IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 7, April 2024

CV Classification using NLP

Mr. Kshirsagar Saurabh S, Mr. Aher Devesh K, Mr. Shinde Vaibhav V,
Mr. Wabale Nilesh N, Prof. Mundhe.B.
Department of Computer Engineering
Sahyadri Valley College of Engineering, Rajuri, Maharashtra, India

Abstract*:* *There has been a large influx of jobs due to the rapid urbanization. This has led to the creation of large organization that requires equally large number of varied employees to perform their tasks. To accumulate talent and useful employees, the organizations post job offers. Most of the time, due to high demand and the low supply of jobs and increasing unemployability, there is an inordinate number of job applications that are received by the organization. The companies are highly motivated to select the candidate with the best personality. Doing it from the large number of resumes or Curriculum Vitae received is a daunting task. Therefore, there is a need for the implementation of an effective and reliable automatic personality assessment system. There are a few approaches that have been researched for this purpose of personality assessment system. Most of these approaches have a number of drawbacks that need to be addressed. Therefore, this publication deals with the implementation of an effective personality evaluation system through the analysis of candidate CVs. The proposed system utilizes the Bag of Words and protocol estimation along with protocol mapping and Decision Tree approach to achieve highly precise Personality classification by using extensive concepts of NLP.*

Keywords: Curriculum Vitae, Natural Language Processing, Bag of Words, Decision Tree

I. INTRODUCTION

The increasing population is also one of the reasons that most of the large corporations receive an increasing number of applications for a particular job posting. Most of these large corporations utilized all different means to achieve applications such as mail, post along with online services which form most of the applications that are provided on this day. The platform of online recruitment has been increasing significantly for the past few years which has led to a large influx of job applications for a particular job vacancy in the organization. This large influx is very difficult to ascertain the right employee for a particular job as manually segregating and classifying these documents gets difficult after a point of time for a person. Therefore, an innovative approach for the automatic segregation of resumes is needed so that there can be an efficient and useful utilization of the recruitment period.

The fuzzy classification paradigm is also one of the most useful platforms that are utilized for classification. The difference between various classification techniques concerning classification is that the fuzzy classification paradigm provides a complete classification of the provided data. This is done by deploying a different classification system that is not dependent on the absolute truth and false values of the data under consideration. Instead, the fuzzy classification paradigm employs the use of various degrees of association to the truthiness of the data associated with that particular label such as high very high low very low.

These values provide an effective fuzzy classification of the resumes along with the introduction of the NLP paradigm.

II. LITERATURE SURVEY

A. Zaroor explains that there has been a shift from the traditional hiring techniques towards the paradigm of online recruitment in most of the job portals online. This is highly useful and convenient for a large number of users as it allows them to apply for a job with increase convenience of their homes. This is also useful for organizations as they can shortlist people according to the details offered in the resume [1]. Due to the increase in the volume of the resumes, it becomes highly difficult for manually classifying the resumes according to the open posts. Therefore, the purpose of automatically classifying job post in resumes a classification system has been proposed by the authors. The experimental results conclude that the proposed methodology achieves high effectiveness and efficiency.

DOI: 10.48175/568

S. Nasser states that there has been an increase in the number of online applications to apply to various job posts. There has been an increase in the population which results in an increased number of job applications being posted on various portals and companies. The job of manually classifying these resumes is a highly difficult and complex procedure which can also result in a large discrepancy [2]. Therefore, the purpose of the classification of the documents the authors in this paper presents an innovative technique for classification of the resumes through the use of glove word embedding and convolutional neural networks.

C. Ayishathahira elaborates on the platform of resume parsing as it is the most novel and currently improving that helps in automatically classifying various resumes. This is highly useful as there have been increased demands for an online approach to uploading resumes and applying for jobs. Various Job Recruitment procedures have been getting streamlined and automatic to promote improved and useful job hiring process [3]. Therefore, for this purpose, the authors in this paper have outlined and innovative techniques for resume parsing through the use of conditional random fields and neural networks.

