NLP BASED RESUME SCREENING SYSTEM

Submitted for the summer internship
On

MACHINE LEARNING AND PYTHON
(5-06-2023 to 24-07-2023)

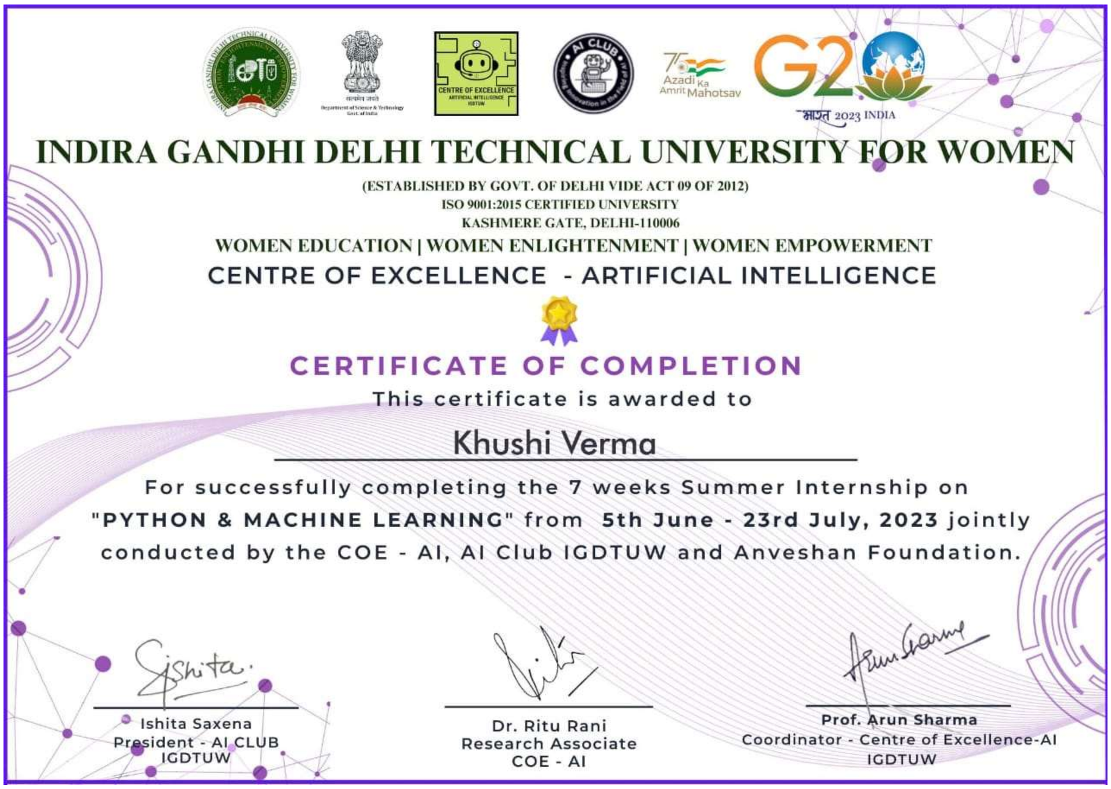
ORGANIZED BY
COE-AI
AI Club IGDTUW
Anveshan Foundation
By

Khushi Verma(092)
Mahak Yadav(100)
Mahika Malik(101)
CSE 2ndYear

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

INDEX

CERTIFICATE



INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

(ESTABLISHED BY GOVT. OF DELHI VIDE ACT 09 OF 2012)
ISO 9001:2015 CERTIFIED UNIVERSITY
KASHMERE GATE, DELHI-110006

WOMEN EDUCATION | WOMEN ENLIGHTENMENT | WOMEN EMPOWERMENT

CENTRE OF EXCELLENCE - ARTIFICIAL INTELLIGENCE

CERTIFICATE OF COMPLETION

This certificate is awarded to

Khushi Verma

For successfully completing the 7 weeks Summer Internship on "PYTHON & MACHINE LEARNING" from 5th June - 23rd July, 2023 jointly conducted by the COE - AI, AI Club IGDTUW and Anveshan Foundation.

