International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

AI Resume Analyzer Using Natural Language Processing and Data Mining

Project Guide: Prof. Jayashri Mankar
1.Abhishek Chauhan, 2.Aniket Gophane, 3.Aditya Karle, 4.Taslimarif Makandar
BE Students
Department of Computer Engineering
Genba Sopanrao Moze College of Engineering Balewadi, Pune-411045, Maharashtra, India

Abstract

Imbalance data conversion into structured data is the very tedious task in data mining techniques, various techniques have been already introduced to extract the data from large text and extract the features using various feature extraction techniques, some machine learning algorithms have been already introduced by various researchers for classification and display the results on heterogeneous data. This work suggests a method of eliminating or resuming important information in a curriculum vitae from the semi-structured text format, and rating it according to the preference and requirements of the client. The whole process was divided into three basic sections to achieve the desired goal. The first section consists of segmenting the whole Summary according to the content of each part, the second section consists of extracting data in a standardized form from unstructured data and the final section consists of analyzing structured data using NLP and Machine learning algorithms. The Stanford NLP rule extraction algorithm ha used to extract the various rules from raw data and select some important feature for classification as well as optimization. Experimental analysis shows the effectiveness of proposed system with classification accuracy.

Keywords : Resume parsing, Data Mining, Machine learning, NLP

Introduction

Classification is specially essential in solutions to data processing and machine-learning. Nowadays, many outlets have generated the numerous types of data in row format, as well as its hard to process from existing environments and algorithms. Text classification requires assigning the text to one or more predefined groups using some kind of classification algorithm performed by the content of the document. A Generic classification corpus has been developed and a single assessment system has been introduced to identify English text based on machine learning, which has now made significant progress. Most of the evidence in the real world is contained in relational bases. Data clustering is an essential machine learning process in which a sub-set of candidate labels is allocated to an entity, the main issue with multi-label clustering is the redundant online clustering method and the offline data set for dealing with this issue. We plan to use unstructured data classification to structured conversion systems and maximize the accuracy of the final sub-cluster. Demonstrate two implementations of our method using logistic regressions and improved gradient trees, along with a simple procedure for Expectation Maximization preparation. We also get an efficient prediction approach dependent on dynamics programming.

Literature Survey

According to [1] a recruiting case study as a basis for a statistical evaluation of several methods for calculating similarity scores. To this end, we suggest using a computer-aided resume evaluator on a group of resumes, then has professionals evaluate the same set of resumes, and finally look for a connection between the two sets of results. Finding the right computer-aided resume evaluator for digital human resources requires a consideration of the various similarity score calculation methodologies now available for processing resumes.

According to [2] an approach to resume writing that is both straightforward and straightforward to apply. We propose a program that, given some basic information about the applicant, may generate a professional-looking résumé. Users may sign up for an account and start working on their resume by entering their login details and receiving a one-time password (OTP).

According to [3] Recruiters have a hard time finding the greatest fit for a job position since the resumes candidates submit vary in format (e.g., font, color, font size, etc.). To combat these issues, recruiters may turn to natural language processing (NLP) to glean the specifics about potential candidates they need to move their candidacy ahead. In this paper, we suggest using the Stanford CoreNLP system's named entity recognition capabilities to glean data useful in the hiring process.

According to [4] Data-driven HR has been shown to significantly enhance the quality and speed of the whole recruitment process by using Natural Language Processing tools. First, a resume parser has been built utilizing natural language processing to evaluate the most important aspects of the hiring process. The computational framework of the parser was then used to create a potent instrument for resume matching based on job requirements, with the candidate's ability to produce a pie chart serving as an input.

According to [5] It might be challenging to find qualified candidates for an open position, particularly if there are numerous applications to choose from. It may be difficult for the team to find the most qualified candidate at the most opportune time if they have to go through each resume manually. An automated method of screening and rating applications might significantly reduce the time spent on the screening process. The KNN algorithm is used to select and rank Curriculum Vitaes (CV) based on job descriptions in large numbers, and the cosine similarity is used to find the CVs that are most relevant to the provided job description in our work.

