International Journal of Research in Engineering, Science and Management
Volume 6, Issue 12, December 2023
https://www.ijresm.com | ISSN (Online): 2581-5792

HR Analytics: Resume Parsing Using NER and Candidate Hiring Prediction Using Machine Learning Model

B. V. Brindashree[1*], T. P. Pushphavathi[2]

[1]*Student, Department of Computer Science Engineering, M. S. Ramaiah University of Applied Sciences, Bangalore, India*

[2]*Associate Professor, Department of Computer Science Engineering, M. S. Ramaiah University of Applied Sciences, Bangalore, India*

***Abstract*: HR analytics is a leveraging technology in which the process of HR business is automated and to develop data driven insights to inform talent decisions. In this project, the main objective is to eliminate or to reduce the HRs job and make the selection process automated. In order to achieve this, the project was divided into two modules. One module processes the resumes and parses them to extract the required data. Firstly, the resume parsing is done using the Named entity recognition (NER) method using the spacy model. The resumes are annotated using NER annotator which labels the required text/ data in the resumes. The second module wants to make the candidate hiring predictions automated. The candidates will be checked on different aspects like experience, their stand towards one company i.e. stability which tells how stable the candidates are in staying with one company. The tenure with the current company they are working for, which tells how many months the candidate is working in the current company when they are applying for the job. The first module gave results that were accurately obtained after using the Named entity recognition method to parse the resumes and extract the skills from the resumes. After successfully extracting the correct skills from resumes with high accuracy, it was matched with the Job ID which matches the skills with the required data in the job description. The second module of the project gave the results based on a decision tree which was plotted and made decisions whether the candidate profile is fitted or not. The prediction is based on training the decision tree ML model which gave an accuracy of 70 %. The precision, recall and f1 score for the data is obtained and checked how well the data is trained. Overall project will make the HR job easier and automate the process which will help to reduce the complete time of recruitment.**

***Keywords*: NLP, HR analytics, Named entity recognition (NER), Resume parsing, Decision tree.**

1. Introduction

HR analytics which is also known as people analytics is the collection and application of talent data to improve all the critical talent and business outcomes. HR analytics leaders will enable the HR leaders to develop most of the data-driven insights to inform talent decisions, and also improve workforce processes which promote positive employee. experience. The main objective is to eliminate or to reduce the HRs job and make the selection process automated. In order to achieve this, the project was divided into two modules. One module processes the resumes and parses them to extract the required data. Resume parsing is a technique in which all the data in the candidates resumes will be parsed and extracted. This technique is used to eliminate the manual process of HR reading each resume and identifying the matching requirements HR wants. The HR's will take the important information from the resumes and move further with the selection process. As the manual process of going through each resume takes more time, this project wants to make the entire process automated by introducing new techniques of Natural language processing and Machine learning. Firstly, the resume parsing is done using the Named entity recognition (NER) method using the spacy model. The resumes are annotated using NER annotator which labels the required text/ data in the resumes. The annotated data is used for training and the results are obtained. After the skill extraction process, the next step will be to find the Job details from different companies and see how to match the requirements. The data of Job ID's and Job descriptions from different companies across different cities will be taken. After which the requirements and skills will be matched and provides the Job ID's which will match the candidates resumes. In simple words the Job ID's will be shortlisted based on the skills extracted from the candidates resume. This process will remove the manual process of HR going through each candidate's resume and find the suitable job role in their company. In the time of mass recruiting this will be more time consuming and there is a high chance of misconceptions and mistakes of choosing the eligible candidate. The second module of the project deals with candidate hiring decision predictions. The HR process will have many rounds of tests and the process of the selection is not just one go. First step will be to schedule an aptitude test, psychometric tests for the candidates, the shortlisted candidates from the test will be taken to further rounds of different tests and interviews, and a final call will be made for the selected candidates. The first module in our project has automated the manual resumes checks and candidates' skills matching the requirements. Now the second module wants to make the candidate hiring predictions automated. The candidates will be checked on different aspects like experience,

*Corresponding author: brindashreebv@gmail.com

their stand towards one company i.e. stability which tells how stable the candidates are in staying with one company. The tenure with the current company they are working for, which tells how many months the candidate is working in the current company when they are applying for the job. All these will be asked by the HR with the candidates who have experience in industry and not a fresher. Data analysis to prepare data from the real facts and make a good decision based on the data. Applying machine learning and new technology-based models for the problem statement. Making the decisions automated, rather than HRs going through all the requirements - which is time taking and not so feasible when there are large number of candidates. Again, checking all these and planning will be time consuming for HR when there are huge candidates who need to be selected. Hence automation is required. Machine learning model (Decision Tree) is built to train with the data of existing employees of the company where the HR is hiring. After the ML model is trained it will give a prediction and decision based on the data trained. Taking the predictions from the Machine learning model, a decision is made whether the candidate is selected based on the profile fitment. A result will be given whether the candidate profile is fit or not.

