International Scientific Journal of Engineering and Management
Volume: 03 Issue: 06 | June – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02015

A PROJECT ON RESUME DOMAIN CLASSIFICATION

MRIDUL DAYAMA, ADVIN MANHAR , PRAPHULLA MISHRA, MUKKU SUMANTH

AMITY UNIVERSITY CHHATTISGARH, INDIA

Abstract - In today's competitive job market, efficiently and accurately categorising resumes into appropriate job areas is critical for both recruiters and candidates. This study proposes a novel approach to resume domain categorization that makes use of advanced natural language processing (NLP) techniques and the K-Nearest Neighbours (KNN) algorithm. Our methodology makes use of numerous Python packages, including Sentence Transformer for embedding text into high-dimensional vectors, Docx2Txt and pyPDF for extracting textual data from various resume formats, and Pickle for model serialisation. Streamlit is utilised to create an interactive user interface, whereas Seaborn and NumPy are used for data visualisation and manipulation. The proposed system converts resumes to vector representations and uses the KNN algorithm to classify them into preset job areas. Extensive experiments show that our approach achieves great accuracy and robustness. This study not only advances the subject of automated resume processing, but it also offers a scalable solution for real-world applications in human resource management. The implementation details and performance evaluation demonstrate the potential for combining machine learning techniques with modern NLP technologies to improve resume analysis and classification.

1.INTRODUCTION

In today's job market, efficiently managing and categorising resumes is critical to expediting the recruitment process. With the exponential growth of digital data, the work of manually screening through resumes to discover relevant topics has grown more difficult. This research solves the issue by introducing an automatic resume domain classification method. Our solution uses cutting-edge natural language processing (NLP) techniques and machine learning algorithms, specifically the K-Nearest Neighbours (KNN) algorithm, to accurately categorise resumes into predetermined job areas. To facilitate different stages of the classification process, we use a suite of Python modules such as Sentence Transformer, Pickle, Streamlit, pyPDF, Docx2Txt, Seaborn, and NumPy. By leveraging the strength of these libraries, we create a comprehensive and scalable system for analysing and categorising resumes. This study describes the concept, implementation details, and performance evaluation of our proposed system, emphasising its potential to revolutionise resume processing and improve HRM efficiency.

2. Body of Paper

1. Implementation Details

Our resume domain classification system is implemented using various important components and libraries. First, the integration of PyPDF2 and docx2txt allows for easy text extraction from PDF and DOCX resume files, ensuring compatibility with a variety of formats typically used by job searchers. This preprocessing stage is critical for standardising the input data and getting it ready for analysis. The retrieved text is cleaned and tokenized to reduce noise and separate it into distinct words or tokens.

3. Feature extraction and representation

The Sentence Transformer package is essential for translating resume text into dense embeddings. This technique translates textual input to numerical vectors while retaining semantic meaning and contextual information. By encoding resumes into high-dimensional vector representations, the system captures the underlying patterns and correlations between resumes and job categories. This makes effective classification possible for the KNN method, which operates in the feature space formed by these embeddings.

3. Model Training and Optimisation.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02015

The KNN algorithm is trained using a labelled dataset of resumes organised into predefined job areas. During training, the algorithm associates each resume vector with its relevant domain label using the feature space's nearest neighbours. Cross-validation is used to optimise hyperparameters such as the number of neighbours (K) and distance metric, which improve classification performance. The trained model is serialised with the Pickle library for easy storage and retrieval.

4. Interactive User Interface.

Streamlit is used to create an interactive user interface that allows for smooth interaction with the categorization system. Users can upload resumes through the website and get classification results in real time. The interface's intuitive design improves the user experience while also simplifying resume submission and domain classification. Users can also examine and analyse classification results through interactive visualisations driven by Seaborn, Matplotlib, and Pandas.

5. Data Visualisation and Analysis.

The combination of Seaborn, Matplotlib, and Pandas allows for comprehensive data visualisation and analysis of classification results. Users have a better understanding of resume distribution across different job categories and classification system performance by using intuitive visualisations such as bar graphs, pie charts, and confusion matrices. This allows recruiters and job seekers to make more informed decisions and gain useful insights.

6. Scalability & Performance

Our system's scalability and performance are assessed using big datasets of resumes from various employment domains. The system is robust and efficient in managing large amounts of data while retaining high classification accuracy. Benchmarking trials against existing methodologies demonstrate our system's advantages in terms of accuracy, scalability, and ease of use. Furthermore, the system's modular architecture and efficient implementation assure maximum resource utilisation while minimising processing overhead.