According to [6], They have propose a substance extraction approach for getting content from news pages that joins a division like methodology and a thickness based methodology. A tool Block Extractor is used to identifies contents in three steps. First, it looks for all Block-Level Elements and Inline Elements blocks, which are designed to roughly segment pages into blocks. Second, it computes the densities of each BLE and IE block and its element to eliminate noises. Third, it removes all redundant BLE and IE blocks that have emerged in other pages from the same site. Compared with three other density-based approaches, our approach shows significant advantages in both precision and recall. BLE and IE blocks to gather related noises or contents. Next, we used this density-based approach and redundancy removal to obtain the final content. Based on our approach, a tool called Block Extractor was developed.

In this paper [7], The issue of naturally removing web information records that contain user generated content (UGC).To solve this problem MiBAT and MDR algorithms are used 1)MiBAT (Mining data records Based on Anchor Trees). They have represented two space imperative guided likeness measures, for example PM and PS. They have propose an information record mining calculation utilizing either PM or PS. Our instinct is exceptionally basic: each record comprises of one or a few sub trees, just one of which contains the rotate. We call such sub-trees that contain turns as grapple trees, since they give stay point data about where information records are found. 2)MDR (Mining Data Records in Web pages) MDR identifies a list of records by conducting a similarity test against a pre-defined threshold for two sub-trees in
the DOM tree of a web page. Such a method is referred to as the similarity-based approach, because the underlying assumption is that data records belonging to the same list usually have similar DOM tree structures MDR and its Limitations:-A group of similar objects, which forms a data region, is usually presented in a contiguous region and format-ted using similar HTML tags. Every record in a data region is formed by the same number of adjacent child sub-trees under the same parent node. Novel mining algorithm called MiBAT which makes use of domain constraints to acquire anchor point information. Our methodology accomplishes an exactness of 98.9% and a review of 97.3%

concerning post record extraction. On page level, it lawlessly handles 91.7% of pages without removing any o_-base posts or missing any brilliant posts.

This paper [8] depicts a framework for robotized continue data extraction to help fast resume search and the board. The framework is equipped for extricating a few significant educational fields from a free arrangement resume utilizing a lot of common language handling (NLP) strategies. We depict a working framework, for programmed continue the board. The framework is equipped for extricating six significant fields of data as characterized by HR-XML In this the main layer is made out of a few general data squares, for example, individual data, instruction and so forth. The second layer of structure is inside the principal layer and contains explicit data comparing to the layer 1. For instance, the layer 1 individual data square comprises of layer 2 data like name, address and email. While this probably won't be valid for every one of the resumes, the structure is by all accounts held in the greater part of resumes. Furthermore, the area of the data (like name, age and so forth) in resumes differs fundamentally from resume to continue. Our framework can chip away at both layered structure and unstructured resumes. Data extraction module is made out of a few sub modules, every one of which plays out the undertaking of removing explicit data. The primary sub modules are (a) Qualification module, (b) Skill module (c) Experience module and (d) individual data extraction. While the capability extraction sub-module separates the graduating college name, degree and the class acquired. The aptitudes extraction module extricates the abilities of the applicant. Experience extraction module is competent or removing the all-out understanding, in any event, when this data isn't expressly referenced in the resume of the candidate. The extraction procedure utilizes a lot of language preparing systems which are part heuristics and part example coordinating. Test results completed on countless resumes demonstrate that the proposed framework can deal with an enormous assortment of resumes in various record positions with an exactness of 91% and a review of 88%.

In this paper [9] we present a near investigation of 5 estimates utilizing distinctive vector loads done over an enormous arrangement of French list of qualifications. The point is to know how these measures act and whether they approve the idea that chose list of qualifications have more in a similar manner as themselves than with the dismissed list of qualifications. We utilize NLP systems and ANOVAs to do the relative examination. The outcomes demonstrate that the determination of measures and vector loads must not be viewed as insignificant in e-Recruitment projects; especially in those where the resumes' resemblance is estimated. Something else, the outcomes may not be dependable or with the normal execution. Four sorts of archives are dissected in this work: pdf, Microsoft Word, Open Document Text and Rich Format Text.