2. Literature Review

This section discusses several existing methodologies and works related to Resume parsing, Named entity recognition model and HR analytics. Few papers which discusses resume parsing are, Sreejith et al [3]., In this paper a model was proposed to classify different segments in a resume using deep learning models such as Convolution neural network (CNN), the work also had CRF and Bi-LSTM-CNN models for sequence labeling in order to tag different entities. The work also uses pre trained glove model for word embeddings. Papiya Das et al [4], this paper the research was mainly about text analysis process and how to extract entities with different big data tools. Satyaki et al [2] this research work involves a resume parsing model using NLP where the similar keywords are clustered and shows most relevant resume to the employee based on the keyword matching , the keywords matched are auto filled and showed to the employees and saves the matched keywords into NoSql database and also in json format. Kanya et al [1] in this paper different NLP techniques like Named entity recognition for information extraction is discussed. The paper presents various efficient NER techniques and modeling . Named entity recognition addresses the problem of locating textual mentions of predefined types of entities, where the entity categories can be very diverse, ranging from people and companies in business applications to cells and proteins in biomedical applications. Michael et al [5], this paper discusses the overview and analysis of IT solutions in which text understanding is applied to the programming domain. Conclusions are drawn by authors with respect to experience in NLP/NLU in recent years. To solve typical problems Binary classification and logistic regression is used. The paper develops the ideas of understanding texts in the software development domain using standard text processing tools. The proposed solution is recommended for HR professionals who search suitable candidates for a job based on their blogs, online presence and code. Kajal et al [6], to optimize the entire process of HR interviews, this paper proposes video analytics which is used to screen candidates. A candidate's emotion is extracted from his speech using Mel-Frequency Cepstral Coefficients (MFCCs) as a major classification feature for the Artificial Neural Network (ANN). Alessandro et al [7], this paper provides a large systematization effort and a research agenda for developing detailed studies in the field of HR analytics. From a practitioner perspective, the study in this paper offers insights to support the design of innovative analytics in the projects within the same organizations. Liyuan et al [8] Basically the HR data is messy and imbalanced, it is hard to harness both structured and unstructured data, some HR managers lack data mining skills and the lack of related empirical research that gives a detailed analytics guideline. This paper tells us the framework which is developed to support an industrial aluminum company to make the decisions and to improve strategy execution. The framework also includes descriptive analysis, predictive analysis, and entity sentiment analysis. The work was analyzed based on the industrial aluminum company's data and found some actionable issues. Roshini et al [9] Data mining is the process of data extraction using particular algorithms. It is mainly used to predict the future scope of the study using existing data. In the field of HR, data mining is mainly used to understand attrition, performance, retention and recruitment. The aim of this paper focuses on executing a decision tree algorithm and K Nearest Neighbor, in order to understand the relationship between chosen variables in the HR data used. Ganga Naga et al [10], this article explores the importance of Data Analytics tools in Human Resource management. Data analytics, Business analytics are new terminologies in business. Predictive analytics are very important in every organization. HR analytics concepts very useful for measuring employee performance, informed decisions about salary and promotions, to increase employee retention, Examining the employee engagement, measuring the employee developments and learning outcome. By using Statistical tool ANOVA we concluded that there is a significant difference between various levels of HR Roles in the organization.

