Automatic Software Engineering Position Resume Screening using Natural Language Processing, Word Matching, Character Positioning, and Regex

Dipendra Pant
Department of Information Technology
Kathford International College of Engineering and Management
Lalitpur, Nepal
Email: dipendrapant778@gmail.com

Dhiraj Pokhrel
Department of Industrial Engineering
University of Arkansas,
Arkansas, Fayetteville, AR 72701, USA
Email: dhirajpokhrel4455@gmail.com

Prakash Poudyal
Department of Computer Science and Engineering
Kathmandu University
Banepa, Nepal
Email: prakash@ku.edu.np

2022 5th International Conference on Advanced Systems and Emergent Technologies (IC_ASET) | 978-1-6654-2735-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/IC_ASET53395.2022.9765916

***Abstract*— Screening candidates' resumes manually is a tedious job, with possibilities of sometimes missing good candidates due to human errors, nepotism, and bias. However, these kinds of mismanagement don't apply to machines. Instead, automatic screening of candidates reduces a lot of effort, time, and cost. Hence this work specifically focuses on extracting technical skills using natural language processing specifically resume label character positioning, data set consisting of software engineering candidate requirements, regular expressions, and word and phrase matching for candidate information retrieval. Character positioning a new technique for information extraction is introduced, which perceives needed data and pulls it out. This methodology creates a summary of the resume from the extracted information. And computes count scores from based recognized skills, and education plus experience level. Finally, upon testing on five random software engineering positions resume correct extraction rate of 33.59% was obtained.**

Keywords*— *Automated Resume Screening, Information Retrieval Using Characters Positioning, Natural Language Processing

I. INTRODUCTION

Resume screening is the process of having a detailed study of candidate resumes to determine the best candidate for the job or a post and disqualify the unfit one. Especially for information technology and software engineering positions, a large number of applicants apply from around the globe. Due to this high effort for search is needed to select an appropriate candidate. According to [1] software engineering means designing, developing quality software and its maintenance utilizing the capabilities of computers. Therefore, anyone that utilizes computer capabilities to make useful software to humans is a software engineer.

According to [2] application software developers crown the rank of the most demanded job. Similarly, based on the [4] artificial intelligence, data science, cloud computing, developers, DevOps are high demanded jobs with an annual growth rate of more than 30 % in cloud computing and more than 40 % in artificial intelligence and data processing related works. To define the software engineering position here it means person working as a software developer, technical architects, chief technology officers, data scientists and artificial intelligence professionals, cloud computing, web or mobile or desktop application developer. The professional developers around the world are 23.9 million [3] and projected to reach 28.7 million by 2024 and among the ten most profitable companies, 3 were the information technology companies as per [4]. They provide a significant number of software engineering jobs and are the most profit-earning companies too. As in almost all tech discussions these days it is said that every company will be a data company and this is what Google also strongly believes. Additionally, artificial intelligence will be dominating every sector in the world which clearly means a rise in software engineering jobs. According to a popular job search company Glassdoor [26] the average time taken for a single manual hire takes twenty-three days. And according to LinkedIn, only 30% of companies fill vacant positions in thirty days whereas 70% of companies take between 1- 4 months. Another amazing fact is according to Ideal [5] manual resume screening takes up to two to three hours for just one hire and 88% of received resumes are of unqualified candidate. So manual recruitment takes a lot of effort, time, and budget and has a high probability to miss potential candidates. Also in many cases, manual screening might not be able to identify the best skills and specialization of the candidate. Hence, there is a great need for automated screening of resumes for selecting the best candidate. So, to mitigate it and make the process fast and automate the process. There has been the concept of using artificial intelligence in recruitment, which means making machines intelligent to hire the best candidates from applicants. This might lower cost, time, and effort to a large extent. Several companies like Google, Microsoft use proprietary software for hiring. According to 98.2% of fortune, 500 companies use the applicant tracking system (ATS). ATS is a software program that collects and stores resumes and handles the recruitment and hiring process.