F. Javed discusses the various problems faced by the job recruitment procedures and relevant approach to maximize job hiring capabilities [4]. These approaches are inherently useful as it allows for an effective methodology and implements techniques for reducing unemployability across a large margin. The authors convey that there has been an increase in the influx of resumes that are being posted for application to a particular job this increase in the volume puts undue pressure on the workers at the recruitment firms. Therefore, the authors in this process outline carotene which is a job classification system that utilizes the machine learning platform for efficient and automatic job recruitment classification.

III. SCOPE OF THE PROJECT

Scope of CV Classification is applying in many areas as mentioned below:

3.1 Resume Parser

A Resume Parser is an AI-based software that parses resumes using NLP (Natural Language Processing). It eliminates manual data entry by extracting candidates' information intelligently and saves it in pre-designed fields.

[1] A resume parser is a compiler or interpreter that converts the unstructured data into a structured form.

[2]It is a component that automatically segregates the information into various fields and parameters like contact information, educational qualification, work experience, skills, achievements, professional certifications to quickly help you identify the most relevant resumes based on your criteria.

[3]A parser takes input in the form of a sequence of program instructions and tends to build a data structure, a "parse tree," or an abstract syntax tree.

- Keep in mind the following features while selecting a CV/resume parser :
- Parses resumes of all formats, such as PDF, doc, docx, HTML, RTF
- Easy to integrate with your existing software
- It contains a detailed library of taxonomies to identify candidate skills
- Choose a multilingual resume parser that automatically identifies region and language to parse information.
- Allows the user to enable or disable fields from resumes as per requirement using 'configuration feature.' It helps in promoting unbiased recruitment.
- It should extract the complete resume information in maximum data fields.
- Creates an executive or management summary so that recruiters can evaluate a candidate by reading this summary.
- Uses deep learning algorithm for improved extraction and smarter identification of resume data for better search results.
- Bulk import allows a resume/job parser to parse multiple resumes/jobs in a go.
- Email inbox integration allows users to parse resumes/jobs from single or multiple email inboxes.
- Option to integrate RScript plugin directly to your web page within 2 min.
- Get the parsed data in a document template designed to bring uniformity to the presentation.

IV. METHODOLOGY

The methodology for CV Classification developed under waterfall model architecture as shown in the below figure 1.

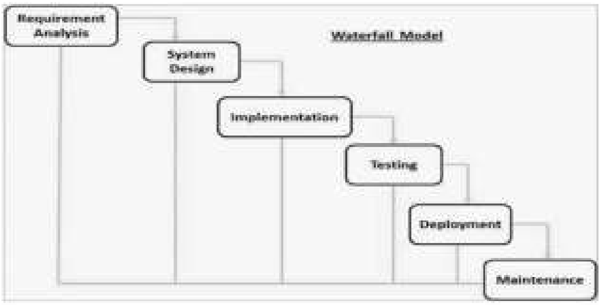

Fig 1 : Water fall model Architecture

The sequence phases in water fall model according to our project are mentioned below.

Requirement Analysis – Here requirement analysis are done based on following points

- Base paper for CV Classification
- Studying on Natural Language Processing

System Design: The System of CV Classification is designed by using the following hardware andsoftware’s

Minimum Hardware Specification:

- CPU : Core i3
- RAM : 4 GB
- HDD : 500 GB

Software Specification:

- Coding Language : Java,
- Development Kit : JDK 1.8
- Front End : Java Swing
- Development IDE : Net beans 8.2
- Data Base : My SQL 5.0
- External API : Mysql Connector

Implementation:

Proposed system is designed by using the following modules

Module A: Preprocessing

- Input: CV Folder and Attributes
- Process: CV String stop word and Stemming
- Output: Preprocessed data

Module B: Bag of Words

- Input: Preprocessed data
- Process: Bag of word frequency Estimation
- Output: Word Frequency List

Module C: Protocol Estimation

- Input: Word Frequency List and CV String
- Process: Protocol Word frequency
- Output: Protocol List

Module D: Decision Tree

- Input: Protocol List
- Process: IF-then Rules
- Output: Personality Classification

Integration and testing:

Testing is done based on the following techniques

- Performance testing
- Recovery testing
- System Testing
- Security testing
- Integrating testing
- Deployment of the system:

The developed software is deployed in the laptop of above mentioned configuration with the help of the mentioned software.