3. About Dataset

Data is important for any model to develop and analyze the working of each concept which is planned. The data in this work is taken and created on own, and tried with different technical models. Firstly, the resume parsing model requires the resumes to extract all the information like education details, experience, and their skills. The personal resumes are taken and trained. In this project skills are the main part of entities which need to be extracted for the model preparation. The NER annotator is used to annotate/ tag the skills in the resumes chosen. The tagged skills are downloaded in the form of a json file. The JSON file is then sent to the model training and results are drawn. The data for Job matching with the skills extracted is taken from the glassdoor website. Web scraping method is used to scrape the required fields of data from the glassdoor website and extract the data. Hence the data consists of Job ID's, required skills for the corresponding Job ID's and the related information like company name, job role etc. Next the data for candidate hiring criteria is prepared with knowledge of experience from professional HR's data. The data was developed by creating a sample data which had 1143 candidate's information where the candidates age, tenure (in months), stability (in years), experience, highest education, productivity, etc., was created with reference to real word HR's data. The target variable for machine learning model prediction was taken based on the productivity of the candidate. The entire data was developed based on the idea of real HR data, with the knowledge of work done in an IT company during training period.

4. Methodology

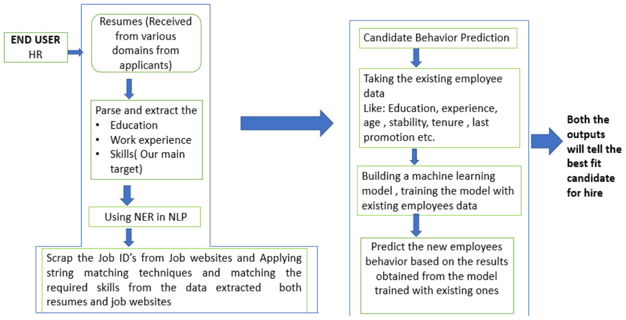
END USER
HR
Resumes (Received from various domains from applicants)
Parse and extract the
• Education
• Work experience
• Skills(Our main target)
Using NER in NLP
Scrap the Job ID's from Job websites and Applying string matching techniques and matching the required skills from the data extracted both resumes and job websites
Candidate Behavior Prediction
Taking the existing employee data
Like: Education, experience, age , stability, tenure , last promotion etc.
Building a machine learning model , training the model with existing employees data
Predict the new employees behavior based on the results obtained from the model trained with existing ones
Both the outputs will tell the best fit candidate for hire

Fig. 1. Framework

Pyresparser is the resume parser which is used for extracting all the information's from resumes. The pyresparser is a resume parser library which is used to extract all the information from the resume. In this project the pyraseparser library was used to extract the skills from the resumes. As the results were not predicted accurately. The skills were not properly extracted and the results were incomplete. This was the disadvantage of existing pyresparser library. Spacy is an NLP library which is used for multiple purposes of applications that processes and understands large volumes of text. It is used to extract information, understand the system in a natural way and to pre-processes text for deep learning. Spacy provides a small English pipeline trained on written web text (blogs, news, comments) which includes vocabulary, entities and syntax called en_core_web_sm. Spacy will predict the text and tell whether the text is a person, location or organization etc. in this way when we provide a resume with skills the spacy library cannot predict the skills instead it predicts as different entities such as person, GPE and ORG. This is because spacy is not trained for technical words. In order to overcome the issue, spacy needs to be trained with custom data. NER annotator is a UI which facilitates to label the text with our desired labels. In this project the resume is converted into a text file and sent to the NER annotator. The annotator reads each sentence. The customized labels are created and used for labeling the text based on individual choice. In this way annotator provides a json file. The json file contains the labeled text, the starting and end index of the text which is labeled and the customized entity names. In spacy pipeline the custom data created using ner annotator is fed. The blank English model in spacy is created and the custom data trained pipeline is combined and trained. After training the custom data in spacy, it gives us two models one is the best model and another is the last model. The model best is used to find and predict the results. The data for Job matching with the skills extracted is taken from the glassdoor website. Web scraping method is used to scrape the required fields of data from the glassdoor website and extract the data. Hence the data consists of Job ID's, required skills for the corresponding Job ID's and the related information like company name, job role etc. Firstly the Job description is selected and made a separate column as test in the dataframe. This will make the process of string matching easy with the extracted skills. The extracted skills which were accurately predicted using the NER method are stored as a list with variable name skills. This variable skill is used for matching the extracted skills with the skills mentioned in the column which have Job ID's and Job description. Now it's time to use string matching techniques to read the text/words and match them accordingly. ftfy fixes Unicode that's broken in various ways. The goal of ftfy is to take in bad Unicode and output good Unicode, for use in your Unicode-aware code. This is different from taking in non-Unicode and outputting Unicode, which is not a goal of ftfy. It also isn't designed to protect you from having to write Unicode-aware code. ftfy helps those who help themselves. Tfidf Vectorizer will convert the collection of row documents into a matrix of TF-IDF vector features. The analyzer is used as n-grams, it is used to tell whether a feature should be made of words or characters. The threshold value is set using min_df=1, when building the vocabulary ignore the terms that have a document frequency strictly lower than the given threshold. The nearest neighbor's method is used for analysis of measuring the spread or distribution of everything under a geographical space. Cosine similarity provides a numerical value describing to which set of points the clusters are or they are uniformly spaced. In this way the similar strings which fall under the same group will be matched, and provide the corresponding Job ID's which matches the skills extracted using NER and skills mentioned in Job description. In the process of hiring a candidate it's always important from the HR perspective to hire a candidate with few eligible criteria or few important keys which HR look into. HR not only wants a candidate who is meeting the job requirements like skills and previous projects, but also HR wants a candidate who has good experience in his/her work life, stability which tells how many years a candidate has stayed in one particular company, tenure which tells how many months the candidate is tenured with the existing company,etc. In this project the Decision tree is used to predict and make decisions on which type of candidate can be hired. Decision tree gave better decision criteria than other machine learning models, because the other ML models gives the overall accuracy of the trained model whereas the decision tree can be plotted and take the decisions based on true positive values. The data is taken and created using industrial experience. Usually, the data can be taken based on the existing employee's details. The details of existing employees can be made used to train the model, and bring out the results. The categorical variables are encoded using sklearn label Encoder method. Every continuous variable like age, experience, stability are made as buckets. Ex: Age - 20-30 years, Experience - 2-5 years, 6 above etc. These buckets further are labeled. Ex: Age: 20-30 years = 1, 30-40 years= 2, 40- 50 years= 3, 50 and above =4. After taking the existing employees data, the model will be trained with the data and results will be obtained. For training the main variables chosen are Age_bucket, Source of hire, Highest education, Experience_bucket, Total companies, Stability bucket. The target variable chosen is based on the productivity, after many trial and errors, one decision label was chosen as the target variable where the productivity is greater than a certain percentile value.