So, in this paper, we have evaluated the possible ways for automating the resume screening process through technological interventions, to extract useful insights from the text file, especially for software engineering positions. This paper uses the natural language processing approach of word matching and regular expressions, character positioning, a skill data set consisting of skills, education, and experience-related text data. Here label character positioning means to keep track of a character and a specific word position in a document.

A. Research Questions

The following research question was used to guide this research work:

1- What can be an architecture to extract technical skills from software engineering position resumes?

978-1-6654-2735-7/22/$31.00 ©2022 IEEE

2- How text mining can be implemented for relevant information extraction to summarize the resume?

II. Literature Review

Time-consuming drawbacks, nepotism, and human bias in manual screening led to the automatic screening of resumes. Different approaches have been used for resume parsing keywords retrieval and keyword matching. Some are ontology mapping tree, resume summarization, information extraction followed by resume classification are some popular ones. Natural language processing has been there since the 50s and new improvements are going on but the evolution of text mining eased and simplified the concept of extracting knowledge from resumes. Maria Heart suggested discovering new facts and knowledge from any unstructured and semi-structured texts using the term text data mining for the first time. As "Text has rich information, but it is encoded in a form that is difficult to decode [6]. Applicant Tracking System (ATS) software that performs automatic electronic recruitment handling [7] has been used since ignited by various recruitment agencies. ATS has improved itself largely by embedding the state of art machine learning approaches and artificial intelligence in segregating suitable candidates for the job. But it mainly focuses on automating the tracking and informing the job applicant and human resource officers about recruiting. Regarding job specification parsing and resume parsing [8] used incremental learning using object-oriented for automatic job classification of job openings but extracting meaningful phrase-based features from the job title and description was not achieved. Resume information extraction followed by resume classification is another effective approach mainly suitable for companies that have diverse job postings. Resume classification supports in determining for which specific job category the resume has been sent. Many works have been done for resume category classification, [21- 22] used the convolution neural networks and the concept of word embedding on a small amount of labeled data. Natural language processing and machine learning techniques were used by [23] to create a resume classification system. Upon experimenting with nine different algorithms, the support vector machine gave the best result for classification. The data-driven approach based on natural language processing for filtering the resume was used by [24], where they ranked the resume based on the scores calculated.

The paper [10] used the ontology mapping technique for finding similarities in different concepts, to screen candidates based on resumes with a tool named EXPERT. It creates ontology of job seekers and job postings, then matches both ontologies to select the best candidate further screens candidates semantically with the score value of the job requirement. In keyword matching approach [10] proposed a hybrid approach consisting of resume classification. Followed by keywords matching through natural language processing techniques and regular expressions for respective job categories. Again, based on the obtained information about resume ranking of resumes to find the best match for the job in a specific job category was done. Continuing the path of text mining approach [11] developed a prototype that assists in the screening of candidates from submitted unstructured resumes in English, thus identifying software development technical knowledge using natural language processing and text mining. It used the ontology model for knowledge profiling which includes (i) Profiling (ii) Evaluation (iii) Valuation and finally compares the knowledge profile with job specification requirements. Amazon made a huge investment for the Amazon AI-enabled resume parsing tool [12] that was being built but was shut down after the algorithm used showed biases against female candidates. Huge demand for automatic intelligent recruiting tools has embraced continued research in the field. Various research has been done in the field [9-12,14-15,19-20], among resume screening for technical jobs like software engineering extracting the skill levels and expertise from the resume is a must. Two-step resume, the information extraction algorithm [13] for Chinese language resume parsing was implemented where the first raw text of the resume is identified as resume blocks through word index and punctuation index, word lexical attribute. Secondly, resume writing style and multiple classifiers are employed to extract information from resumes which consume significantly less annotation time. A detailed systematic review on resume screening by [25] mentions resume screening as an easier task in the case of structured resumes but a challenging task for unstructured ones. Semantic search was used to understand the context of the language but it still possesses many research challenges in making more better. However, all works mentioned above are limited to specific file formats and require huge human efforts, so this paper focuses on exploring a new mechanism of character positioning and integrating it with natural language processing and text mining techniques to extract the skill level, personal details, expertise of candidates from a resume for software engineering position jobs.