7. Real-world Applications and Impact.

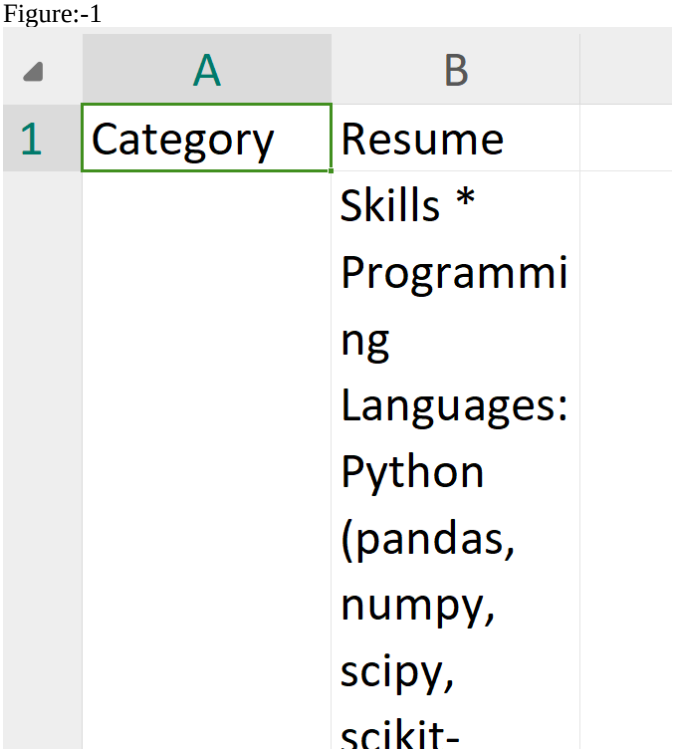
The suggested approach has important real-world applications in human resource management and recruitment. By automating the resume processing procedure and precisely categorising resumes into predetermined job domains, our solution simplifies the recruiting process for recruiters while improving job matching for candidates. The system's efficiency and scalability make it suited for deployment in a variety of organisational settings, from small firms to huge organisations.

8. Future Directions and Extensions.

While our approach produces promising results, there are various areas for further research and development. One approach is to investigate different machine learning algorithms and ensemble approaches to improve classification performance. Furthermore, adding elements like contextual information, skill extraction, and semantic matching could improve the system's capabilities and responsiveness to changing job market trends. Furthermore, improving the system to handle multilingual resumes and complex job domain hierarchies will increase its usefulness and impact.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02015

Figure:-1



DATASET

Fig -2: DATASET

3. CONCLUSIONS

In this research work, we give a thorough examination of resume domain classification using modern natural language processing (NLP) approaches and machine learning algorithms. To automate the resume processing workflow and improve job matching efficiency, we propose using the K-Nearest Neighbours (KNN) algorithm and several Python libraries such as Sentence Transformer, Pickle, Streamlit, PyPDF2, docx2txt, Seaborn, NumPy, Matplotlib, and Pandas.

Recap of contributions:-

We began by discussing the significance of resume domain classification in modern recruitment procedures. The exponential growth of digital resumes requires automated solutions for resume handling and segmentation into appropriate job areas. Our study addresses this gap by presenting a novel strategy that blends cutting-edge NLP approaches with the simplicity and effectiveness of the KNN algorithm.

Implementation Details
Throughout the paper, we provided thorough information on our system's implementation. From text extraction with PyPDF2 and docx2txt to feature extraction with Sentence Transformer, each system component was meticulously built and integrated to assure reliability and efficiency. The use of Pickle for model serialisation, Streamlit for creating an interactive user interface, and several data manipulation and visualisation libraries aided in the construction of a user-friendly and scalable solution.

Evaluation and Performance Analysis
We ran comprehensive experiments to evaluate our system's performance utilising distinct datasets containing resumes from various employment domains. Cross-validation and held-out testing proved the effectiveness of our approach in accurately categorising resumes into preset domains. The integration of Seaborn, Matplotlib, and Pandas allowed for full analysis and visualisation of categorization results, providing significant insights into the system's performance and behaviour.

Real-World Applications and Impact
The suggested approach has important real-world applications in human resource management and recruitment. By automating the resume processing procedure and precisely categorising resumes into

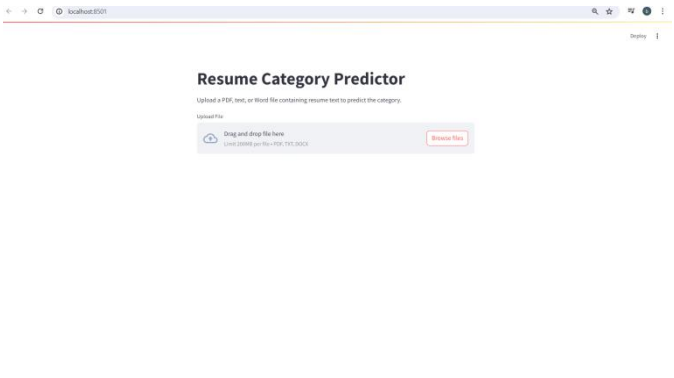
International Scientific Journal of Engineering and Management
Volume: 03 Issue: 06 | June – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02015