5. Experimental Results

After parsing the resume, the skills are labeled using the NER annotator and trained with the NER spacy English model. The above snippet shows the results obtained after training the NER model with the labeled data. The skills mentioned in the resumes or the skills parsed from the resumes are passed to the model, the model gives the prediction and tells how many words skills in the text. In our model, all most all the skills are predicted correctly as skills. There is also a statement passed to ner spacy model for testing, and the model has correctly predicted the education details, skills and date. In this project, the job information from different companies is scrapped and extracted from the glassdoor website using the web scraping method. After extracting the job details, openings in the companies across different cities, the positions which are open, and the job description which contains what qualifications, skills, etc., are required. After scraping the data from the glassdoor website, the extracted data is stored in the form of dataframe. The above snippet shows how the extracted data is stored in the form of dataframe. After predicting the skills using NER and extracting the Job ID’s from the Glassdoor website, the next step is to match the skills with matching job ID’s. Using nearest neighbor and cosine similarity methods the string matching is done. Similar strings which fall under the same groups are identified and the similar vectors of strings are matched and the results are given. The results give the position open, the company name and the location. The results are based on skills extracted from a resume which matches the skills in the job description. The candidates hiring decisions are based on a few important required fields such as experience, highest education, stability, age etc. The machine learning model - Decision tree is trained and the results are obtained. Firstly, the accuracy of the model is 69.9 % approximately 70%. Due to less data obtained, and highly skewed data the accuracy was not able to improve. The precision, recall and f1 score is obtained from the data. The values are not so accurate as the data is highly skewed. The model results are obtained for the productivity greater than 6 months after the candidate joins. After many different trials the final results were obtained from this value. The candidate’s data in the real world will never be highly positively skewed data as there will be many ups and downs in the work growth of individuals in a company. After checking the accuracy of the model, a decision tree is plotted to take the results and make a precise prediction whether the candidate can be hired. Training the existing employee data and making the model learn what is the best pattern to choose the real approach.

Example: A candidate having:
Education= MBA,
Experience> 2Years,
Age > 25,
Stability > 2
Is selected and the Profile fitment is Consider
Otherwise, the Profile Fitment is Not Consider

The results are obtained based on true positive values where the decision tells 1 is traced and conditions are taken and fed to the backend connection to predict the candidate who matches these decisions. The tree is pruned and only the true positive values branches are retained and used for decision making.