III. Architecture

The overall architecture used for automatic screening of software engineering resumes is shown in figure 2. The architecture uses the keyword matching technique. Initially, the job seeker submits the resume, then the obtained resume is converted into a text file using a Tika Parser 1.25 which converts its contents into a string type. Then the skills are extracted from the text file after preprocessing the unwanted text from the string using the predefined dictionary of the skill data set. In the case of education and experience set of keywords associated with education and experience are used. Along with that character positioning technique supports the extraction of skills, education, experience details using a trained character position model.

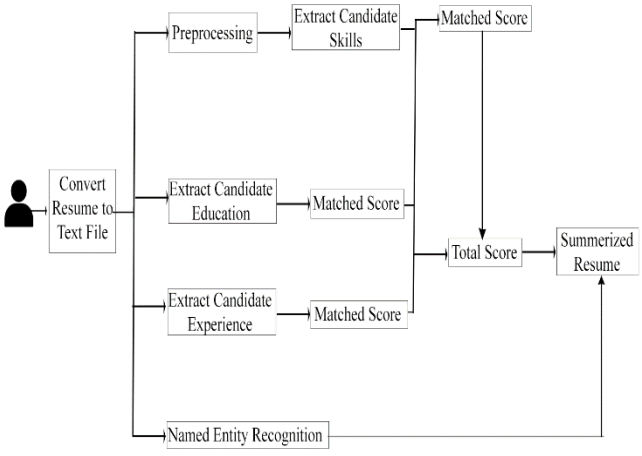

Fig. 1. Architecture for Technical resume Parsing

Simultaneously skills mentioned on the job specification are also extracted. The same iterative process is applied for

experience and education. Then each domain match score is calculated using job specification extracts and resume extracts. Thus the obtained score of each domain is used to compute the overall score of the resume. Based on the calculated total score for a specific job posting a resume list is formed based on the ascending order of the obtained total score. Along with that a summary formed after summarizing the submitted resume is also attached.

A. Dataset

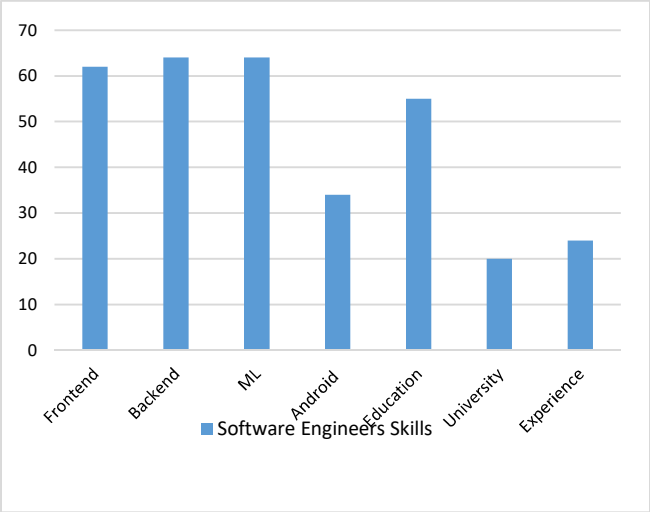

Fig. 2. Architecture for Technical resume Parsing

The first iteration data set were formed by web scraping from the Nepalese online job posting site *Merojob*. The job specification mentioned in the job posting was analyzed to include them in the features. The formed data set consists of skills for software developers, academic degree details in the education, and university name. Dataset consist of keywords to a string. The data set has seven different features that are relevant for information extraction from the resumes. As skills, we took four main skills which consist of Back-end, Front-end, Mobile application development, and machine learning skills which consist of sixty-four, sixty-two, thirty-four, and sixty-three different skill sets respectively. Along with that the education level that determines academic qualification, years of experience, and name of the university were also included.

