IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 11, No. 4, April, 2024, pp. 1044-1052
Journal homepage: http://iieta.org/journals/mmep

Rule-Based Information Extraction from Multi-format Resumes for Automated Classification

Dhiaa A. Musleh

Department of Computer Science (CS), College of Computer Science and Information Technology (CCSIT), Imam Abdulrahman Bin Faisal University (IAU), P.O. Box 1982, Dammam 31441, Saudi Arabia

Corresponding Author Email: daMusleh@iau.edu.sa

https://doi.org/10.18280/mmep.110422

Received: 2 July 2023
Revised: 10 October 2023
Accepted: 22 October 2023
Available online: 26 April 2024

Keywords:
information extraction, text and data mining, document classification, PDF resume, rule based system, Python language, NLP

ABSTRACT

Nowadays, with the expansion of the Internet, a lot of people publish their resumes on the internet and social media networks. Large companies receive hundreds of resumes per day, which comes in several formats such as Joint Photographic Experts Group (JPG), Portable Document Format (PDF) and Word files. Therefore, information extraction from resumes can be applied automatically by several methods. In this research, the important details that are taken from resumes are: name, date of birth, email, phone number, GPA, gender, nationality, and address. The private resumes dataset used is taken from different sources including open source as well as personally annotated. The processes of information extraction for resumes have been performed in different phases such as: pre-processing, converting the resumes files into PDF and information extraction by the rule-based method to extract the eight elements from resumes. To carry out the experiment, the Python language is used, particularly the spacy library and word2vec technique. Consequently, the experimental results demonstrate that the testing phase achieved 96.4% information extraction precision which is quite considerable in contrast to the techniques in the literature. The scheme is then extended to classify the resume based on the extracted information fields and exhibited classification accuracy, precision, recall and F1-score as 98.02%, 98.01%, 98% and 98%, respectively.

1. INTRODUCTION

At present, the number of applicants for different jobs has increased dramatically, resulting in an abundance of resumes greatly. Therefore, the individual may write different versions of their resume depending on the job they aim to apply for or because of updating their personal information. For example, when an individual intends to apply for a faculty position, they more accurately list their teaching experience while when applying for an industrial position, such as companies, they write their list of practical experiences more comprehensively. Also, social media platforms like LinkedIn and Twitter allow individuals to share their digital identities online to communicate with different societies around the world.

The process of extracting information is the automatic retrieval of specific information related to a specific subject from a text using information extraction tools, which may be from an unstructured, semi-structured, or structured text. Therefore, the process of extracting information from resumes and retrieving them accurately is not an easy process. The reason behind the difficulty of this matter is the lack of a specific format for all resumes as they come in different languages and different types of files, such as JPGs, PDFs, and Word. A resume can be described as a multi-section document, with a description for each part, focusing on various aspects of an individual's professional details [1]. The information extraction from resumes has become more significant to many companies. It is more important for large companies because they receive hundreds of resumes daily. Writing resumes has different reasons, but often used is for applying for jobs. Also, resumes are written in several languages and several formats such as PDF and Word, and it also has several patterns such as tables and drawings. Therefore, it is difficult to extract information from resumes because it does not have a specific predefined format [2].

Various approaches have been proposed in the literature to information extraction and classification. The classification is categorized as single and multilabel whereas the documents are categorized as structure, semi-structured and unstructured. The classification accuracy is mainly based on the quality of extracted information [2-4].

This paper focuses on extracting information from resumes by using rule-based methods. The format used in this research for all resumes is PDF. Furthermore, Python language is used for coding the regular expression to extract information. Finally, the comparison between our approach and some studies are presented. Furthermore, to make the process of extracting information from resumes easier, we seek to make it automatically by exploiting the main information that is extracted from resumes. There are many ways to extract

information from resumes such as machine learning and rule-based methods. In this research, we relied on a regular expression by using the Python language to extract basic information from resumes. The important details that are taken from resumes are: name, date of birth, email, phone number, GPA, gender, nationality, and address. Additionally, the dataset used is private resumes from different sources. To extract information from resumes, the extraction process can be done in three phases: pre-processing, converting the resumes files into PDF and information extraction by the rule-based method to extract the eight elements from resumes. Also, the Python language is used to implement the experiment. The result showed that the training set has 97.1% while the testing set has 96.4% accuracy. The objectives of this work are stated as follows:

- Automating the process of extracting information from resumes.
- The challenges to extract information from resumes accurately.
- Increased demand for resumes of job applicants in the world.
- Facilitating the process of selecting candidates for jobs in companies.

The rest of this paper is organized as follows: related work is presented in Section 2 followed by the proposed approach methodology given in Section 3. Section 4 highlights the experimental part and contains a description of the dataset and the result and discussion are presented. Finally, Section 5 presents the conclusion and future work.

2. LITERATURE REVIEW

Information extracted from resumes would help to save time and effort. Much research has been carried out on extracting information from resumes in various formats. The previous related works are presented in this section chronologically.

Li and Sun [5] studied a framework that contained two main processes: the first used the Tika tool to remove the formats from resumes, then dividing the raw text of resumes; the second extracted the knowledge from the semi-structured data. Jayaraj et al. [6] proposed an approach to select specific information from a resume by feature extraction model using the Term Frequency-Inverse Document Frequency (TF-IDF) method. The TF-IDF measure contains two disconnected measures: the first is that the TF part cares about the frequency of a term occurring within a document, and the second, the IDF part cares about the occurrence of the inverse frequency of a term through all documents in a set. So, Celik [7] made use of an ontology-driven system for extracting information from resumes written in English and Turkish languages. The system converted the resumes into the ontological format. Moreover, they used semantic information extraction instead of syntactic information extraction. On the other hand, Ahmad et al. [8] extracted information from research papers in PDF format based on a rule-based system. They converted the PDF format of a research paper into XML and in a plain text format to extract the information. The Information extracted from the research paper was authors, their country and affiliation, table caption, first level section heading funding agency name, figure caption, and the project name.