In this paper [10] the general target of this examination was to concentrate such information as experience, highlights, and business and training data from resumes put away in HR archives. In this article, we propose a philosophy driven data extraction framework that is intended to work on a few million free-design printed resumes to change over them to an organized and semantically improved rendition for use in semantic information mining of information basic in HR forms. The engineering and working instrument of the framework, similitude of the idea and coordinating strategies, and a deduction system are presented, and a contextual investigation is displayed.

According to [11] a keywords extraction based on CRF is proposed and implemented. As far as we know, using CRF model in keyword extraction has not been investigated previously. Experimental results show that the CRF model outperforms other machine learning methods such as support vector machine, multiple linear regression model etc. in the task of keywords extraction. In keyword extraction, words occurred in the document are analyzed to identify apparently significant ones, on the basis of properties such as frequency and length. In keyword assignment, keywords are chosen from a controlled vocabulary of terms, and documents are classified according to their content into classes that correspond to elements of the vocabulary.

In this paper [12] a hybrid approach that employs conceptual-based classification of resumes and job postings and automatically ranks candidate resumes (that fall under each category) to their corresponding job offers. In this context, we exploit an integrated knowledge base for carrying out the

classification task and experimentally demonstrate - using a real-world recruitment dataset- achieving promising precision results compared to conventional machine learning based resume classification approaches. In this context, each and every resume in the resumes collection will be matched to the offered job post instead of matching only those that fall under the corresponding occupational category.

According to [13] a novel framework, not depending on the file format, to extract knowledge about the person for building a structured resume repository. The proposed framework includes two major processes: the first is to segment text into semi-structured data with some text pretreatment operations. The seconds to further extract knowledge from the semi-structured data with text classifier. This work aim to improve the accuracy of building resume repository for head-hunters and companies focus on recruiting.

According to [14] a machine learned solution with rich features and deep learning methods. Our solution includes three configurable modules that can be plugged with little restrictions. Namely, unsupervised feature extraction, base classifiers training and ensemble method learning. In our solution, rather than using manual rules, machine learned methods to automatically detect the semantic similarity of positions are proposed. Then four competitive "shallow" estimators and "deep" estimators are selected. Finally, ensemble methods to bag these estimators and aggregate their individual predictions to form a final prediction are verified.

According to [15] Based on the information we ranked individual skills of the user. Using Natural Language Processing(NLP) and (ML)Machine Learning to rank the resumes according to the given constraint, this intelligent system ranks the resume of any format according to the given constraints or the following requirements provided by the client company. We will basically take the bulk of input resume from the client company and that client company will also provide the requirement and the constraints according to which the resume shall be ranked by our system. Moreover the details acquired from the resumes, our system shall be reading the candidates social profiles which will the more genuine information about that candidate

Literature Survey on tools

Apache Tika

- Apache Tika is a content type detection and content extraction framework. Tika provides a general application programming interface that can be used to detect the content type of a document and also parse textual content and metadata from several document formats.
- The Tika works on various existing parser libraries such as Apache POI for Microsoft formats, PDFBox for Adobe PDF, Neko HTML for HTML etc.
- The Tika API is stream oriented so that the parsed source document does not need to be loaded into memory all at once but only as it is needed.
- Tika supports directly around 30 document formats. See list of supported document formats.

1. Package Formats :- .tar , .jar , .zip , .bzip2 , .gz , .tgz
2. Text Document Formats :- .doc , .xls , .ppt , .rtf , .pdf , .html , .xhtml , open document
3. Image Formats :- .bmp , .gif , .png , .jpeg , .tiff
4. Audio Formats :- .mp3, .aiff, .au , midi , .wav
5. Misc Formats :- .pst, .xml ,.class.

Functionality

The two main functionalities Tika offers are

Mime Type detection :- Tika contains a class named AutoDetectParser that uses mime type detection functionality to find out the mime type of a file and then uses that information to dispatch the parsing task to a parser that can understand the format.