B. Named Entity Recognition

Named entity was coined by R. Grishman & Sundheim [14] 1996. In [16] named entity recognition is used to extract the names of persons, locations, organizations, and other entities from text documents. The names can be either a single word or multiple words. It simply works by chunking the text into words then algorithms are used to determine if the entity is a name or not. Maximum entropy and hidden Markov model have been used [17] to find the named entity from a chunk of text. Generally, it involves sentence segmentation, tokenization, part of speech tagging, and then entity detection. For western names, a lot of parsers exist and great research is there but extracting Nepali names is a tedious task and research is continuing. For email and phone number extraction, regular expression was used. In the case of Nepali names identified in the resume, there may be multiple named entities like university name, address, reference name. So, the approach used was evaluating the writing style, extracting the first two lines from the resume, and parsing names from those first two lines of the resume.

C. Characters Positioning

Personal details associated with applicants like name, address, phone numbers are usually mentioned at the start of the resume. Similarly experience, skills, and education in the middle. And finally followed by references. This was the general assumption that we used for implementing the character positioning concept. Character position use label name which denotes the type of details. In the case of name, the keyword can be a label name, similarly, in the case of address, the address can be used as a label name. And while defining character positioning the tentative position in resume from where the characters associated with personal details are mentioned. As in figure 3 below, a block having character position details includes start character position, end character position, and a label name for it. In this way, the character positioning technique uses a large collection of information where the possible start and end positions for a specific label are mentioned. This was further used for training a model for information extraction.

D. Algorithm

1) Create a collection of skills for different software engineering positions.
2) Transform the submitted resume in any format into the text file and extract text from the file.
3) Extract each character from the text file in form of the arranged as the word.
4) Pre-process the text by removing unwanted text, symbols, characters, stopwords for technical skill extraction.
5) Calculate the term frequency index of the extracted texts.
6) Compare software engineering position skills in the skillset with the words extracted from the resume text file.
  6.1) Train the model using character position *CSV* file.
  6.2) Use the model to extract personal details and others
  6.3) Provide the extracted skills and score count.
7) Identify the matching words from the skillset list with the words in text file and increase the count score of the matching category.
  7.1) Use extracted skills from characters position model only if they exist in the defined skill setlist.
8) Display the score in each category and the overall skill score and use it to calculate the total score for ranking
9) Repeat the step from (3) to (8) for experience, education, university details.

IV. Results and Discussion

a) Resumes can be written in any writing style based on the job type and their school of knowledge. Extracting skills and other details from a structured resume is the easiest task. But it may restrict the person's creativity in presenting him/herself. Hence in the case of an unstructured resume, it's a difficult task. Regarding a structure's resume, we created a data set of resumes by identifying positions of specific details based on the position of the character, words in the resume and trained the model using the spaCy [17] tool to test on random samples. But it showed very good results for the resume following the resume that used the structure of the training resume but for others, it showed bad results. Hence we moved

towards the keyword and phrase matching teaching to summarize the resumes. The proposed method provides satisfactory results among all other tested methods only, as it extracted the skills and personal details. But it requires a larger amount of keywords and phrases for matching. Extracting experiences from the resume is a challenging task and our method may deviate in extracting the accurate experience of the user.

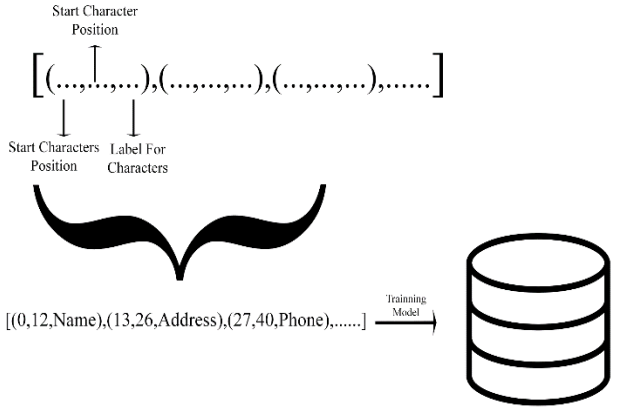

Fig. 3. Label Characters Position-Based Technique

The used methodology successfully extracted skills and personal details, a sample output shown in figure 4.

```
Email Address: ['cwse@fastmail.com']
Phone Number: 555-234-2345
Academic Qualifications: [['Bachelor']]
                    Scores       Skills
Front End                2  [HTML, CSS]
Back End                 2  [Java, ORM]
Machine Learning         2   [R , Java]
Android Developer        2   [GO, Java]
```