Mhapasekar [9] proposed a system for extracting information from resumes in three main phases: firstly, reading the Rich Text Format (RTF) document, which includes unstructured information, secondly, reading the ontological structure and comparing it with the unstructured document, thirdly, storing the information that has been extracted in the Microsoft Access (MS ACCESS) database. To generate the ontological structure, the regular expression in Practical Extracting and Reporting (PERL) language is used. Yan and Qiao [10] proposed a Chinese Resume Information Extraction System (CRIES). The system used two extraction algorithms which are the Hypertext Markup Language (HTML) document abstraction algorithm and the plain text abstraction algorithm to take the information out from resumes. Additionally, they did an experiment by using the Java programming language. Ghufran et al. [11] used a dataset of 153 resumes written in English. They proposed an approach of five modules: the section delimiter module, the n-gram generator module, the n-gram interpretation graph constructor which is for each n-gram, and the interpretation filters module which is to get rid of some interpretations according to two criteria. At the end, they have annotated resumes and these resumes can be used by the web resource getter module. Jayaram and Sangeeta [12] reviewed several techniques for extracting information from research documents in a PDF format. They found that machine learning was more important in achieving more accurate results and a lower propagation rate of errors.

Chen et al. [13] suggested an algorithm in two stages; they converted the different text files to a raw text to classify the text block to obtain the information, and then they added information following the new writing style forms. Ayishathahira et al. [14] produced a method with four steps: first, extracting the raw text. The Second is preprocessing to delete noise in a text such as punctuation. The third is splitting the text into blocks. Finally, information extraction after merging all information and converted it into a JSON file. They applied two deep learning models, namely, convolutional neural networks (CNN) and Bi-directional Long Short-Term Memory (Bi-LSTM) to extract the information. The result obtained for the personal class was that CNN has 93% precision which is higher than Bi-LSTM.

Sajid et al. [15] presented a study to classify the research articles based on the references inside. The scheme exploits the extracted references for sake of article classification, because the references contain the most relevant topics to the article's scope and area of the field. They further extended their work to multi-label document classification using metadata only and proven their work to be comparable to that of full text-based approaches [16].

Similarly, Zaman et al. [17] presented a very interesting approach to information extraction from the published research article by means of a novel framework. The framework is named as ontological information extraction equipped with several interesting ingredients such as fuzzy regular expressions, word sense disambiguation, word2vec and ontology building [18]. The scheme was quite interesting and has achieved good accuracy and precision scores.

Two natural language processing (NLP) models were developed by Vukadin et al. [19] to extract meaningful information, such as name, address, and role, from unstructured (free form) resumes. The proposed models were built using resumes written in five different languages, including English, Swedish, Norwegian, Polish, and Finnish. The first model collected "hard" information from the resumes, such as names and organizations, and it contextualized the information by dividing the classified items

into sections. This model archives F1 score of 0.825 at the item level and 0.833 at the section level. The second model, which checked if each skill found in a resume has a corresponding quality, was used to identify the level of self-assessed skill competence and F1 score of 0.616 was achieved by this model.

Gaur et al. [20] have utilized a resume database of 550 resumes to capture the educational qualifications, knowledge, and skills relevant to the job. The proposed model was trained using BILOU encoding scheme, CNN Layer and Bi-LSTM layers. The authors have used this proposed model to extract only the education section of the resumes which includes the degree and institute name. The proposed model achieved high performance with an F1 score of 73.28 and an accuracy of 92.06%. Wosiak [21] proposed a hybrid, multi-module, system for the automated extraction of information from Polish resume documents in the IT industry. The proposed system combines methods for named entity recognition with dictionary methods. Three Name Entity Recognition tools were used, namely Liner2, NERF and Babelfy. To verify the quality of the proposed hybrid module comparing with the individual modules, different experiments were conducted on real data provided by an IT recruiting company. The obtained results show that the hybrid solution achieves a much higher efficiency compared with the best single solution.

Deep Learning models have recently proved to be very effective for Named Entity Recognition because they can handle different document formats, languages, and layouts. Barducci et al. [22] proposed end-to-end resume information extraction framework for Italian Labor Market. The proposed system uses linguistics patterns to segment the extracted raw data into semantically consistent parts. Each segment is further processed with a Named entity recognition (NER) algorithm, based on pre-trained language models. The utility of the proposed framework was assessed in a real company's human resource department. Bhoir et al. [23] investigated how well deep learning models work at extracting relevant information from resumes and putting it in structured formats for future examination. The proposed parser used pre-trained deep learning model called Spacy Transformer BERT to capture the text's semantic meaning, and Spacy to employ NLP to glean relevant information from it. The findings demonstrated that the proposed parser was extremely accurate in locating relevant data, such as candidate names, email, contact information, qualifications, job experience, and other crucial details helpful in the hiring process. Musleh et al. [24] presented an Arabic key phrases extraction approach using various Arabic documents as the dataset. By the experiment the authors revealed their technique was superior to the key phrase (KP)-Minor approach in the literature. Similarly, Batool et al. [25] presented an approach to extract the references from Wikipedia dump file using rule-based system and classify them into the category such as conference, journal, book, and chapter. The scheme exhibits an average accuracy of 97.5%.

Rahman et al. [26] and Zaman et al. [27] presented assessment of information extraction approaches, model, and techniques by surveying more than a decade literature in the area. In this regard, they performed a comprehensive and systematic literature review.