Content parsing :- The most important capability of Tika is parsing content. Tika provides a thin wrapper/adapter on top of existing parsers, defined by the Parser interface.

Parameters :-

It takes in just three parameters.

1. The (input) parameter stream is needed so the parser can read the raw data of document.

2. The (output) parameter handler is used to send callback notifications about the logical content of a document back to your application.

- The handler interface is of type org.xml.sax.ContentHandler and it is exactly same interface that is used in Java SAX 2.0 API.

Finally,

3. The metadata (input/output) parameter provides additional data to the parser as input and can return additional metadata out from the document. Examples of metadata include things like author name, number of pages, creation date, etc.

Tools of NLP

1. **CoreNlp**:-

➢ Advantages
-Simple to use
-Fast Serialization
-Maintains Thread Safety

➢ Disadvantages:-
- Less Customizability
- More Clunky than other interface
- Non Deterministic

2. **NLTK(Natural Language Tool Kit)** :-

➢ Advantages:-
-The most well-known and full NLP library
-Many third party Extensions
-Fast Sentence tokenization
-support largest number of languages.

➢ Disadvantages
-works very slow
-Low accuracy
-tokens do not align to original strings
-Models return lists of strings
-no word vector support
-complicated to learn and use

3. **SpaCy** :-

➢ Advantages:-
-Fastest NLP
-Easy to learn
-uses neural networks for training some models

➢ Disadvantages
-not supported multiple languages
-Lexical Resources are not available
-Lacks Flexibility comparing to NLTK
-Sentence tokenization is slower

4. **Gensim** :-

➢ Advantages:-
-works with large database
-Provides TFIDF vectorization
-Support deep learning

➢ Disadvantages
-Doesn't have enough features to provide NLP pipeline
-Model wrappers are not working properly
- less semantic search features
-data streaming are low captures with Tika

Proposed System

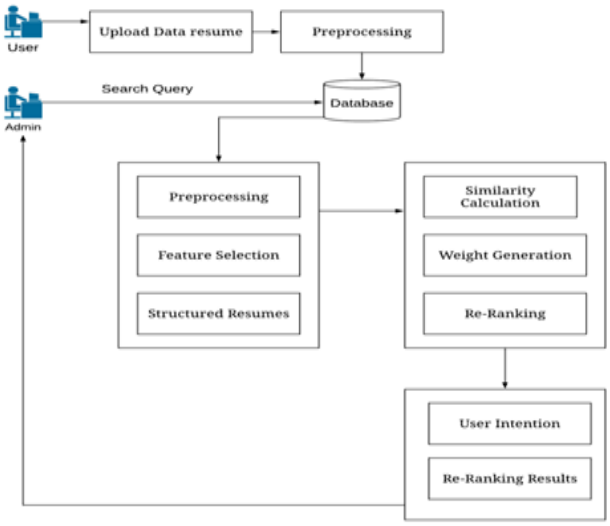

Figure 1: Proposed system architecture

The proposed system follows below procedure for entire execution to convert unstructured data to structured conversion in entire process.

- Initially some raw resume data has given input to system, it should be unstructured format. (it should be doc, pdf file)
- Read data from entire document and apply stop word removal as well as porter stemming algorithm to get the lemmas features.
- Natural Language Processing (NLP) is another technique has used to extract the features from text

using dependency parser.

- To identify the name of specific entity has used Name Entity Recognizer of Stanford NLP.
- Once semi-structured format has done, selected feature has place in structured format and using any respective machine learning algorithm.
- Once clarification has done we calculate the confusion matrix for entire test data and predict the precision, recall, accuracy etc. respectively.

Algorithm Design

1 : Stop word Removal Approach

Input: Stop words list L[], String Data D for remove the stop words.

Output: Verified data D with removal all stop words.

Step 1: Initialize the data string S[].

Step 2: initialize a=0,k=0

Step 3: for each(read a to L)

If(a.equals(L[i]))

Then Remove S[k]

End for

Step 4: add S to D.

