition)
Description:

The OCR module is responsible for converting scanned or image-based resumes, typically in PDF format, into machine-readable text. This module is crucial for extracting textual content from resumes uploaded by job seekers.

Algorithm:

1. PDF to Image Conversion:
   The module begins by converting the pages of the PDF resume into images. This step is accomplished using the pdf2image library.

2. Text Extraction with OCR:
   Once the PDF pages are converted to images, the easyocr library is employed for Optical Character Recognition (OCR).
   The OCR algorithm identifies and extracts text from these images, transforming the visual content into machine-readable text.

3. Output:
   The output of this module is the textual content extracted from the resume images, which can be further processed and analyzed using NLP techniques.

Module 2: Named Entity Recognition (NER) for Structured Information Extraction
Description:

The NER module focuses on identifying and classifying named entities within the textual content of resumes. Named entities can include names of persons, organizations, locations, dates, numerical values, and more. This module helps in structuring and categorizing information from unstructured text.

Algorithm:

Text Preprocessing:

The textual content extracted from resumes is preprocessed to remove any noise or irrelevant information.

Named Entity Recognition:

The spacy library, known for its NER capabilities, is used to perform NER on the preprocessed text.
The NER algorithm identifies and categorizes entities by assigning them labels (e.g., PERSON, ORGANIZATION, DATE).

Output:

The output of this module is a structured representation of the extracted named entities, along with their respective labels. This structured data can be used for various purposes, such as populating candidate profiles.

Module 3: Chatbot for User Interaction

Description:
The chatbot module is responsible for facilitating interactions between recruiters (users) and the system. It serves as the user interface through which recruiters can submit their job requirements, receive recommendations, and provide feedback.

Algorithm:

1. User Input Processing:
    - The chatbot processes user inputs, which include job category selection, feedback submission, and other queries related to the recruitment process.

2. NLP-Based Intent Recognition:
    - Utilize natural language processing (NLP) techniques to understand the intent behind user inputs.
    - Recognize keywords, phrases, and context to determine what actions the user wants to perform.

3. Backend Communication:
    - Communicate with the backend of the system to initiate relevant actions based on user intent.
    - For example, if a recruiter selects a job category, the chatbot communicates this information to the recommendation module for candidate retrieval.

4. Response Generation:
    - Generate responses to user queries or actions, providing information and recommendations.
    - Present recommendations to recruiters in a user-friendly and informative manner.

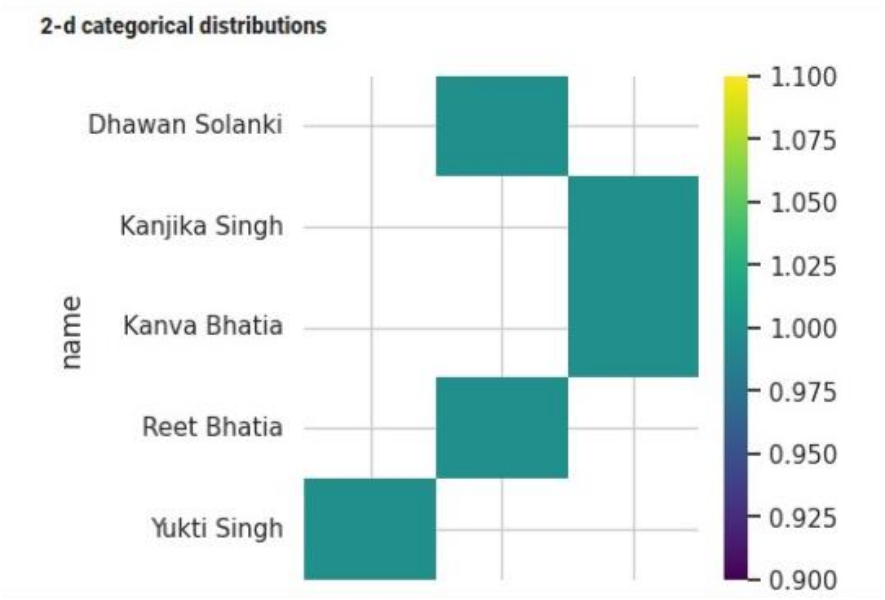
5. Feedback Collection:
    - Enable recruiters to provide feedback on recommended candidates or the system's performance.
    - Collect and process this feedback for system improvement.