Likewise, Alamoudi et al. [28] presented a rule-based information extraction approach using regular expressions to extract metadata from PDF books for sake of better searching, indexing and classification. The scheme exhibits promising results in terms of accuracy. Similarly, an approach proposed by Alghamdi et al. [29], further extended the work to extract the table of contents (ToC) from the PDF books for sake of generating the tree index for the chapters. Information extraction is one of the potential applications of digital libraries and web-based systems [30]. Alqahtani et al. [31] used the automated approach for resume or curriculum vitae parsing for sake of automated human resource (HR) center especially for sake of electronic recruitment (E-recruitment). The proposed was capable of selecting the candidates with fully aligned requirements as well as with partially aligned requirements. That was the case where the decision was taken for most than one potential candidate to be hired. The proposed approached was presented as a potential recommender system or decision support system (DSS) in the HR hiring process to induct the most suitable employee for the said job posting. Information extraction has been used an important tool for not only extracting the useful information from the diverse source documents but also sets the basis for several advanced text and data mining systems for the document classification. It is pretty obvious that finer the information extraction better will be the document classification. For instance, Dash et al. [32] used information extraction as tool to extract the useful information from the user logs separated as network, machine and web usage duly collected from a deployed proxy server in the organizational network. By investigating a fuzzy rule-based system [33, 34] together with a Gaussian radial basis function neural network (GRBF-NN) [35, 36] to classify the network user as safe or unsafe depending upon his web, machine, and network activities over the time. Further, the individual 360 feedback was contrasted to the classifier outcome to overall present his/her credibility over the year. The scheme was potentially useful for the administration to take suitable action in case the user exhibits a suspicious behavior.

According to these studies discussed above, various methods to extract information have been proposed. However, none of them use regular expressions in Python language. Hence, we applied regular expressions in Python language to extract the main information from resumes. Furthermore, we want to compare our method in terms of information that has been extracted with the information that was extracted from the previous study [14] to improve the performance of the proposed method.

3. METHODOLOGY

In this research, we aim to extract the main information for the authors of resumes. The information has been extracted from the resume includes: name, date of birth, email, GPA, gender, phone number, nationality, address. The input of our system is a collection of resumes in PDF Format while the output is the information that is extracted. The system approach goes through different phases such as pre-processing, converting the resumes files into PDF and information extraction by the rule-based method to extract the elements. Figure 1 shows the phases of the proposed approach.

3.1 Pre-processing

In the pre-processing, we carried out some processes to unified the format and helped for extracting information easily. For example, annotations and labels to help identify the system about the said token whether it is a name or salutation.

Moreover, where the said text token resides in the target document.

This phase contained many processes including:

- Added a label to some information that did not have a label in a resume such as some resume containing the name without the label (name).
- Unified all labels to the name of the label then followed by (:) such as (Name:).
- Removed the icons that represent specific labels such as phones (☎).
- Changed and unified the format of the date of birth to (day/month/year).
- Unified the Grade point Average (GPA) shortcut from CGPA to GPA for each resume.
- Changed all resumes have an (E-mail) label to (Email).
- Some resumes have the mobile number and others have a phone number, thus unified the label to phone number for all resumes.
- Edited all the spaces in resumes.
- Any resume has a table, we converted the table to text then applied all processes that are mentioned in previous steps.

After applying all the processes in resumes, all resumes have to be converted to PDF Format. Figure 2 shows the final format of the resumes.

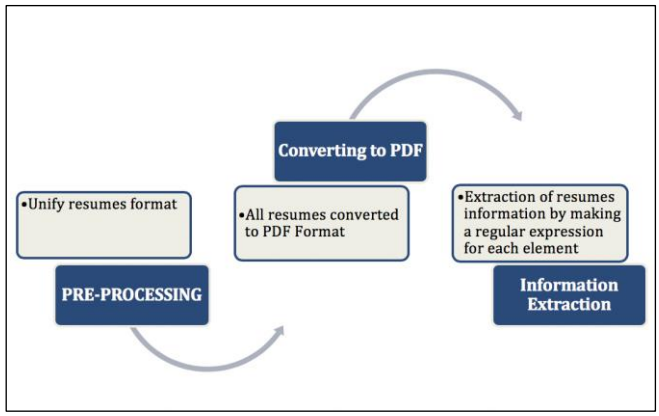
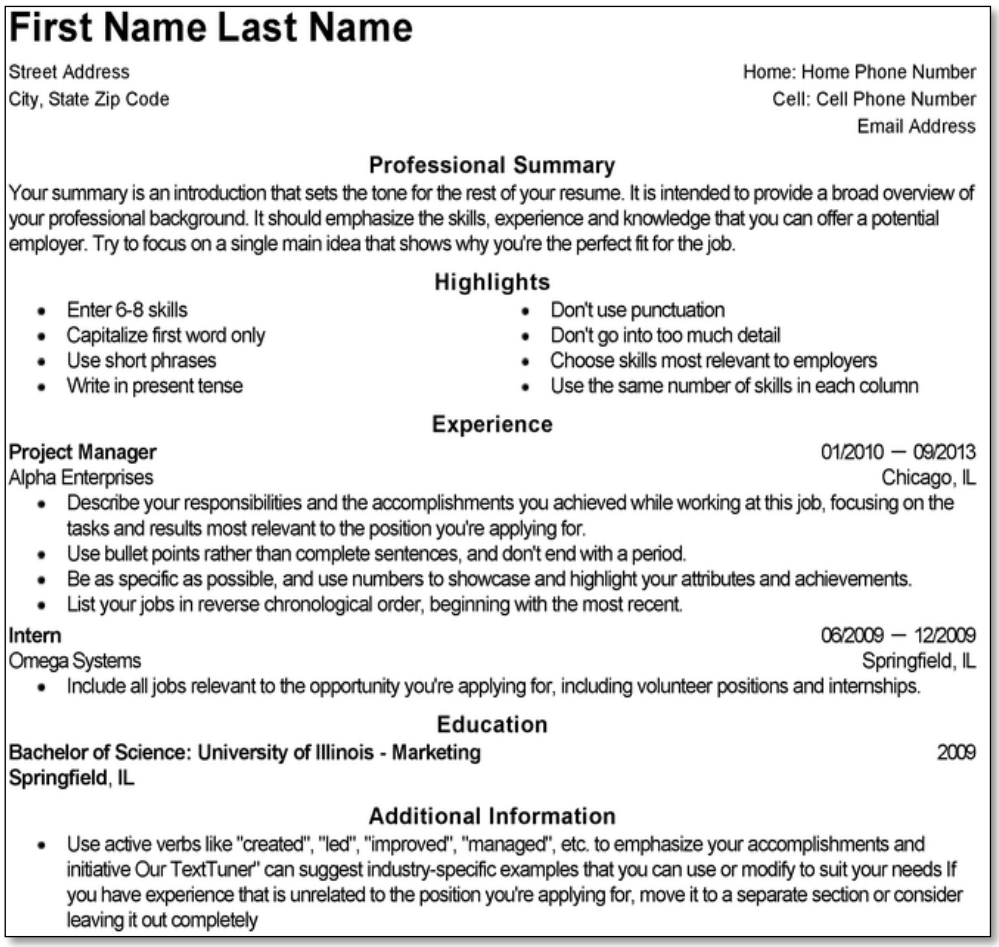


