



SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING & TECHNOLOGY | SCHOOL
OF COMPUTING
DEPARTMENT OF COMPUTING TECHNOLOGIES

18CSC305J ARTIFICIAL INTELLIGENCE – MINI PROJECT

AI RESUME PARSER

RA2111003010596 - PULKIT SHRINGI

RA2111003010608 - PARTH AGARWAL

RA2111003010612 - BRYAN ABRAHAM

ABSTRACT

This project addresses the complexity of analyzing semi-structured documents, especially resumes. It proposes a Resume Parser system with tailored algorithms for block analysis, enhancing accuracy and efficiency. The system automates reading, analysis, retrieval, and storage of resume information, demonstrating practical viability and offering guidance for future research in similar domains.



INTRODUCTION

This project focuses on classifying text into structured, free style, and semi-structured categories based on specific criteria in documents. Semi-structured documents, exemplified by resumes and scientific papers, exhibit distinctive characteristics, making them suitable for investigation. By developing a Resume Parser system and employing various algorithms, this research aims to enhance information retrieval from semi-structured documents, contributing to advancements in document analysis and processing technologies.

CHALLENGES/MOTIVATION

The primary challenge in developing a resume parser stems from the nature of resumes as semi-structured documents that lack uniform formatting. Resumes often contain diverse expressions and complex structures which makes it difficult for traditional parsing methods to accurately and efficiently extract and categorize information.

The motivation behind developing a resume parser is to automate and enhance the recruitment process by using robust algorithms to accurately extract and categorize information, thereby increasing efficiency and decision-making quality in hiring.

PROBLEM STATEMENT

The challenge lies in effectively parsing semi-structured documents, such as resumes, which exhibit discrete items with independent semantics and non-punctuation segmentation. Current document analysis methods often struggle to accurately extract and categorize information from these documents due to their complex structure and varied expression styles. Developing robust algorithms and a systematic approach to parsing such documents is crucial for improving information retrieval and automating tasks like resume processing for recruitment platforms.

LITERATURE SURVEY

Authors	Title	Dataset	Methods	Remarks
Chuang Zhang Ming Wu Chun-Guang Li Bo Xiao	Resume Parser: Semi-structured Chinese Document Analysis	No Specific Dataset used	Resume parsing method involves iterative segmentation and categorization using feedback loop algorithms for comprehensive item identification and convergence.	Semi-structured resume information is extracted using regular expressions, text classification, and fuzzy matching for enhanced accuracy and broader applicability.
A. K. Maheshwaran P. Saurav Arvind G. Vijayaragavan	On-Demand Job-Based Recruitment For Organisations Using Artificial Intelligence	The IBM HR employee Attrition dataset was generated by IBM data scientists	Method involves HR data analysis, ML algorithms, resume parsing, NLP/AI for screening, and AI recommendations.	The study presents a comprehensive system for HR tasks, including attrition prediction, resume parsing, and candidate matching.

EXISTING MODEL

RChilli Resume Parser is an existing versatile solution designed to streamline the recruitment process by extracting structured data from resumes in various formats, including PDF, DOC, DOCX, and HTML. Its key features include format agnosticism, multi-language support, and customization options, making it adaptable to diverse candidate pools and specific business requirements. By automating the parsing process, RChilli helps save time for recruiters, improves the candidate experience, and enables data-driven decision-making in recruitment. With seamless integration capabilities and a focus on accuracy and efficiency, RChilli Resume Parser offers businesses a reliable tool to enhance their hiring efforts and make informed hiring decisions.