Following are three decisions for profilement obtained after training the model:

1. Age_bucket<=0.5 —> which falls under label 0, i.e. 20-25 years
   SourceofHire<=0.5 —> which falls under label 0, i.e. consultant
   Total companies<=1.5 —> which falls under label 1, i.e., 1-3 companies
   Profile fitment = Yes
2. SourceofHire <=3.5 —> which falls under label 3, i.e., walkin
   Total companies<=2.5 —> which falls under label 2, i.e. 3-5 companies
   Exp_bucket<=2.5 —> which falls under label 2, i.e. 2-3 years
   Age_bucket<=0.5 —> which falls under label 0, i.e. 20-25 years
   Highest Education <=1.5—> which falls under label 1, i.e., Degree
   Stability_bucket<=0.5 —> which falls under label 0, i.e. 1-2 years
   Profile fitment = Yes
3. Total _companies <=1.5 —> which falls under label 1, i.e. 1-2 companies
   Age_bucket<=1.5 —> which falls under label 1, i.e. 25-30 years
   Stability_bucket<=0.5 —> which falls under label 0, i.e. 1-2 years
   Profile fitment =No

6. Conclusion

The main aim of the project was to automate the overall HR process and decision-making responsibilities in order to save time from the manual process/the traditional method. The first module gave results that were accurately obtained after using the Named entity recognition method to parse the resumes and extract the skills from the resumes. After successfully extracting the correct skills from resumes with high accuracy, it was matched with the Job ID which matches the skills with the requirement data in the job description. The second module of the project gave the results based on a decision tree which was plotted and made decisions whether the candidate profile is fitted or not. The prediction is based on training the decision tree ML model which gave an accuracy of 70 %. The precision, recall and f1 score for the data is obtained and checked how well the data is trained. The data obtained was based on the sample real HR data from the industry . As the data is highly skewed and there is less data for training the results couldn't be improved more. Overall project will make the HR job easier and automate the process which will help to reduce the complete time of recruitment.

References

[1] N. Kanya and T. Ravi, "Modelings and techniques in Named Entity Recognition-an Information Extraction task," IET Chennai 3rd International on Sustainable Energy and Intelligent Systems (SEISCON 2012), Tiruchengode, 2012, pp. 1-5.

[2] S. Adhikary, "Resume Parser with Natural Language Processing," International Journal of Engineering Science and Computing, February 2017.

[3] C. H. Ayishathahira, C. Sreejith and C. Raseek, "Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing," 2018 International CET Conference on Control, Communication, and Computing (IC4), Thiruvananthapuram, India, 2018, pp. 388-393.

[4] Das, P., Pandey, M., & Rautaray, S.S. (2018). A CV Parser Model using Entity Extraction Process and Big Data Tools. International Journal of Information Technology and Computer Science.

[5] Ivaschenko, A., Milutkin, M. (2019). HR Decision-Making Support Based on Natural Language Processing. In: Kravets, A., Groumpos, P., Shcherbakov, M., Kultsova, M. (eds) Creativity in Intelligent Technologies and Data Science. CIT&DS 2019. Communications in Computer and Information Science, vol 1083. Springer, Cham.

[6] Jewani, K., Bhuyar, A., Kaul, A., Mahale, C., Kamat, T. (2021). Smart Employment System: An HR Recruiter. In: Senjyu, T., Mahalle, P.N., Perumal, T., Joshi, A. (eds) Information and Communication Technology for Intelligent Systems. ICTIS 2020. Smart Innovation, Systems and Technologies, vol 195. Springer, Singapore.

[7] A. Margherita, "Human resources analytics: A systematization of research topics and directions for future research," in Human Resource Management Review, 2022.

[8] L. Liu et al., "Using HR Analytics to Support Managerial Decisions: A Case Study," ACM SE '20: Proceedings of the 2020 ACM Southeast Conference, 2020.

[9] S. Roshini, S. Prakash, J. Shilpha Dharshini, M. N. Saroja and J. Dhivya, "Decision Tree and KNN Analysis for HR Analytics Data," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2021, pp. 1-4.

[10] G. N. S. Bandi, T. S. Rao and S. S. Ali, "Data Analytics Applications for Human Resource Management," 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2021, pp. 1-5.