Figure 1. The phases of the system approach



First Name Last Name

Street Address
City, State Zip Code

Home: Home Phone Number
Cell: Cell Phone Number
Email Address

Professional Summary

Your summary is an introduction that sets the tone for the rest of your resume. It is intended to provide a broad overview of your professional background. It should emphasize the skills, experience and knowledge that you can offer a potential employer. Try to focus on a single main idea that shows why you're the perfect fit for the job.

Highlights

- Enter 6-8 skills
- Capitalize first word only
- Use short phrases
- Write in present tense
- Don't use punctuation
- Don't go into too much detail
- Choose skills most relevant to employers
- Use the same number of skills in each column

Experience

Project Manager 01/2010 — 09/2013
Alpha Enterprises Chicago, IL

- Describe your responsibilities and the accomplishments you achieved while working at this job, focusing on the tasks and results most relevant to the position you're applying for.
- Use bullet points rather than complete sentences, and don't end with a period.
- Be as specific as possible, and use numbers to showcase and highlight your attributes and achievements.
- List your jobs in reverse chronological order, beginning with the most recent.

Intern 06/2009 — 12/2009
Omega Systems Springfield, IL

- Include all jobs relevant to the opportunity you're applying for, including volunteer positions and internships.

Education

Bachelor of Science: University of Illinois - Marketing 2009
Springfield, IL

Additional Information

- Use active verbs like "created", "led", "improved", "managed", etc. to emphasize your accomplishments and initiative Our TextTuner" can suggest industry-specific examples that you can use or modify to suit your needs If you have experience that is unrelated to the position you're applying for, move it to a separate section or consider leaving it out completely

Figure 2. The general resumes format/template

3.2 Proposed rule based approach for IE

In the proposed approach, the extraction of resumes information includes several types such as name, date of birth, email, phone number, GPA, gender, nationality, and address. Figure 3 shows the extracted information from the resume. After standardizing the form of resumes a regular expression is written or a rule is carved to match the location of the information in the document and obtain the content of information. The regular expression is a string of text used in programming languages that helps to search within a text for pattern matching to make a comparison and then extract it or verify it [17].

These elements are identified by regular expression identifiers to extract the eight elements from the content of resumes. In the following sections we discuss each element:

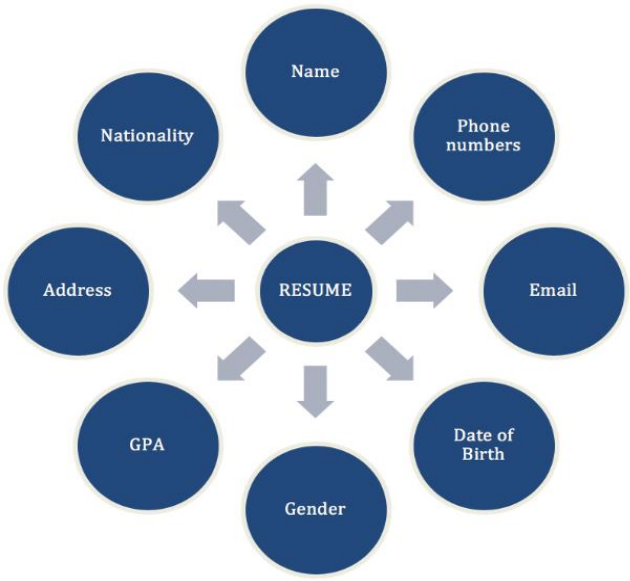


Figure 3. The extracted information from the resumes

Name:

Names in a resume may consist of two to four words. Therefore, we extracted the first two components of the name from all resumes. The following regular expression of the name takes the first name and second name of the resume author:

"name:\s*(.*?)(?=\s*name:|:)"

Example token:
Name: John Smith
Output: John Smith

Phone numbers:

Phone numbers in resumes can be in different forms, depending on the town in which each person resides. Hence, the regular expression is used to extract the phone numbers from resumes with all symbols that may be contained in phone numbers such as (+ -). While the space between the numbers is not accepted.

"phone number:\s*(.*?)(?=\s*phone number:|)"

Example token:
Phone number: 009665XXXXXXX
Output: 009665XXXXXXX

Email:

The email is split into three parts: The first part of an email address is the username then the symbol "@ " is the second part.

The last part is the domain divided into two: the mail server and the top-level domain. However, to identify email in resumes can apply the following carved regular expression:

'[\w\.-]+@[\w\.-]+'

Example token:
alpha@bravo.charlie
Output: alpha@bravo.charlie

Date of birth:

The date of birth extracted from the resumes in this research is in the form of (day, month, year). Day and the month may contain two numbers/digits and the year from four numbers/digits. It can also be expression as dd/mm/yyyy date format. The resultant regular expression for extracting the date of birth from the resume is:

'(?:[0-9]{2}/){2}[0-9]{4}'

Example token:
01/10/1982
Output: 01/10/1982

Gender:

The gender for the author's resumes is taken out by applying the regular expression:

"gender:\s*(.*?)(?=\s*gender:|)"

To control the output to be as male or female we used if condition after applying the regular expression of gender.