Step 5: End Procedure

2 Stemming Algorithm.

Input : Word w

Output : w with removing past participles as well.

Step 1: Initialize w

Step 2: Intialize all steps of Porter stemmer

Step 3: for each (Char ch from w)

If(ch.count==w.length()) && (ch.equals(e))

Remove ch from(w)

Step 4: if(ch.endswith(ed))

Remove ‘ed’ from(w)

Step 5: k=w.length()

If(k (char) to k-3 .equals(tion))

Replace w with te.

Step 6: end procedure

Conclusion

Based on the proposed experimental analysis this system will provide better and efficient solution to current hiring process. This will provide potential candidate to the organization and the candidate will successfully be placed in an organization which appreciates users skill set and ability and speed up the whole hiring process. To work with various kinds of large unstructured data will be future work for such systems

References:-

[1] Özçevik, Yusuf, Fatih Yücalar, and Murat Demircioğlu. "Determining a Proper Text Similarity Approach for Resume Parsing Process in a Digitized HR Software." Celal Bayar University Journal of Science 18.4 (2022): 371-378.

[2] Tyagi, Rinki, et al. "Resume Builder Application." International Journal for Research in Applied Science and Engineering Technology (IJRASET) Volume 8 (2020).

[3] Mittal, Vrinda, et al. "Methodology for resume parsing and job domain prediction." Journal of Statistics and Management Systems 23.7 (2020): 1265-1274.

[4] Deepak, Gerard, Varun Teja, and A. Santhanavijayan. "A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm." Journal of Discrete Mathematical Sciences and Cryptography 23.1 (2020): 157-165.

[5] Tejaswini, K., et al. "Design and development of machine learning based resume ranking

system." Global Transitions Proceedings 3.2 (2022): 371-375.

[6] Shuang Lin, Jie Chen, Zhendong Niu, \Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction \ , August 2017

[7] Xinying Song, Jing Liu, Yunbo Cao, Chin-Yew Lin, and Hsiao-Wuen Hon , "Automatic Extraction of Web Data Records Containing User-Generated Content", Jan. 2015

[8] Sunil Kumar Kopparapu , \Automatic Extraction of Usable Information from Unstructured Resumes to Aid Search", Oct. 2016

[9] Luis Adri_an Cabrera-Diego1,2, Barth_el_emy Durette2, Matthieu Lafon2, Juan-Manuel Torres-Moreno1,3 and Marc El-B_eze1, "How Can We Measure the Similarity Between R_esum_es of Selected Candidates for a Job? Luis Adri_an Cabrera-Diego1,2, Barth_el_emy Durette2, Matthieu Lafon2, Juan-Manuel Torres- Moreno1,3 and Marc El-B_eze1" , Jan. 2014

[10] Duygu C_EL_IK,"Towards a semantic-based information extraction system for matching resumes to job openings", June 2016

[11] Zhang, Chengzhi. "Automatic keyword extraction from documents using conditional random fields." *Journal of Computational Information Systems* 4.3 (2008): 1169-1180

[12] Zaroor, Abeer, Mohammed Maree, and Muath Sabha. "A hybrid approach to conceptual classification and ranking of resumes and their corresponding job posts." *International Conference on Intelligent Decision Technologies*. Springer, Cham, 2017.

[13] Chen, Jie, Zhendong Niu, and Hongping Fu. "A novel knowledge extraction framework for resumes based on text classifier." *International Conference on Web-Age Information Management*. Springer, Cham, 2015.

[14] Lin, Yiou, et al. "Machine learned resume-job matching solution." *arXiv preprint arXiv:1607.07657* (2016).

[15] Sayed Zainul Abideen Mohd Sadiq, Juneja Afzal Ayub, Gunduka Rakesh Narsayya, Momin Adnan Ayyas, Prof. Khan Tabrez Mohd. Tahir , “Intelligent Hiring with Resume Parser and Ranking using Natural Language Processing and Machine Learning " , April 2016