e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 11, November 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 11, November 2024| DOI: 10.15680/IJMRSET.2024.0711102

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Application of NLP for Information Extraction from Unstructured Documents

Karrem Praneeth Reddy[1], Pawar Pranay[2], Tekumatla Pranay[3], Vadde Pranay[4], Goli Praneeth Sai Saran[5], Ms.Maddi Sri V.S.Suneeta[6]

Department of (AI & ML), School of Engineering, Malla Reddy University, Hyderabad, India[1,2,3,4,5]

Project Guide, Department of (AI & ML), School of Engineering, Malla Reddy University, Hyderabad, India[6]

ABSTRACT: The increasing interest in data has led to significant investments in developing tools that can analyze and extract useful information from various sources. However, when it comes to applicant tracking systems(ATS) that gather information from candidates' resumes and job descriptions, most approaches are still rule-based and do not fully utilize modern techniques. This is surprising because, although the content of these documents may vary, their structure is usually quite similar. In this paper, we introduce a Natural Language Processing (NLP) pipeline designed to extract structured information from a wide range of textual documents, with a focus on those used in applicant tracking systems, such as resumes and job postings. The pipeline employs several NLP techniques, including document classification, segmentation, and text extraction. To classify the documents, we use algorithms like Support Vector Machines (SVM) and XGBoost, which help in accurately identifying the type of document based on its content. After classification, the documents are divided into different sections using methods such as chunking, regular expressions, and Part-of-Speech (POS) tagging. These techniques allow us to identify and focus on the most important parts of the document. Finally, we use tools like Named Entity Recognition (NER), regular expressions, and pattern matching to extract relevant information from each section. The structured data obtained can be used to improve various processes, such as document organization, scoring, matching, and auto-filling forms, making ATS systems more efficient and effective for both job seekers and employers.

I. INTRODUCTION

The introduction outlines the challenges of extracting relevant information from the vast amounts of unstructured data available online. With the growth of the Internet, manual extraction of information from diverse documents—such as news articles, research papers, CVs, and job postings—has become impractical. **Information Extraction (IE)**, an important Natural Language Processing (NLP) technique, addresses this issue by converting unstructured text into structured formats for easier analysis. This research specifically focuses on automating the extraction of key details from **IT-related CVs** and **job vacancies**, as both are vital for recruitment. The aim is to streamline the hiring process by efficiently identifying personal information, educational background, skills, and work experience from CVs, while extracting job titles, required skills, and responsibilities from vacancies. The methods developed for this extraction are discussed in detail in subsequent sections, ultimately aiming to enhance candidate selection and simplify recruitment.

II. LITERATURE SURVEY

A literature survey for the project on the application of NLP for information extraction from unstructured documents would include an exploration of existing research and methodologies in related areas. Here's a structured overview:

1. Information Extraction in NLP:

"Natural Language Processing for Information Extraction" (arXiv, 2018): This paper reviews various NLP techniques essential for effective information extraction from unstructured text sources. It discusses rule-based, statistical, and hybrid approaches, as well as the role of syntactic and semantic parsing for identifying relevant entities and relationships. The paper provides a foundation for understanding how traditional and modern NLP methods can be applied across different fields, such as finance, law, and recruitment, highlighting the adaptability of these

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

techniques.

2. CV Parsing Techniques:

"Application of Machine Learning Algorithms to an Online Recruitment System" (International Conference on Internet and Web Applications and Services, 2015): This study investigates various machine learning techniques that enhance the efficiency of online recruitment platforms, particularly through CV parsing. The paper explores algorithms like decision trees, support vector machines (SVM), and ensemble methods for classifying and extracting relevant information from candidate CVs. It demonstrates how these methods can automate the categorization of job experiences, skills, and qualifications, thus reducing HR workload and improving candidate-matching accuracy.

3. Named Entity Recognition (NER) in Healthcare and Other Domains:

"FoodIE: A Rule-based Named-entity Recognition Method for Food Information Extraction" (Proceedings of the International Conference on Pattern Recognition Applications and Methods, 2019): This paper presents a rule-based approach to Named Entity Recognition (NER) in the context of food information extraction, which is particularly relevant for highly specialized domains. It highlights the strengths of rule-based methods in scenarios where domain-specific knowledge enhances accuracy, such as healthcare, legal, and recruitment contexts. The paper also discusses challenges in developing NER systems that need to handle nuanced terms and phrases, which is a common issue in CV parsing and job description analysis.

4. Machine Learning and Deep Learning in Information Extraction:

"Deep Q-Learning for Medical Decision Making" (Proceedings of the IEEE International Conference on Healthcare Informatics, 2020): While focused on healthcare, this paper explores the application of deep reinforcement learning (specifically Q-learning) in decision-making, which has implications for NLP-based information extraction tasks. By illustrating how deep learning can model complex relationships in unstructured data, this work emphasizes the potential of using neural networks to extract, classify, and predict outcomes based on textual information. Techniques from this domain could be adapted to recruitment, such as by predicting candidate fit or automating decision points in applicant tracking systems (ATS).