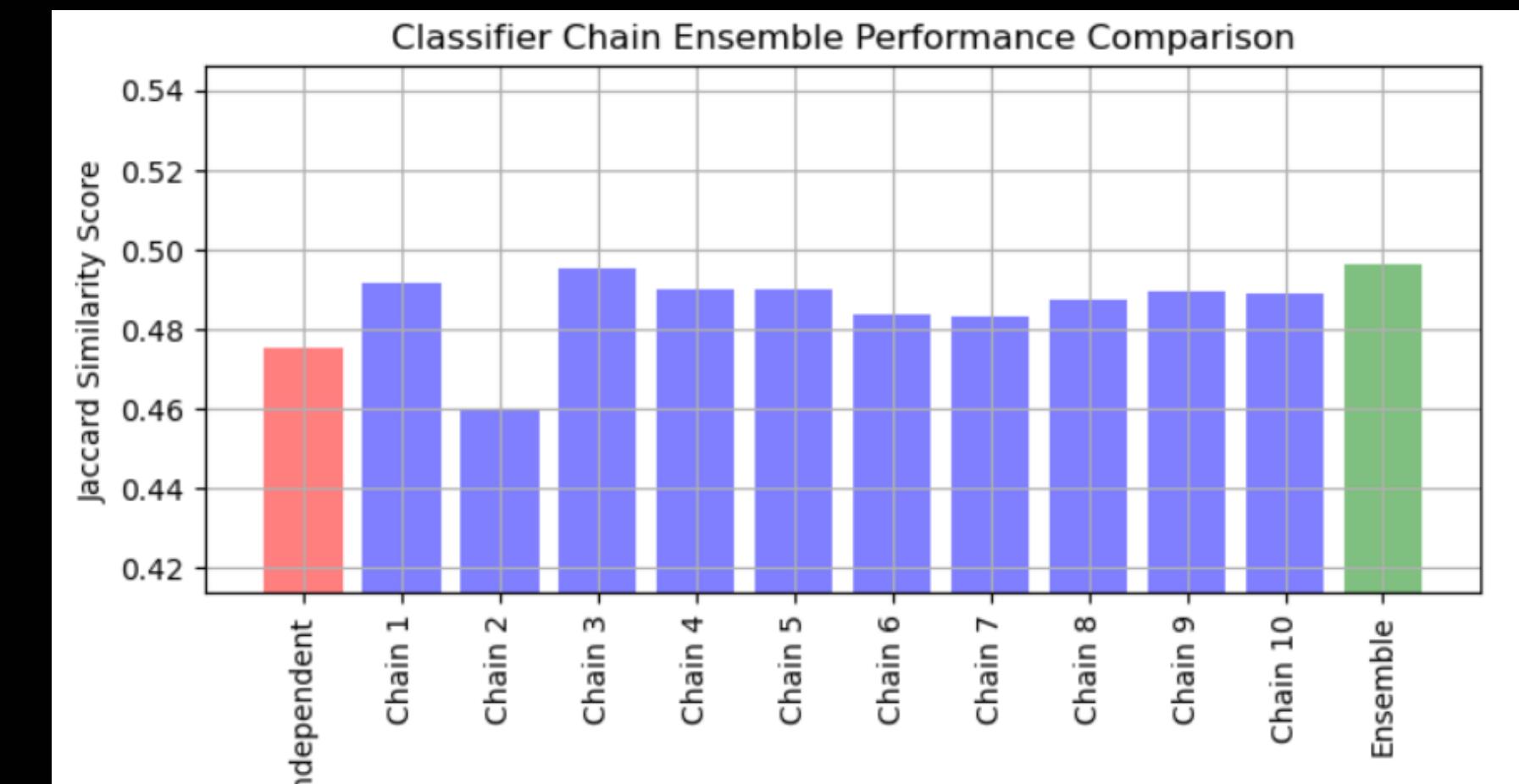