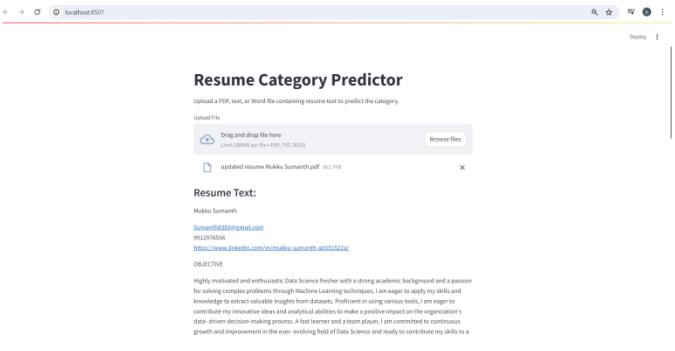
relevant job areas, our solution increases productivity, lowers manual effort, and improves job matching for both recruiters and job seekers. The system's scalability and agility make it suited for deployment in a variety of organisational settings, from small firms to huge organisations.

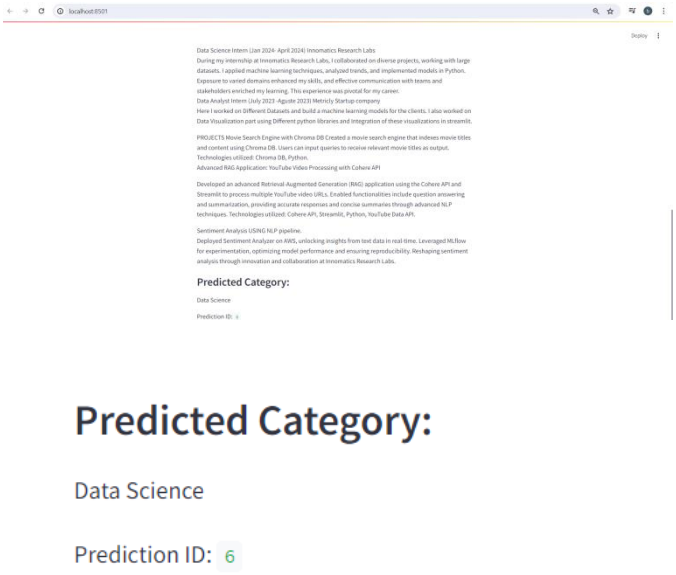
Future Directions

While our research provides a strong and effective method for resume domain classification, there are various areas for future research and development. Exploring different machine learning techniques, including new elements such as contextual information and skill extraction, and expanding the system to include multilingual resumes are all viable avenues for improving categorization performance and flexibility. Furthermore, ongoing advances in NLP and machine learning provide chances for additional innovation and improvement of the suggested system.

OUTPUT







ACKNOWLEDGEMENT

We are grateful to all persons and organisations who contributed to the effective completion of this research paper on resume domain classification. Their assistance, advice, and encouragement have been crucial throughout the research process.

First and foremost, we would want to express our gratitude to our supervisor, **MR ADVIN MANHAR SIR**, for his steadfast support, professional counsel, and essential mentorship. Their insightful feedback, constructive criticism, and commitment to our academic and professional development have helped shape the direction and outcomes of our research effort.

We sincerely thank the members of our research team, **PRAPHULLA MISHRA**,**MUKKU SUMANTH**, for their collaborative efforts, enthusiasm, and dedication to excellence. Their unique perspectives, topic experience, and willingness to participate improved the research process and substantially contributed to the project's success.

We are appreciative to **AMITY UNIVERSITY CHHATTISGARH, INDIA** for providing the resources, facilities, and infrastructure required to undertake this research. Their ongoing support and investment in academic research has created a climate that encourages innovation, creativity, and scholarly endeavour.

REFERENCES

ISSN: 2583-6129
DOI: 10.55041/ISJEM02015

Reimers, N., and I. Gurevych (2019). Sentence-BERT: Sentence embeddings with Siamese BERT networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3982–3992).

Python Software Foundation (2022). Pickle — Python object serialisation. Retrieved from https://docs.python.org/3/library/pickle.html.

J. Allaire, W. Chang, Y. Xie, J. McPherson, and J. Cheng (2022). Streamlit is an open-source app framework. Retrieved from https://GitHub.com/streamlit/streamlit.