Example token:
Gender: Male
Output: Male

GPA:

Grade Point Average (GPA) consists of (x.x) format, the number before dot contains one number only and the numbers after dot contain two to three numbers. Also, the regular expression accepts (x.x/x) format.

In the current study, the GPA that is an extraction from resumes is a bachelor's degree by using the regular expression:

"gpa:\s*(.*?)(?=\s*gpa:|)"

It is noteworthy, the regular expression accepts both out of 4.0 and out of 5.0 GPA.

Example token:
GPA: 4.56/5
Output: 4.56/5

Address:

In this study, the address in the resumes come in the form of (Street/Road name city name, country name). It is written to identify the location of the author. The address's regular expression ends the process of extracting when there is a comma or double space. The country in the address did not extract because it appears after the comma.

"address:\s*(.*?)(?=\s*address:|)"

Example token:
Address: 23 Beverly hills Murree.
Output: 23 Beverly hills Murree

Nationality:

Nationality in resumes can be from any country. To retrieve the nationality can apply the following regular expression:

"nationality:\s*(.*?)(?=\s*nationality:|)"

Example token:
Nationality: American.

Output: American

These experimental analyses showed that each format requires a pattern for the required information.

Table 1 enlists all the regular expressions used for the extraction of the said information from various sections of the resume document.

Table 1. Regular expressions for each element in the resume

Element	Regular Expression (Pattern)
Name	"name:\s*(.*?)(?=\s*name:|:)"
Phone numbers	"phone number:\s*(.*?)(?=\s*phone number:|)"
Email	'[\w\.-]+@[\w\.-]+'
Date of Birth	'(?:[0-9]{2}/){2}[0-9]{4}'
Gender	"gender:\s*(.*?)(?=\s*gender:|)"
GPA	"gpa:\s*(.*?)(?=\s*gpa:|)"
Address	"address:\s*(.*?)(?=\s*address:|)"
Nationality	"nationality:\s*(.*?)(?=\s*nationality:|)"

4. IMPLEMENTATION

In this section, we discuss the dataset description, the carved and customized regular expressions pertaining to the fields of the resume. Moreover, the results of the experiment in detail and the further discussions to make a comparison between the proposed study and the earlier study [14].

4.1 Description of dataset

In In this research, the dataset used in the experiment is a private resume from multiple sources. The dataset contains 75 resumes; 50 resumes are for the training sets and the other 25 are for testing sets. The resumes came in PDF format, text, and Word Document format.

4.2 Extraction algorithm

The proposed information extraction algorithm is given in Algorithm 1. It starts with taking a resume in the PDF form as an input and produce the set of extracted information tokens as the output.

Algorithm 1: Extraction algorithm

Input: A Resume File
Output: Extracted Information

```
While (end of file)
        If (String matches RE for Name)
                Extract Name
        If (String matches RE for Cell)
                Extract Cellphone number.
        If (String matches RE for Email)
                Extract Email ID
        If (String matches RE for DoB)
                Extract DoB
        If (String matches RE for Sex)
                Extract Sex
        If (String matches RE for CGPA)
                Extract CGPA
        If (String matches RE for Add)
                Extract Address
        If (String matches RE for Nat)
                Extract Nationality
End While
```

4.3 Results and discussion

The experiment in this research was applied using the Python programming language. The proposed approach was investigated on the private resumes dataset. The regular expressions are used in the Python code to extract information from resumes. In the experiment, we take 75 resumes in a 2:1 ratio. The experiment contained 50 resumes for a training set and 25 for testing sets. We evaluate the effectiveness and the efficiency of our system by using a precision indicator. In order to evaluate the experiment to achieve a better possible precision, the regular expressions are made for every element after re-formatted the resumes.

The regular expressions mentioned in detail in the previous section. The name elements during the extraction process from the training and testing dataset obtained 100% precision. Also, the rest of the elements achieved the same results of the name precision except the address element. The address precision has achieved 77.02% in the training set while the testing set achieved 71.60%. Table 2 shows the comparison of the performance of the training set and testing set for elements extraction from the resumes.

To calculate the precision for the element in each resume, following steps were applied as given in Eq. (1) and Eq. (2):

$$\text{Accuracy for each element} = [(\text{total of character in output of the program}) / (\text{total of character in resume})] * 100 \quad (1)$$

$$\text{Precision} = [(\text{sum of the accuracies for each element}) / (\text{total number of the resumes})] * 100 \quad (2)$$

After computing the precision for the element, the overall average of precision calculated for all resumes in an experiment is by Eq. (3):

$$\text{Total precision} = [(\text{total of precision all element}) / (\text{number of the element})] * 100 \quad (3)$$

Table 2. Experimental results

Element	Precision	
	Training	Testing
Name	100%	100%
Email	100%	100%
Phone Number	100%	100%
Date of birth	100%	100%
GPA	100%	100%
Address	77.02%	71.60%
Nationality	100%	100%
Gender	100%	100%
Average	97.126%	96.451%

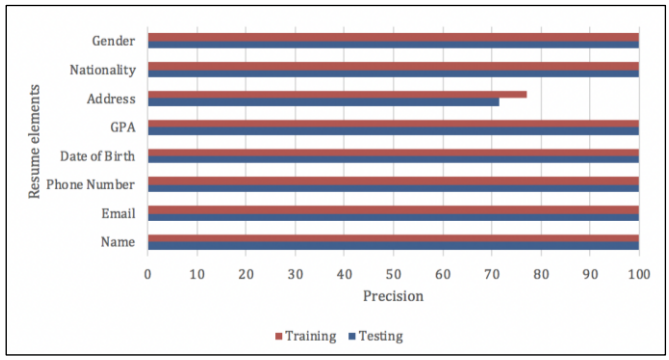


Figure 4. The precision of each element in resumes

Figure 4 presents the precision against each extracted field in the resume in both training and testing phases. While Figure 5 presents the average precisions from all the field in both training and testing phases. Except the address field, that exhibited 77.02% and 71.60% precision for the training and testing phases, respectively; all the other fields exhibited 100% precision in both training and testing phases.