Maintenance of the system:

As this software is tested for the quick recovery, so maintenance of the system is not a challenging task. This is because the tools and the software used are open source, so there is no question of licensing the required software.

V. WORKING AND PROCESSES

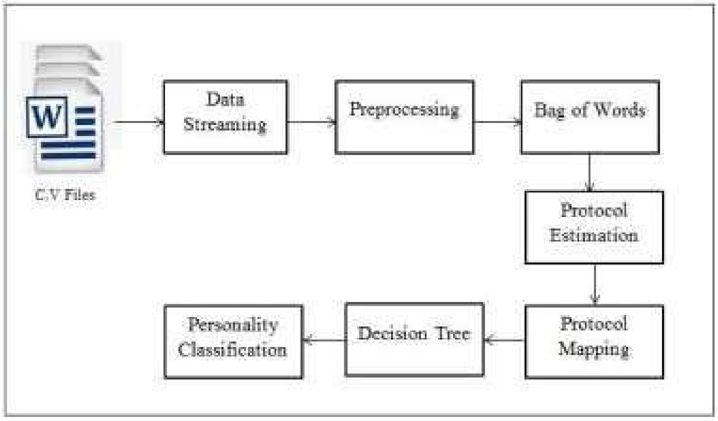

FIG. SYSTEM OVERVIEW

The presented technique for Classification of the personalities through the CVs is illustrated in figure given above and the procedure that is executed to implement this system is detailed below.

Step 1: Data Streaming –This is the first step of the proposed methodology in which the requisite data is provided to the system for the purpose of utilizing the Curriculum Vitae for personality classification. In this step of the procedure a number of CV's are provided as in input from a folder. This folder contains a number of CVs that are supposed to be classified. The classification is done through a given criteria. The criteria for the classification of the CVs are also provided as an input to the system, where the gender and qualification form the basis of classification.

The Apache POI is used for the purpose of reading the resume doc files in a line by line manner. All the CVs in the folder given as an input are read through the Apache POI line by line. The extracted contents of the resume are then subsequently stored in a list and transferred to the next step of the procedure.

Step 2: Pre-processing – The output list of the CVs achieved in the previous step is taken as an input for the pre-processing step. The objective section from the CV is extracted as a string and subjected to the preprocessing procedure as detailed below.

DOI: 10.48175/568

Special Symbol Removal – Special Symbols are symbols that are used for the purpose of providing a structure to the English language while speaking. These are not as important in our implementation as they do not provide any additional information and are subsequently purged from the string.

Tokenization – The special symbol removed string is then given as input for this procedure which performs the tokenization of the string. This process divided the string into smaller tokens for the creation of a well indexed string. This is done to enable the easy conversion of the string into an array list for utilization in the further processing in the proposed methodology.

Stopword Removal – The English language has a collection of words that are purely to provide aesthetic to the language and the sentence formation. These words are known as stop words that denote pauses. These are specifically used for joining two sentences together to form an uninterrupted flow of the conversation. These words do not provide anyadditional meaning to the string and can be removed without any change in the meaning of the string.

For example, the phrase "going to walk" if subjected to the stopword removal procedure of the preprocessing. The stopword in this phrase is "to" which is be removed and the phrase transformed into "going walk". This example depicts that the stopwords removal procedure does not change the meaning of the string.

Stemming – In the English language there are a lot of words that are added with a suffix to denote the timing or the tense of the words. This is done to ensure that the conversation is precise and to the point for the individual. This is not necessary in out implementation as it does not change the semantics of the string in any way. Therefore, in the stemming process the words are purged off their suffixes and this makes processing the newly shortened string significantly easier and faster.

For example, "sleeping" will be reduced into "sleep" by the removal of the substring "ing" which is replaced with an empty character in its position. It can be noticed that there is no semantic difference between "sleeping" and "sleep". But stemming can have a significant impact on the time and resources needed for the processing of the string in the methodology further.

Step 3: Bag of Words and Protocol Estimation – The bag of words is a collection of predefined words that are selected for the purpose of personality estimation. The words describing the personality of the candidate from the objectiveand hobbies are utilized in this step. The extracted words are utilized for the comparison through the link https://grammar.yourdictionary.com/style-and- usage/words-that-describe-personality-traits.html.