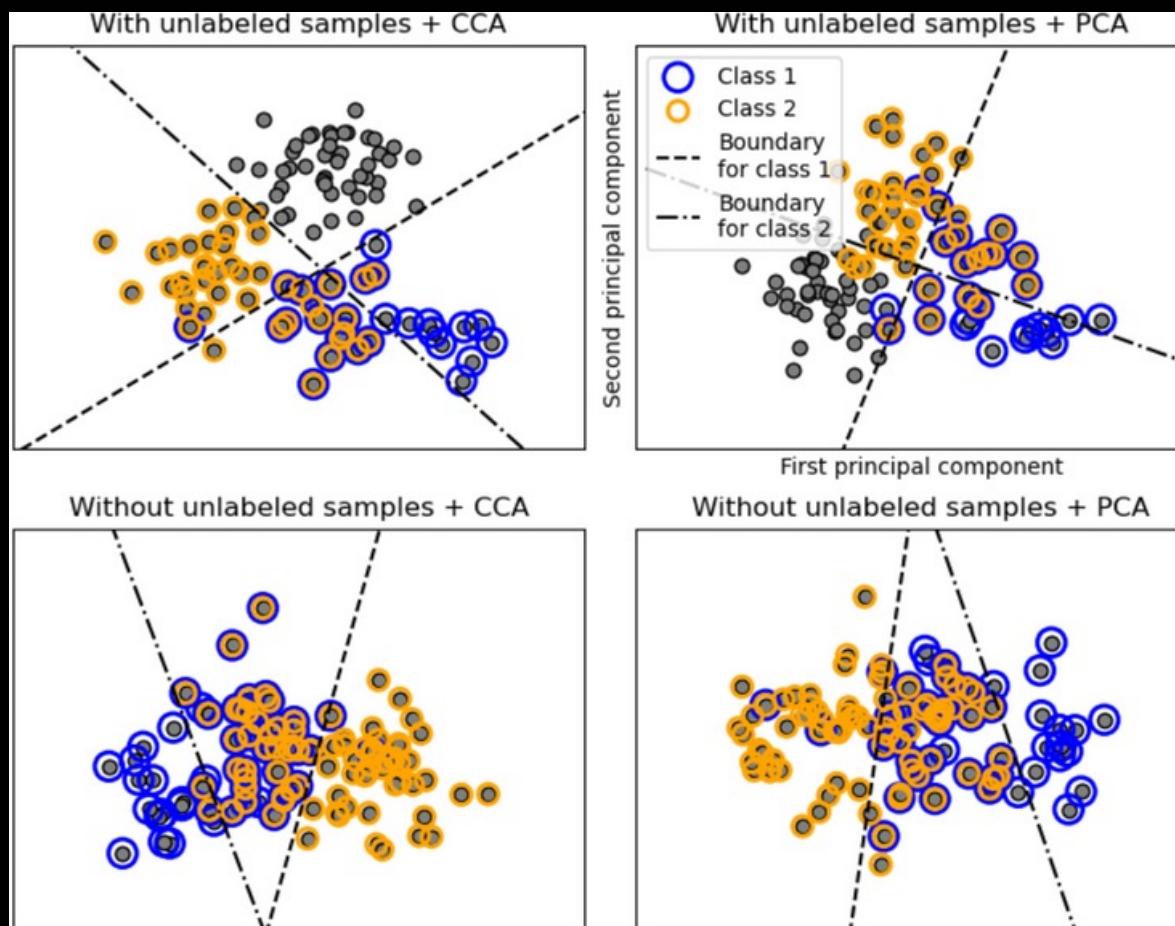
PROPOSED MODEL