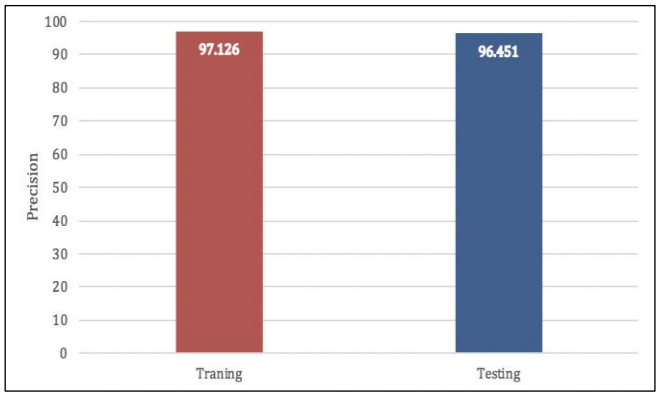


Figure 5. The average of precision for all resumes

Further comparison between the proposed study and the previous study in information extraction is provided in Table 3. The proposed technique exhibits better accuracy in the relevant fields of meta data and consequently a better average accuracy.

Same is expressed in Figure 6 where the existing studies' precision is contrasted to the proposed study field by field.

Table 3. Comparison with state of the art studies

Field	Element	Earlier Study [14]	Proposed Method
1	Name	85%	100%
2	Email	75%	100%
3	Phone num.	85%	100%
4	Date of birth	98%	100%
5	GPA	X	100%
6	Address	80%	71%
7	Nationality	100%	100%
8	Gender	100%	100%
9	Father name	79%	X
10	Mother name	68%	X
11	Passport	67%	X
12	Location	56%	X
13	Marital status	100%	X

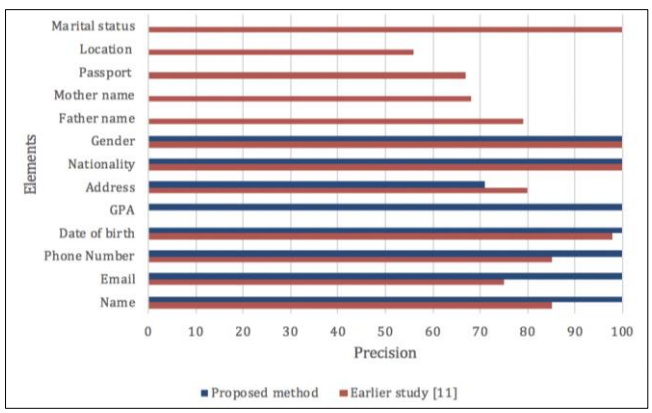


Figure 6. The precision of the proposed method & the earlier study for each element that extracted from resumes

Table 4 presents the comparison between the proposed study and state-of-the-art study in the literature [14], in terms of average precision for the extracted information from all the provided resumes in the dataset. Proposed scheme outperforms with 3.4% difference.

Table 4. Comparison for average precision

Metric	Earlier Study [14]	Proposed Method
Precision	93%	96.4%

In light of the findings in the above tables, the elements that have been extracted by the earlier study are more than the ones that have been extracted in our proposed method. The reason behind that is some of the elements extracted from the earlier study are uncommon in all resumes, and also it is not necessary, such as (Father name, Mother name, Passport and Marital status). The name element in the proposed work included the first and second name, while the earlier work has three kinds of the names separately which are (author name, author's father name and author's mother name). Furthermore, GPA is one of the important elements in resumes. It is extracted from resumes in the proposed work but the earlier work does not extract such information. The earlier study has extracted the passport element from resumes but due to the sensitivity of the passport information, it may not be important to extract this item from the resumes. The final results reveal that the precision of the proposed method outperformed the earlier works. The scheme can be used as tool to automatically extract the information from the resume to help classify them with respect to various criteria, that will make the job easier and less vulnerable to errors. Moreover, the scheme has an ability to be scaled to the next level in order to incorporate other fields such as experience type and number of years etc. As far as the limitations of the study are concerned, the scheme is vulnerable to extensive changes in the resume format. Moreover, the current approach addresses the information extraction problem only in English language.

4.4 Document classification

This section presents the document/resume classification based on the extracted information. The classification is based on address, gender and nationality that is automatically extracted already. The performance measures can be implemented using a confusion matrix, that consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Furthermore, the loss rate stands for the total errors or the variance between the predicted and the actual values. Using a variety of metrics will significantly help in defining the performance of the proposed study [37-40]. The Eqs. (4)-(7) of the previously mentioned measures are seen below.

$$\text{Accuracy (A)} = (TP+TN)/(TP+TN+FP+FN) \quad (4)$$

$$\text{Precision (P)} = TP/(TP+FP) \quad (5)$$

$$\text{Recall (R)} = TP/(TP+FN) \quad (6)$$

$$\text{F1}-\text{Score (F1)} = (2*P*R)/(P+R) \quad (7)$$

Table 5 presents the accuracy, precision, recall and F1-score

of the proposed scheme. The proposed scheme possesses classification accuracy, precision, recall and F1-score as 98.02%, 98.01%, 98% and 98%, respectively.