Fig. 4. Summary of Extracted Skills and Information from a Resume

We tested our methodology on five different resumes as represented in table 1, in general, the methodology and approach seem promising for automating resume screening. But in the case of other details and skills that are not in the data set but are important for software engineering jobs couldn't be automatically identified. Personal details which consist of the phone number, email, academic qualification were extracted in all of the five resumes, But regarding name and experience details extraction it could only extract the name in only one resume which has followed a certain structure but couldn't extract name in others. Similarly for the experience details extraction out of five resumes it only extracted experience-related details from two resumes only.

TABLE 1: SKILLS EXTRACTED VS ACTUAL SKILLS

ID	FrontEnd		Backend		Mobile		M.L	
	Act	*Obt*	*Act*	*Obt*	*Act*	*Obt*	*Act*	*Obt*
R1	7	2	6	2	5	2	8	2
R2	13	5	8	3	5	2	7	3
R3	14	6	9	4	5	2	10	3
R4	10	4	8	2	6	3	9	2
R5	15	5	9	5	6	2	10	4

where ID represents the resume Id, Act is total number of skills in the resume and Obt is number of skill extracted, Frontend denotes the front-end development skill sets. Similarly, Backend, Mobile and M.L. denotes back-end development, mobile development and machine learning skill sets respectively.

V. CONCLUSION

Automating resume screening is the need of time and state of art technologies are capable of achieving that already implemented by many tech giants and recruiting agencies. Classifying the resume into categories to determine resume that suits the job specification, followed by extracting skills using character positioning technique and natural language processing has the capability to reduce effort and manual hard work. For structured resumes there exist multiple easy methods and one of them can be the label character position-based technique to train a model that gives a very good output. But for the unstructured resumes, the keywords and phrase matching gave a desirable summary that can ease the recruitment process. A random test of five different resumes produced the summarized form of their resume by extracting 33.59% of skills correctly. The methodology used automates the manual process of resume screening for software engineering-related jobs by creating a summary of the resume. Through identifying of personal details using named entity recognition and natural language processing and text mining techniques to extract the skill levels.

A. Future Work

In the future we aim to include the following:

1- Incorporate more software engineering skills and categories and incorporate resume classification followed by resume parsing
2- Seek ways in enhancing the character positioning technique for more accurate information extraction.

ACKNOWLEDGMENT

The author is grateful to the Department of Computer Science and Engineering, Kathmandu University, Nepal for their valuable inputs and guidance.

REFERENCES

[1] B. W. Boehm, “Software engineering economics,” *IEEE Transactions on Software Engineering*, vol. SE-10, no. 1, pp. 4–21, 1984.

[2] S. O’Brien, "Here are the most in-demand jobs for 2019", *cnbc*, 2021. [Online]. Available: https://www.cnbc.com/2019/01/24/here-are-the-most-in-demand-jobs-for-2019.html. [Accessed: 31- Oct- 2021]

[3] "Jobs of Tomorrow: Mapping Opportunity in the New Economy", *World Economic Forum*, 2021. [Online]. Available: https://www.weforum.org/reports/jobs-of-tomorrow-mapping-opportunity-in-the-new-economy. [Accessed: 31- Oct- 2021]

[4] "Worldwide Professional Developer Population of 24 Million Projected to Grow amid Shifting Geographical Concentrations", 2021. [Online]. Available: https://evansdata.com/press/viewRelease.php?pressID=278. [Accessed: 31- Oct- 2021]

[5] "Global 500", *Fortune*, 2021. [Online]. Available: https://fortune.com/global500/2019/search/?profits=desc. [Accessed: 31- Oct- 2021]

[6] "Resume Screening: A How-To Guide For Recruiters | Ideal", *Ideal*, 2021. [Online]. Available: https://ideal.com/resume-screening/. [Accessed: 31- Oct- 2021]

[7] Marti A Hearst. Untangling text data mining. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pages 3–10. Association for Computational Linguistics, 1999.