Ishita Saxena
President - AI CLUB
IGDTUW

Dr. Ritu Rani
Research Associate
COE - AI

Prof. Arun Sharma
Coordinator - Centre of Excellence-AI
IGDTUW

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

(ESTABLISHED BY GOVT. OF DELHI VIDE ACT 09 OF 2012)
ISO 9001:2015 CERTIFIED UNIVERSITY
KASHMERE GATE, DELHI-110006

WOMEN EDUCATION | WOMEN ENLIGHTENMENT | WOMEN EMPOWERMENT

CENTRE OF EXCELLENCE - ARTIFICIAL INTELLIGENCE

CERTIFICATE OF COMPLETION

This certificate is awarded to

Mahika Malik

For successfully completing the 7 weeks Summer Internship on "**PYTHON & MACHINE LEARNING**" from **5th June - 23rd July, 2023** jointly conducted by the COE - AI, AI Club IGDTUW and Anveshan Foundation.

Ishita Saxena
President - AI CLUB
IGDTUW

Dr. Ritu Rani
Research Associate
COE - AI

Prof. Arun Sharma
Coordinator - Centre of Excellence-AI
IGDTUW

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

(ESTABLISHED BY GOVT. OF DELHI VIDE ACT 09 OF 2012)
ISO 9001:2015 CERTIFIED UNIVERSITY
KASHMERE GATE, DELHI-110006

WOMEN EDUCATION | WOMEN ENLIGHTENMENT | WOMEN EMPOWERMENT

CENTRE OF EXCELLENCE - ARTIFICIAL INTELLIGENCE

CERTIFICATE OF COMPLETION

This certificate is awarded to

Mahak Yadav

For successfully completing the 7 weeks Summer Internship on "**PYTHON & MACHINE LEARNING**" from **5th June - 23rd July, 2023** jointly conducted by the COE - AI, AI Club IGDTUW and Anveshan Foundation.

Ishita Saxena
President - AI CLUB
IGDTUW

Dr. Ritu Rani
Research Associate
COE - AI

Prof. Arun Sharma
Coordinator - Centre of Excellence-AI
IGDTUW

ACKNOWLEDGEMENT

I would like to thank my internship supervisor at External Company, Mr. Mohit Uniyal, Coding Minutes, for closely working with me and supervising me, hosting regular meetings (twice a week) with me to see my progress. I would like to thank Dr. Ritu, concerned faculty, IGDTUW for helping out with my work. I would also like to thank Ishita Saxena, president of AI Club, IGDTUW for giving me an opportunity to expand my knowledge through this internship. I would also like to thank Anveshan Foundation, IGDTUW , the internship coordinator for the Artificial Intelligence Internship/Course, for helping me find the internal assessor, and also helping me with all the documentation and administration part of the internship.

ABSTRACT

Technology has been advancing at an exponential rate and so are modern businesses. Everything is powered by tools and machines reducing human dependency to a minimal scale. In the rapidly evolving job market, Every year, a substantial number of candidates actively pursue career opportunities in a diverse range of industries, for which the first crucial step is a well-structured resume. With such large numbers of entries, manual resume parsing has become an impractical and unfeasible task which necessitates the adoption of AI-powered parsing and analysis systems.

In this paper, we will discuss a resume parsing and skill-gap analysis system powered by NLP techniques and ML models that'll extract data from unstructured resumes employing cutting-edge Natural Language Processing (NLP) techniques followed by a comprehensive analysis. By leveraging the power of NLP, the system efficiently interprets and organises candidate information, and identifies key strengths and areas for improvement via machine learning. Through this analysis, our model suggests potential enhancements to the resume identifying skills that can be upgraded to boost the candidate's profile.

Moreover, the system offers more than just resume analysis but also provides valuable insights through ML algorithms recommending suitable job profiles based on the similarities found between candidate's qualifications, experience, and skill set and the job requirements. With our model's integrated capabilities, recruiters and job seekers can make informed decisions, ensuring a better match between candidates and job opportunities, ultimately leading to enhanced career prospects and saving valuable hours of effort.

INTRODUCTION

In today's dynamic and technologically advancing business world, millions of individuals actively search for jobs that would align with their skill set and aspirations. Well framed resumes are the primary requisite for the purpose, different organisations might even have different format requirements which needs to be strictly followed.
Despite this, there is still a chance of not being selected for further stages of the process or not getting the desired role in the company. Gap between the job requirements and the candidates qualifications also known as the "skill gap" is the primary reason for this. The problem at hand is to develop an integrated resume parsing and skill gap analysis system that will allow the users , which can be employers as well as applicants, to comprehend the job demands and carefully tailor their resumes to highlight relevant skill set, and simplify employers' job making the whole selection process more accurate and productive, Moreover, it should recommend job roles according to the input skill set allowing the candidates to make careful decisions for their career.