Table 5. Classification results

Metric	Score
Accuracy	98.02%
Precision	98.01%
Recall	98.0%
F-Score	98.0%

5. CONCLUSIONS

In this research, we presented regular expressions to extract useful information from resumes. Regular expressions have been applied by using the Python programming language. Also, the extracted information from resumes includes eight elements such as name, date of birth, email, phone number, GPA, gender, nationality, and address. The experiment showed that the testing set has 96.4% in term of precision. Moreover, we compared the proposed work with earlier study. Future research should further be applied to extract other elements of the resumes such as skills and experience information. Applying other regular expressions to further improve the outcomes of the experiments in the address element and to achieve more accurate results. Also, the ability to apply information extraction from resumes in various formats to increase the flexibility to accept all possible formats for resumes. Consequently, the resume is classified based on the extracted information mainly gender, address and nationality. The scheme exhibits a classification accuracy of 98.02%. It is worth mentioning that the accuracy solely based on the information extraction quality. In future, deep learning, transfer learning and other state of the art approaches can be used in the hybrid manner to increase the performance to the next level with more fault tolerance and with diverse resume format with additional fields and better classification.

REFERENCES

[1] What Is a Resume? A Brief Overview | Pongo. https://www.pongoresume.com/articles/391/what-is-a-resume-a-brief-overview.cfm, accessed on 21 Apr. 2020.

[2] Sajid, N.A., Rahman, A., Ahmad, M., Musleh, D., Basheer Ahmed, M.I., Alassaf, R., Chabani, S., Ahmed, M.S., Salam, A.A., AlKhulaifi, D. (2023). Single vs. multi-label: The issues, challenges and insights of contemporary classification schemes. Applied Sciences, 13(11): 6804. https://doi.org/10.3390/app13116804

[3] Rahman, A.U., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., Khan, M.A., Mosavi, A. (2022). Rainfall prediction system using machine learning fusion for smart cities. Sensors, 22(9): 3504. https://doi.org/10.3390/s22093504

[4] Talha, M., Sarfraz, M., Rahman, A., Ghauri, S.A., Mohammad, R. M., Krishnasamy, G., Alkharraa, M. (2023). Voting-based deep convolutional neural networks (VB-DCNNs) for M-QAM and M-PSK signals classification. Electronics, 12(8): 1913. https://doi.org/10.3390/electronics12081913

[5] Dong, X.L., Yu, X.H., Li, J., Sun, Y.Z. (editors). (2015). Web-Age Information Management. In 16th International Conference, WAIM 2015 Qingdao, China, pp. 540-543, Springer.

[6] Jayaraj, V., Mahalakshmi, V., Rajadurai, P. (2015). Resume information extraction using feature extraction model. American International Journal of Research in Sciences, Technology, Engineering and Mathematics, 201-206.

[7] Celik, D. (2016). Towards a semantic-based information extraction system for matching résumés to job openings. Turkish Journal of Electrical Engineering and Computer Sciences, 24(1): 141-159. https://doi.org/10.3906/elk-1304-130

[8] Ahmad, R., Afzal, M.T., Qadir, M.A. (2016). Information extraction from PDF sources based on rule-based system using integrated formats. In: Sack, H., Dietze, S., Tordai, A., Lange, C. (eds) Semantic Web Challenges. SemWebEval 2016. Communications in Computer and Information Science, vol. 641. Springer, Cham. https://doi.org/10.1007/978-3-319-46565-4_23

[9] Mhapasekar, D.P. (2017). Ontology based information extraction from resume. In 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, India, pp. 43-47. https://doi.org/10.1109/ICOEI.2017.8300962

[10] Yan, W., Qiao, Y. (2017). Chinese resume information extraction based on semi-structured text. In 2017 36th Chinese Control Conference (CCC), Dalian, China, pp. 11177-11182. https://doi.org/10.23919/ChiCC.2017.8029141

[11] Ghufran, M., Bennacer, N., Quercini, G. (2017). Wikipedia-based extraction of key information from resumes. In 2017 11th International Conference on Research Challenges in Information Science (RCIS), Brighton, UK, pp. 135-145. https://doi.org/10.1109/RCIS.2017.7956530

[12] Jayaram, K., Sangeeta, K. (2017). A review: Information extraction techniques from research papers. In 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, pp. 56-59. https://doi.org/10.1109/ICIMIA.2017.7975532

[13] Chen, J., Zhang, C., Niu, Z. (2018). A two-step resume information extraction algorithm. Mathematical Problems in Engineering, 2018: 5761287. https://doi.org/10.1155/2018/5761287

[14] Ayishathahira, C.H., Sreejith, C., Raseek, C. (2018). Combination of neural networks and conditional random fields for efficient resume parsing. In 2018 International CET Conference on Control, Communication, and Computing (IC4), Thiruvananthapuram, India, pp. 388-393. https://doi.org/10.1109/CETIC4.2018.8530883

[15] Sajid, N.A., Ahmad, M., Afzal, M.T., Atta-ur-Rahman. (2021). Exploiting papers' reference's section for multi-label computer science research papers' classification. Journal of Information & Knowledge Management, 20(1): 2150004. https://doi.org/10.1142/S0219649221500040

[16] Sajid, N.A., Ahmad, M., Rahman, A.U., Zaman, G., Ahmed, M.S., Ibrahim, N., Ahmed, M.I.B., Krishnasamy, G., Alzaher, R., Alkharraa, M., AlKhulaifi, D., AlQahtani, M., Salam, A.A., Saraireh, L., Gollapalli, M., Ahmed, R. (2023). A novel metadata based multi-label document classification technique. Computer Systems Science & Engineering, 46(2): 2195-2214. https://doi.org/10.32604/csse.2023.033844