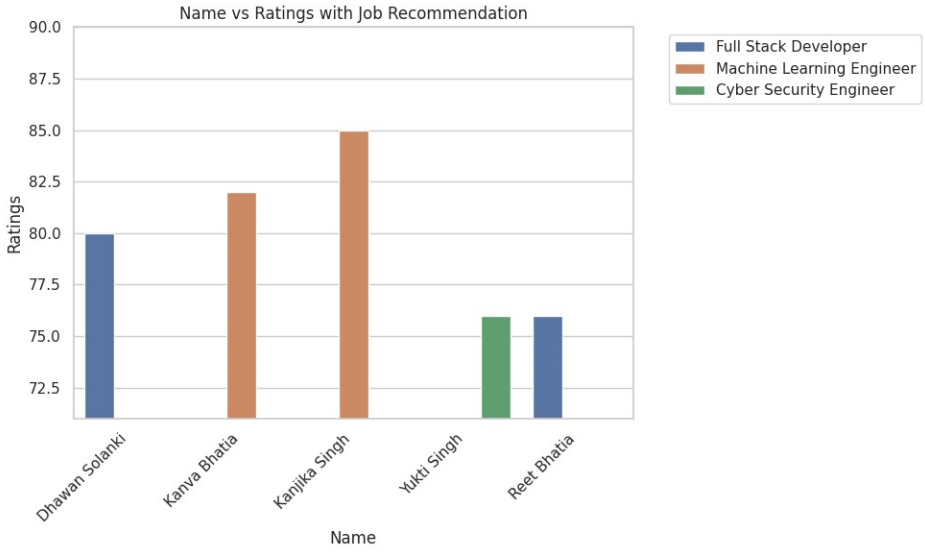
6. Output:
    - The chatbot module provides a user-friendly interface for recruiters to interact with the system. It processes user inputs, communicates with other modules, generates responses, and collects feedback.

RESULTS:

The model proposed above segregates the candidates resume into suitable job categories as soon as the applicant uploads their resume.

The following graphs show the results:





Chapter 6: Learnings and Takeaways from the Study

During the technical internship, I had the opportunity to work on an exciting project that leveraged Natural Language Processing (NLP) techniques for revolutionizing the recruitment process. This experience has been incredibly enriching, and we've gained several valuable insights and skills along the way.

1. **NLP Proficiency:** One of the most significant takeaways from this internship is a deepened proficiency in NLP. We've had hands-on experience with various NLP methods, from tokenization to named entity recognition, and learned how to apply them effectively to analyze and process resumes.

2. **Handling Unstructured Data:** Working with large volumes of unstructured data in the form of resumes has been a valuable learning experience. We've become adept at preprocessing and structuring data efficiently, a skill that is vital across various data-driven projects.

3. **Collaboration:** Collaborating with a diverse team on a complex project has enhanced teamwork, communication, and project management skills. It has reinforced the importance of effective collaboration in achieving project goals.

4. **Continuous Learning:** NLP and technology are rapidly evolving fields, and this internship has underscored the importance of continuous learning and staying updated with the latest advancements.
5. **Communication Skills**: Working on this group project has taught me the importance of effective communication. I've learned how to express my ideas clearly, actively listen to my team members, and resolve conflicts through open and honest communication.
6. **Teamwork:** Through this project, I've had the opportunity to experience the power of teamwork. I've learned how to collaborate with others, coordinate our efforts, delegate tasks, and work harmoniously towards our common goal.

REFERENCES AND ANNEXURES

[1] Sunhao Dai, Ninglu Shao "Uncovering ChatGPT's Capabilities in Recommender Systems", Gaoling School of Artificial Intelligence, Renmin University of China,August 2023.

[2] "CV Parsing Using NLP" K. Bhavya Sai 1, G. Kavya Sree2, S. Sai Soundarya3, C. Sai Pranathi4, Y. Durga Bhargavi5, Volume 2, Issue 1, April 2022.

[3] Gunawardana, Stephan. "Resume Parser and Job Search." (2022).

[4] Vukadin, Davor, et al. "Information extraction from free-form CV documents in multiple languages." IEEE Access 9 (2021): 84559-84575.

[5] Bhor, Shubham, et al. "Resume parser using natural language processing techniques." Int. J. Res. Eng. Sci 9.6 (2021).

[6] Satheesh, K., et al. "Resume Ranking based on Job Description using SpaCy NER model." International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395–0056 7.05 (2020).

[7] "Resume Analyzer Using Text Processing" B.Kelkar1, R.Shedbale2, D.Khade3, P.Pol4, A.Damame51.(2020)

[8] Pham Van, Long, Sang Vu Ngoc, and Vinh Nguyen Van. "Study of Information Extraction in Resume." Conference, 2018.

[9] International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 20 (2018) pp. 14644-14649 "Chatbot as a Personal Assistant"

[10] Sanyal, Satyaki, et al. "Resume parser with natural language processing." International Journal of Engineering Science 4484 (2017).

[11] Chandola, Divyanshu, et al. "Online resume parsing system using text analytics." Journal of Multi-Disciplinary Engineering Technologies 9 (2015).