Resume parsing is a computerised procedure involving extraction of relevant data from unstructured input resumes and curriculum vitae (CVs) employing natural language processing and machine learning algorithms, for the evaluation of masses of applicants. This time efficient process significantly simplifies and speeds up the screening task while enhancing the accuracy.
the technical underpinnings of resume parsing involve a combination of NLP, machine learning, data structuring techniques, and continuous learning mechanisms. These elements collectively contribute to the efficiency, accuracy, and scalability of the resume parsing process.

Skill gap analysis is the process of careful examination of a candidates' skills, qualifications, education and background etc. to compare them with the requisites of the job role at hand and identify the similarities between the two, based on which the candidate can draw a conscious conclusion for their career choice or identify the need of training in specific areas.
The goal is to automate and streamline a large number of resumes and identify qualified candidates.The automation of skill gap analysis not only expedites the identification of qualified candidates but also provides a comprehensive and dynamic approach to career development. The integration of advanced technologies ensures a more personalized and forward-looking assessment, benefiting both candidates and organizations in their pursuit of talent and skill alignment.

OBJECTIVES

- Design and implement a powerful resume parsing system capable of extracting key information, such as skills, education, and work experience, from a variety of resume formats

- Train a Classification machine learning model to accurately classify and categorize resumes based on relevant job skills, experience, and qualifications
.
- Fine-tune the machine learning model to achieve a high level of accuracy in classifying resumes, ensuring that the system can effectively match candidates with appropriate job postings.

- Ensure that the resume parser can handle a diverse range of resume formats, including both digital and scanned documents, to maximize its applicability and usability.

- Connect the system to a comprehensive job postings database, ensuring that the model outputs relevant job opportunities that closely align with the skills and qualifications extracted from the input resumes.

METHODOLOGY AND IMPLEMENTATION

Developing a model that matches resumes to job postings and provides suggestions for improving resumes based on job requirements involves a multi-step process. Below is a step-by-step guide to help you achieve this:

1. Data Preprocessing:

a. Data Cleaning:
- Ensure both datasets are clean and free of errors.
- Handle missing values, if any.

b. Text Preprocessing:
- Tokenize the text in both datasets.
- Remove stop words, punctuation, and other irrelevant characters.
- Convert text to lowercase for consistency.

2. Data Integration:

Combine the job postings and resume datasets based on the common attribute (e.g., job title).

3. Feature Engineering:

a. TF-IDF (Term Frequency-Inverse Document Frequency):
- Convert text data into numerical vectors using TF-IDF.
- This will represent the importance of each word in the context of the entire dataset.

4. Model Development:

a. Build a Recommendation Model:
- Use a similarity-based model (e.g., cosine similarity) to find the similarity between resumes and job postings.
- Train the model on the integrated dataset.

b. Machine Learning Model:
- Train a classifier to predict the relevance of a job posting to a resume.
- Use features like TF-IDF vectors and other relevant attributes.

5. Upgrade Suggestions:

a. Extract Keywords:
- Analyze job requirements to extract key skills and qualifications.

b. Resume Analysis:
- Compare the extracted keywords with the content of the resume to identify gaps.

c. Recommendation System:
- Provide suggestions on how to upgrade the resume based on the identified gaps.
- Suggest additional skills or modifications to existing content.

6. User Interface:

a. Develop a User-Friendly Interface:
- Create a user interface that accepts resumes and displays relevant job postings along with upgrade suggestions.
- Use a web application or any other suitable platform.

7. Testing and Evaluation:

a. Split the Dataset:
- Split the dataset into training and testing sets for model evaluation.

b. Evaluation Metrics:
- Use metrics like precision, recall, and F1-score to evaluate the model's performance.

8. Model Deployment:

a. Deploy the Model:
- Once satisfied with the model's performance, deploy it to a production environment.

b. Continuous Improvement:
- Monitor user feedback and continuously improve the model based on real-world usage.

9. User Feedback:

Encourage users to provide feedback on the relevance of job suggestions and the helpfulness of upgrade recommendations. Use this feedback to improve the model iteratively.

Note:
This is a high-level overview, and the specific tools and libraries you use will depend on your preferences and the nature of your data. It's recommended to use popular machine learning libraries such as scikit-learn or TensorFlow for model development. Also, consider using natural language processing (NLP) libraries like NLTK or spaCy for text processing tasks.

LITERATURE SURVEY

NLP in Resume Parsing

Gupta, N., & Jain, A. (2019): "Resume Parsing using Natural Language Processing."

Objective:

This study likely explores the integration of Natural Language Processing (NLP) techniques in the process of resume parsing. NLP is employed to understand and extract meaningful information from unstructured text data in resumes.