III. PROBLEM STATEMENT

The recruitment process heavily relies on extracting relevant information from unstructured documents, particularly CVs and job descriptions. Traditional methods, primarily rule-based systems, struggle to adapt to the variability in document formats and often lead to inefficient parsing and low accuracy. These systems are unable to leverage modern Natural Language Processing (NLP) techniques, resulting in suboptimal performance in extracting critical data.

As a consequence, HR professionals spend excessive time manually reviewing documents, hindering the efficiency of the recruitment process. There is a pressing need for a more robust solution that can accurately categorize, segment, and extract structured information from diverse textual inputs. The goal is to implement an NLP pipeline that automates this extraction process, significantly improving the speed and accuracy of candidate evaluation, thereby enhancing overall recruitment effectiveness.

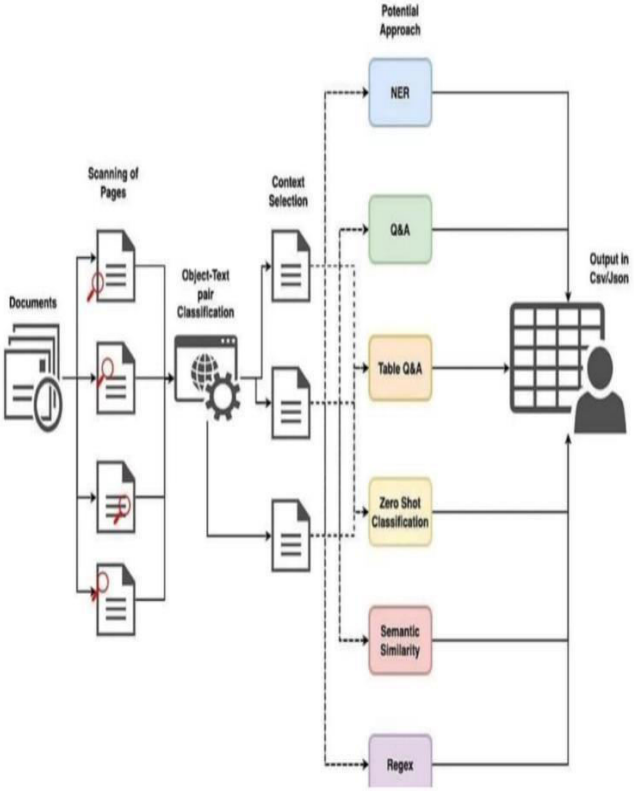
© 2024 IJMRSET | Volume 7, Issue 11, November 2024| DOI: 10.15680/IJMRSET.2024.0711102

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. SYSTEM DESIGN



V. METHODOLOGY

Our project focuses on extracting information from unstructured documents, particularly CVs and job vacancy details in the IT sector, through the following steps:

1. **Data Preprocessing**: Text is normalized (lowercased and cleaned), and missing values are handled via imputation or removal. Outliers are detected, and numerical features are scaled while categorical variables are encoded.
2. **Document Classification**: A Support Vector Machine (SVM) model is trained on 10,670 documents to classify them into CVs, job-vacancy details, and others, achieving an accuracy of 98.7%.
3. **Document Segmentation**: Gaussian Naive Bayes is used to segment CVs into sections like personal information and education.
4. **Named Entity Recognition (NER)**: A Conditional Random Field (CRF) model tags entities in segmented CVs, such as names and qualifications.
5. **Feature Engineering**: Age is categorized into bins, and quality-of-life metrics are combined into a composite health score.
6. **Model Evaluation**: k-fold cross-validation assesses model performance, and accuracy is evaluated using a confusion matrix.
7. **Deployment**: The model undergoes testing and real-time monitoring post-deployment to ensure performance and facilitate updates based on user feedback.