The extracted words are utilized for the purpose of comparison through the use of the protocol estimation module in the proposed methodology. The protocol estimation extracts the relevant hobbies along with the words used in the objective according to the link and stored into a list. The module then compares it with the bag of words and produces a list of comparisons according to the candidates.

Step 4: Protocol mapping and Decision tree – This list of comparisons produces a consolidated list which eventually contains each candidate's name and obtained personality, respectively. Now a unique list of personality is created using hash set class of java. This obtained unique list examined thoroughly for the user's list for the obtained personalities. This process ultimately yields the classified personalities of the candidate for whom the CV's are being uploaded.

This process can be depicted in the below mentioned algorithm 1.

ALGORITHM 2: CV CLASSIFCATION

```
//INPUT : UNIQUE PERSONALITY LIST UPL
//INPUT : ALL CANDIDATE PERSONALITY LIST ACL
//OUTPUT:PERSONALITY CLASSIFIED LIST PCL FUCNTION: PERSONALITYCLASSIFICATION (ACL, PCL)
START
PCL = ∅
FOR I=0 TO SIZE OF UPL
    SL= ∅ [SL = SINGLE LIST]
    PER = UPL[I]
    FOR J=0 TO SIZE OF ACL
        RL = ACL[J]
        NAME = RL[0]
```

DOI: 10.48175/568

```
9:PL = RL[1] [PL =PERSONALITY LIST]
10:IF (PL CONTAINS PER), THEN
11:SL= SL + NAME 12: END IF
13: END FOR
14:PCL = PCL + SL
15: RETURN SL
16: STOP
```

VI. CONCLUSION

So, this is the last chapter of report where there is whole conclusion of the report. We herby conclude the importance of user management system in every website. How necessary they are for better interaction with the users. We learnt the importance of confidentiality of the user data. How website access should be restricted to the members who have registered. The marketing value of adding more users to the website and better interaction of user with the website.

VII. ACKNOWLEDGMENT

We would like to take this opportunity to express our sincere gratitude to our Project Guide Prof. Mundhe Bhalchandra b. for her encouragement, guidance, and insight throughout the project and in the preparation of this project. She truly exemplifies the merit of technical excellence and academic wisdom. Her extensive knowledge, serious research attitude and encouragement were extremely valuable to us. We also appreciate not only for her professional, timely and valuable advices, but also for her continuous scheduled follow up and valuable comments during our project work. We are also very grateful to Prof. Mundhe Bhalchandra B. (HOD, Computer Science Department) for devoting time from her busy schedule to provide valuable comments and suggestions on this work. Special thanks go to Prof. Balaramadu P. (Principal) for his encouragement to complete Engineering in time.

REFERENCES

[1]. A. Zaroor et al, “JRC: A Job Post and Resume Classification System for Online Recruitment”, International Conference on Tools with Artificial Intelligence, 2017.

[2]. S. Nasser et al, “Convolutional Neural Network with Word Embedding Based Approach for Resume Classification”, International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR), 2018.

[3]. C. Ayishathahira et al, “Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing”, International CET Conference on Control, Communication, and Computing (IC4), 2018.

[4]. F. Javed et al, “Carotene: A Job Title Classification System for the Online Recruitment Domain”, IEEE First International Conference on Big Data Computing Service and Applications, 2015.

[5]. A. Taherkhani et al, “Multi-DL-Resume: Multiple neurons Delay Learning Remote Supervised Method”, International Joint Conference on Neural Networks (IJCNN), 2015.

[6].Z. Jiang et al, “Research and Implementation of Intelligent Chinese resume Parsing”, International Conference on Communications and Mobile Computing, 2009.

[7].Z. Chuang et al, “Resume Parser: Semi-structured Chinese document analysis”, World Congress on Computer Science and Information Engineering, 2009.

[8].Y. Wentan et al, “Chinese resume information extraction based on the semi- structured text”, 36th Chinese Control Conference July 26-28, 2017.

[9]. J. Chen et al, “A Two-Step Resume Information Extraction Algorithm”, Hindawi Mathematical Problems in Engineering Volume 2018.

[10]. P. Das et al, “A Review on Text Analytics Process with a CV Parser Model”, 3rd International Conference for Convergence in Technology (I2CT), 2018.