Key Points:

- **Tokenization:** The paper may discuss the use of tokenization to break down resumes into individual words or phrases. This step is crucial for subsequent analysis.
- **Entity Recognition:** NLP techniques, such as Named Entity Recognition (NER), may be applied to identify entities like names, addresses, skills, and work experience in resumes.
- **Semantic Analysis:** The study may highlight how semantic analysis is utilised to understand the context and relationships between different elements in a resume, providing a more nuanced interpretation.
- **Coreference Resolution:** Addressing coreference (e.g., resolving "he" or "she" to a specific person) could be discussed, enhancing the cohesiveness of extracted information.
- **Contextual Understanding:** NLP may be employed to understand the context of phrases and sentences, aiding in the accurate extraction of relevant details.

Significance:

Integrating NLP in resume parsing enhances the system's ability to comprehend and interpret the content of resumes more accurately, improving the overall parsing efficiency.

Evaluation Metrics:

Ahmad, S., & Sarwar, S. (2019): "Evaluation Metrics for Resume Parsing Systems."

Objective:

This paper likely focuses on the evaluation metrics used to assess the performance of resume parsing systems. Choosing appropriate metrics is crucial for gauging the effectiveness and reliability of these systems.

Key Metrics:

- **Precision:** Precision measures the accuracy of the system in correctly identifying relevant information. It is the ratio of true positives to the sum of true positives and false positives.

- **Recall:** Recall, or sensitivity, measures the system's ability to capture all relevant information. It is the ratio of true positives to the sum of true positives and false negatives.
- **F1 Score:** The F1 score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives.
- **Accuracy:** The overall correctness of the system, measured as the ratio of true positives and true negatives to the total number of instances.

Significance:
These metrics collectively provide a comprehensive assessment of the resume parsing system's performance. Precision and recall are particularly important as they address the trade-off between false positives and false negatives.

RESULT AND CONCLUSION

Assessing the performance of a machine learning model, the accuracy score is a commonly used metric, and Scikit-learn's accuracy_score function provides a convenient way to quantify the model's correctness. The accuracy score is calculated by dividing the number of correct predictions by the total number of predictions, providing a percentage that represents the model's overall correctness.

Is crucial to interpret the accuracy score in the context of the dataset and the nature of the problem being addressed. In cases of imbalanced classes, where one class significantly outnumbers the others, accuracy alone may not provide a comprehensive evaluation. In such scenarios, additional metrics like precision, recall, and F1 score might be considered to gain a more nuanced understanding of the model's performance, especially with respect to false positives and false negatives.

The accuracy of our model is 54.545454545454, thus indicating that the model correctly predicted the outcome of 54% of the instances in the dataset.A higher accuracy score generally suggests better model performance, but it's important to consider the specific characteristics of the dataset and the problem at hand.

The 2 month Machine Learning internship has given us a glimpse of what all lies in the depth of the field named ,Machine Learning. But being short on time we couldn’t magically turn to be the experts and since it is our first project so it is outcome of an untiring first attempt of designing machine learning model.

CONCLUSION AND FUTURE SCOPE

In conclusion, the resume screening machine learning project has provided valuable insights into the challenges and opportunities associated with automating the hiring process. The model's performance, as evaluated through metrics such as accuracy and possibly precision-recall, has shed light on its current efficacy. However, it is crucial to acknowledge the limitations and potential biases that may exist in the model, such as underrepresentation or overemphasis on certain qualifications. The project's success in expediting the initial stages of candidate selection underscores the potential for innovation in the recruitment domain.

Looking ahead, there are several avenues for future exploration and enhancement. Firstly, the model could benefit from continuous training and updates to adapt to evolving job market trends and changing recruitment criteria. Additionally, incorporating natural language processing (NLP) techniques to better understand and contextualise the content of resumes could enhance the model's comprehension of nuanced qualifications and soft skills. Collaboration with industry experts and HR professionals would provide valuable insights for refining the model's criteria and aligning it with real-world hiring practices.

Moreover, addressing issues of fairness and bias in the model should be a priority. Implementing fairness-aware algorithms and conducting regular audits to ensure equitable treatment of diverse candidates will be essential. Transparency in the decision-making process of the model is also crucial for building trust among users, both within the organisation and among job applicants.

In conclusion, while the resume screening machine learning project marks a significant step forward in automating recruitment processes, ongoing efforts in model refinement, ethical considerations, and collaboration with stakeholders will be key to unlocking its full potential and ensuring a fair and effective hiring tool for the future.