[17] Zaman, G., Mahdin, H., Hussain, K., Abawajy, J., Mostafa, S.A. (2021). An ontological framework for information extraction from diverse scientific sources. IEEE Access, 9: 42111-42124. https://doi.org/10.1109/ACCESS.2021.3063181
[18] Regular Expressions. https://alhazmy13.net/regex/, accessed on 21 Apr. 2020.
[19] Vukadin, D., Kurdija, A.S., Delač, G., Šilić, M. (2021). Information extraction from free-form CV documents in multiple languages. IEEE Access, 9: 84559-84575. https://doi.org/10.1109/ACCESS.2021.3087913
[20] Gaur, B., Saluja, G.S., Sivakumar, H.B., Singh, S. (2021). Semi-supervised deep learning based named entity recognition model to parse education section of resumes. Neural Computing and Applications, 33: 5705-5718. https://doi.org/10.1007/s00521-020-05351-2
[21] Wosiak, A. (2021). Automated extraction of information from Polish resume documents in the IT recruitment process. Procedia Computer Science, 192: 2432-2439. https://doi.org/10.1016/j.procs.2021.09.012
[22] Barducci, A., Iannaccone, S., La Gatta, V., Moscato, V., Sperlì, G., Zavota, S. (2022). An end-to-end framework for information extraction from Italian resumes. Expert Systems with Applications, 210: 118487. https://doi.org/10.1016/j.eswa.2022.118487
[23] Bhoir, N., Jakate, M., Lavangare, S., Das, A., Kolhe, S. (2023). Resume Parser using hybrid approach to enhance the efficiency of Automated Recruitment Processes. Authorea Preprints. https://doi.org/10.22541/au.168170278.82268853/v1
[24] Musleh, D., Othman, R., Alhaidari, F. (2019). A novel approach to Arabic keyphrase extraction. ICIC Express Letters, Part B: Applications, 10(10): 875-884.
[25] Batool, A., Amnah, A., Meera, A., Safa, A. (2022). A novel approach to Wikipedia references classification. ICIC Express Letters: Part B, 13(12): 1321-1330.
[26] Rahman, A., Musleh, D., Nabil, M., Alubaidan, H., Gollapalli, M., Krishnasamy, G., Almoqbil, D., Khan, M.A.A., Farooqui, M., Ahmed, M.I.B., Ahmed, M.S., Mahmud, M. (2022). Assessment of information extraction techniques, models and systems. Mathematical Modelling of Engineering Problems, 9(3): 683-696. https://doi.org/10.18280/mmep.090315
[27] Zaman, G., Mahdin, H., Hussain, K., Rahman, A. (2020). Information extraction from semi and unstructured data sources: A systematic literature review. ICIC Express Letters, 14(6): 593-603. https://doi.org/10.24507/icicel.14.06.593
[28] Alamoudi, A., Alomari, A., Alwarthan, S. (2021). A rule-based information extraction approach for extracting metadata from PDF books. ICIC Express Letters, Part B: Applications, 12(2): 121-132. https://doi.org/10.24507/icicelb.12.02.121
[29] Alghamdi, H., Dawwas, W., Almutairi, T.H., Rahman, A. (2022). Extracting ToC and metadata from PDF books: A rule-based approach. ICIC Express Letters, 13(2): 133-143. https://doi.org/10.24507/icicelb.13.02.133
[30] Alhaidari, F.A. (2019). The digital library and the archiving system for educational institutes. Pakistan Journal of Information Management and Libraries, 20: 94-117.
[31] Alqahtani, A., Alhaidari, F.A., Mahmud, M., Sultan, K. (2019). Decision support system assisted e-recruiting system. Journal of Computational and Theoretical Nanoscience, 16(2): 335-340. https://doi.org/10.1166/jctn.2019.7955
[32] Dash, S., Luhach, A.K., Chilamkurti, N., Baek, S., Nam, Y. (2019). A Neuro-fuzzy approach for user behaviour classification and prediction. Journal of Cloud Computing, 8(1): 1-15. https://doi.org/10.1186/s13677-019-0144-9
[33] Qureshi, I.M., Malik, A.N., Naseem, M.T. (2016). QoS and rate enhancement in DVB-S2 using fuzzy rule based system. Journal of Intelligent & Fuzzy Systems, 30(2): 801-810. https://doi.org/10.3233/IFS-151802
[34] Rahman, A., Qureshi, I., Malik, A., Naseem, M. (2014). A real time adaptive resource allocation scheme for OFDM systems using GRBF-neural networks and fuzzy rule base system. International Arab Journal of Information Technology (IAJIT), 11(6): 590-598.
[35] Rahman, A. (2023). GRBF-NN based ambient aware realtime adaptive communication in DVB-S2. Journal of Ambient Intelligence and Humanized Computing, 14(5): 5929-5939. https://doi.org/10.1007/s12652-020-02174-w
[36] Rahman, A.U., Dash, S., Luhach, A.K. (2021). Dynamic MODCOD and power allocation in DVB-S2: A hybrid intelligent approach. Telecommunication Systems, 76(1): 49-61. https://doi.org/10.1007/s11235-020-00700-x
[37] Basheer Ahmed, M.I., Zaghdoud, R., Ahmed, M.S., Sendi, R., Alsharif, S., Alabdulkarim, J., Albin Saad, B.A., Alsabt, R., Rahman, A., Krishnasamy, G. (2023). A real-time computer vision based approach to detection and classification of traffic incidents. Big Data and Cognitive Computing, 7(1): 22. https://doi.org/10.3390/bdcc7010022
[38] Gollapalli, M., Musleh, D., Ibrahim, N., Khan, M.A., Abbas, S., Atta, A., Khan, M.A., Farooqui, M., Iqbal, T., Ahmed, M.S., Ahmed, M.I.B., Almoqbil, D., Nabeel, M., Omer, A. (2022). A neuro-fuzzy approach to road traffic congestion prediction. Computers, Materials & Continua, 73(1): 295-310. https://doi.org/10.32604/cmc.2022.027925
[39] Ahmed, M.I.B., Alotaibi, S., Dash, S., Nabil, M., AlTurki, A.O. (2022). A review on machine learning approaches in identification of pediatric epilepsy. SN Computer Science, 3(6): 437. https://doi.org/10.1007/s42979-022-01358-9
[40] Ibrahim, N.M., Gabr, D.G., Rahman, A., Musleh, D., AlKhulaifi, D., AlKharraa, M. (2023). Transfer learning approach to seed taxonomy: A wild plant case study. Big Data and Cognitive Computing, 7(3): 128. https://doi.org/10.3390/bdcc7030128