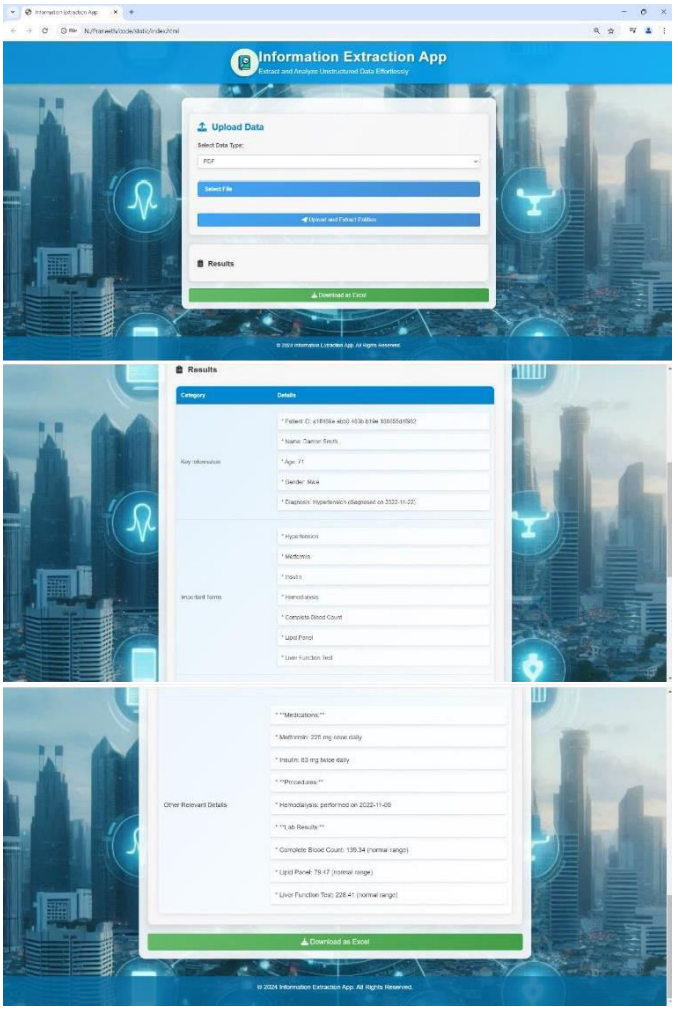
© 2024 IJMRSET | Volume 7, Issue 11, November 2024| DOI: 10.15680/IJMRSET.2024.0711102

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. RESULTS



VII. CONCLUSION

The conclusion of the project states that the authors have demonstrated an efficient and accurate method for structured-information extraction from textual documents. This capability is achieved through the application of Natural Language Processing (NLP) techniques, combined with Machine Learning (ML) and Deep Learning (DL) models. The research specifically references the use of CVs and job vacancy information, which are commonly utilized in applicant tracking systems. The results indicate that the system provides high evaluation metrics and improved execution times for information extraction. The authors suggest that such.

optimized and accurate systems could be beneficial in various fields, including research publications and job portals. They also acknowledge the need for future adaptations of their methods to accommodate changes in requirements or new types of data, and they express interest in exploring additional techniques for information extraction from unstructured documents across different domains.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VIII. FUTURE ENHANCEMENTS

The evolution of information extraction methods holds significant potential for advancing recruitment processes beyond the current capabilities. Future enhancements to our system may include:

a. **Integration of BERT Models**: By utilizing transformer-based architectures like BERT for embedding, we can improve the model's ability to understand context and semantic nuances in text, leading to more accurate information extraction from CVs and job descriptions.
b. **Expansion to Other Domains**: While our current focus is on CVs and job vacancies in the IT field, future research could adapt the system to extract information from various types of unstructured documents across different industries, enhancing its applicability and utility.
c. **Exploration of Document Similarity**: Investigating document similarity metrics, rather than solely relying on token similarity, could allow for improved clustering and classification of documents, facilitating better candidate matching.
d. **Real-time Processing Capabilities**: Enhancing the system to process documents in real-time will enable quicker responses in recruitment scenarios, allowing for more agile hiring processes.
e. **User Feedback Mechanism**: Implementing a feedback loop from recruiters can help refine the extraction algorithms and improve the overall system performance, ensuring it adapts to the changing requirements of the recruitment landscape.
f. **Incorporation of Multi-modal Data**: Exploring the integration of multi-modal data sources, such as video or audio interviews, alongside traditional text-based CVs can provide a more holistic view of candidates, enhancing the decision-making process.

REFERENCES

1. Singh, S., 2018. Natural Language Processing for Information Extraction. arXiv,
2. Zeroual, I. and Lakhouaja, A., 2018. Data science in light of natural language processing: An overview. Procedia Computer Science, 127, pp.82-91.
3. Sovren.com. 2020. Home.
4. Technologies, D., 2020. Daxtra - CV Parsing.
5. AkkenCloud. 2020. Top Staffing And Recruiting Software Solution — Akkencloud.
6. Chandola, D., Garg, A., Maurya, A. and Kushwaha, A., 2015. ONLINE RESUME PARSING SYSTEM USING TEXT ANALYTICS.
7. Das, P., Pandey, M. and Rautaray, S., 2018. A CV Parser Model Using Entity Extraction Process And Big Data Tools.
8. Faliagka, E., Ramantas, K. and Tsakalidis, A., 2012. Application of Machine Learning Algorithms to an online Recruitment System. The Seventh International Conference on Internet and Web Applications and Services
9. Malmasi, S., Sandor, N., Hosomura, N., Goldberg, M., Skentzos, S. and Turchin, A., 2017. Canary: An NLP Platform for Clinicians and Researchers. Applied Clinical Informatics, 08(02), pp.447-453.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com