[8] "Applicant tracking system - Wikipedia", *En.wikipedia.org*, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Applicant_tracking_system. [Accessed: 31- Oct- 2021]

[9] Stephen Clyde, Jianping Zhang, and Chih-Chung Yao. An object-oriented implementation of an adaptive classification of job openings. In Proceedings the 11th Conference on Artificial Intelligence for Applications, pages 9–16. IEEE, 1995.

[10] V Senthil Kumaran and A Sankar. Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping (expert). International Journal of Metadata, Semantics and Ontologies, 8(1):56–64, 2013.

[11] Abeer Zaroor, Mohammed Maree, and Muath Sabha. A hybrid approach to conceptual classification and ranking of resumes and their corresponding job posts. In International Conference on Intelligent Decision Technologies, pages 107–119. Springer, 2017

[12] Rogelio Valdez-Almada, Oscar M Rodriguez-Elias, Samuel Gonz, et al. Natural language processing and text mining to identify knowledge profiles for software engineering positions: generating knowledge profiles from resumes. In 2017 5th International Conference in Software Engineering Research and Innovation (CONISOFT), pages 97– 106. IEEE, 2017.

[13] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women", *U.S.*, 2021. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. [Accessed: 31- Oct- 2021]

[14] Jie Chen, Chunxia Zhang, and Zhendong Niu. A two-step resume information extraction algorithm. Mathematical Problems in Engineering, 2018, 2018.

[15] Ralph Grishman and Beth M Sundheim. Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996.

[16] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1):3–26, 2007.

[17] Venkat N Gudivada. Natural language core tasks and applications. In Handbook of statistics, volume 38, pages 403–428. Elsevier, 2018.

[18] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[19] G. Deepak, V. 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*, vol. 23, no. 1, pp. 157-165, 2020.

[20] A. Sinha, M. Amir Khusru Akhtar and A. Kumar, "Resume Screening Using Natural Language Processing and Machine Learning: A Systematic Review", *Machine Learning and Information Processing*, pp. 207-214, 2021.

[21] L. Sayfullina, E. Malmi, Y. Liao and A. Jung, "Domain Adaptation for Resume Classification Using Convolutional Neural Networks", *Lecture Notes in Computer Science*, pp. 82-93, 2017.

[22] S. Nasser, C. Sreejith and M. Irshad, "Convolutional Neural Network with Word Embedding Based Approach for Resume Classification", in *2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR)*, Ernakulam, India, 2018.

[23] I. Ali, N. Mughal, Z. Khand, J. Ahmed and G. Mujtaba, "Resume Classification System using Natural Language Processing and Machine Learning Techniques", 2022. [Online]. Available: https://doi.org/10.22581/muet1982.2201.07. [Accessed: 09- Feb- 2022].

[24] M. Alamelu, D. Kumar, R. Sanjana, J. Sree, A. Devi and D. Kavitha, "Resume Validation and Filtration using Natural Language Processing", *2021 10th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)*, 2021. Available: 10.1109/iemecon53809.2021.9689075 [Accessed 1 February 2022].

[25] A. Sinha, M. Amir Khusru Akhtar and A. Kumar, "Resume Screening Using Natural Language Processing and Machine Learning: A Systematic Review", *Machine Learning and Information Processing*, pp. 207-214, 2021. Available: 10.1007/978-981-33-4859-2_21 [Accessed 9 February 2022].

[26] A. Chamberlain, "How Long Does it Take to Hire? Interview Duration in 25 Countries", 2022. [Online]. Available: https://www.glassdoor.com/research/time-to-hire-in-25-countries/. [Accessed: 09- Feb- 2022].