This resume parser project, built with Python and employing the K-Nearest Neighbors (KNN) algorithm, is hosted on Streamlit. It offers a simple interface for users to upload resumes and instantly view predicted category. Utilizing machine learning, the algorithm categorizes resume data into sections like skills and work experience. With its user-friendly design, the application streamlines the resume screening process for recruiters, saving time and effort. By providing quick access to relevant information, it enhances efficiency in the hiring process, empowering recruiters to make informed decisions.



ARCHITECTURE OF MODEL

Our resume parsing system utilizes a powerful blend of NLP and machine learning, employing the one-vs-rest algorithm alongside TF-IDF vectorization. This combination enables precise extraction of candidate details and skills from resumes, facilitating streamlined analysis for faster and more informed hiring decisions. By leveraging these advanced techniques, we ensure efficient handling of diverse resume formats and extraction of invaluable insights from candidate profiles.



PROTOTYPE

```
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75

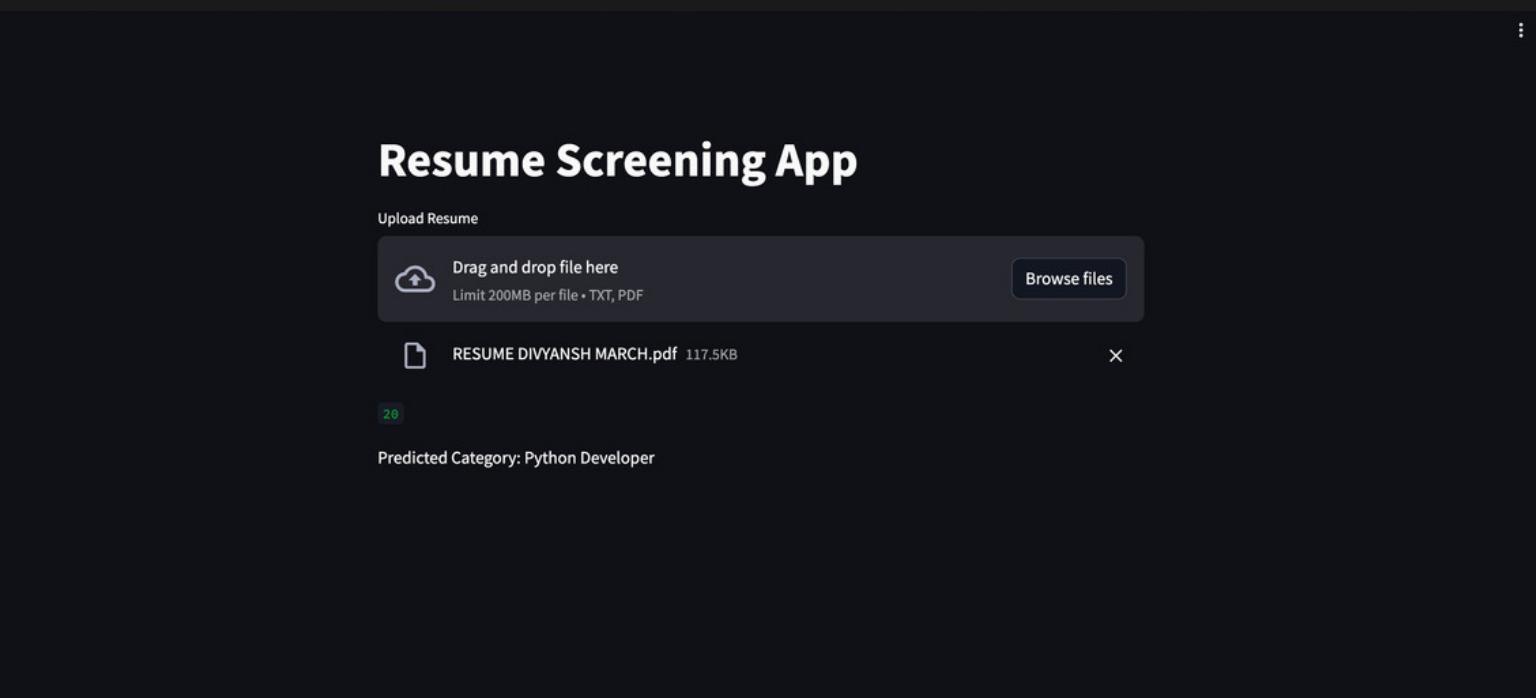
3: "Blockchain",
10: "ETL Developer",
18: "Operations Manager",
6: "Data Science",
22: "Sales",
16: "Mechanical Engineer",
1: "Arts",
7: "Database",
11: "Electrical Engineering",
14: "Health and fitness",
19: "PMO",
4: "Business Analyst",
9: "DotNet Developer",
2: "Automation Testing",
17: "Network Security Engineer",
21: "SAP Developer",
5: "Civil Engineer",
0: "Advocate",

category_name = category_mapping.get(prediction_id, "Unknown")

st.write("Predicted Category:", category_name)
print(category_name)

# python main
main() > if uploaded_file is not None
```

Our software operates by receiving resumes with read permissions as input. It then proceeds with a process of vectorization, transforming the resume's content into numerical representations. Subsequently, the software employs classification techniques to determine a domain of interest for the applicant or potential domains they could pursue.



REFERENCES

- N. Jayakumar, A. K. Maheshwaran, P. S. Arvind and G. Vijayaragavan,"On-Demand Job-Based Recruitment For Organisations Using Artificial Intelligence," 2023 International Conference on Networking and Communications (ICNWC), Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICNWC57852.2023.10127551.** keywords: {Visualization; Machine learning algorithms; Resumes; Organizations; Machine learning; Forestry; Market research; Employee attrition; Recruitment; Random Forest}
- Z. Chuang, W. Ming, L. C. Guang, X. Bo and L. Zhi-qing,Resume Parser: Semi- structured Chinese Document Analysis," 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, USA, 2009, pp. 12-16, doi: 10.1109/CSIE.2009.562.** keywords: {Resumes; Text analysis; document analysis; semi-structured; resume parsing; pattern matching},