IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303223|

Resume Parsing Using Named Entity Recognition and Hugging Face Model

Rimmalapudi Rajesh, Rankela Sai Sri Harsha, Busi Joseph, Vara Lakshmaiah Duggineni, Janardhana Rao Alapati

Department of CSE (Artificial Intelligence & Machine Learning), Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

Asst. Professor, Department of CSE (Artificial Intelligence & Machine Learning), Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

ABSTRACT: New methods in networking and communication have paved the way for enhancing the recruitment process by creating e-recruitment recommender systems. The growing popularity of online recruitment has led to a high volume of resumes being saved in recruitment platforms. Resumes are typically designed in various styles to grab the attention of recruiters, using different sizes, colors, and table formats. Nevertheless, the diversity of formats significantly impacts data mining tasks like extracting resume information, matching profiles automatically, and ranking applicants. Rule-based, supervised, and semantics-based techniques have been developed for precise extraction of resume facts, but they rely heavily on extensive annotated data, which can be challenging to gather. Moreover, the time-consuming nature of these techniques and the incomplete knowledge they possess greatly impact the accuracy of resume parsing. In this manuscript, we introduce a framework for parsing resumes that addresses the constraints encountered in previous methods. Initially, the unprocessed content is taken from CVs and sections are divided by classifying text blocks. Moreover, the entities are identified using named entity recognition and enhanced using Hugging Face. The resume parser proposed accurately grabs details from resumes which directly help in selecting the top candidate.

KEYWORDS: Resume parsing, Data Enrichment, Text Extraction, Hugging Face, Boolean Naive Bayes

I. INTRODUCTION

Recent advancements in networking and communication have led to increased information exchange on the internet, with job seekers using online job boards to share their resumes, resulting in a large number of job postings and CVs online. A study found that over one million resumes are submitted annually to platforms like Monster.com and LinkedIn.com. Recruiters typically spend around 2 minutes reviewing each resume, leading to the quick scanning for key information and potential oversight of important details. This has driven the development of e-recruitment recommender systems, which have become the primary method of hiring in many industries. While these systems have helped streamline the hiring process and reduce time and costs, they also pose challenges such as candidates applying for unsuitable roles due to the limitations of parsing resumes accurately.

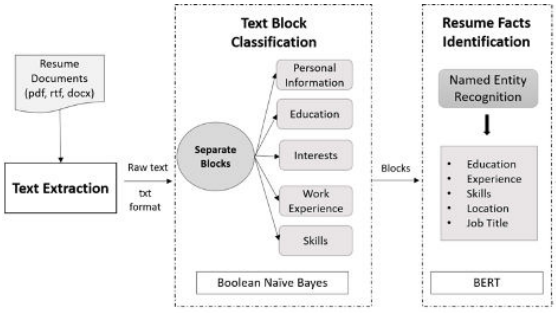

Fig. 1: Proposed Framework for Resume Parsing

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303223|

Resume parsing, crucial for effective hiring, encounters challenges due to varying formats (PDF, docx, rtf) and layouts (list-style, table-style) with elements like info-graphics requiring image processing. This complexity makes candidate recommendation parsing crucial for evaluation. Machine parsing struggles due to complexity, prompting research into plain text, heuristic, and neural network methods. Gathering data across domains remains a focus, despite research gaps in layout details and domain knowledge. This study introduces a framework converting unstructured resumes into structured, data-enriched outputs. It employs Named Entity Recognition and a domain-neutral ontology to enhance skill entities, improving candidate profile matching and ranking.

II. LITERATURE REVIEW

Research has long been centered around resume parsing. Various methods have been developed for precise information extraction and are categorized as follows.

A. Heuristics-Based Methods:

Chen et al. [7] introduced a two-phase algorithm for extracting data from unformatted resumes, leveraging the 'Writing Style' approach, Naive Bayes, and NER. Gunaseelan et al. [5] developed a machine learning-based system for segment extraction from resumes, employing two categorization methods. Despite successes, challenges persisted in handling diverse document structures. Additionally, Chen et al. [12] proposed a hierarchical method for PDF resume extraction, utilizing heuristic rules and a CRF model. While effective for list-formatted resumes, it lacked compatibility with table-formatted ones and primarily focused on specific resume sections like personal details and education.

B. Framework Based Methods:

A different method for parsing resumes and matching them with job profiles was suggested by Deepak et al [13]. The system for parsing resumes had 4 stages: text segmentation, NER, merging co-references and conflict resolution, and text normalization. Initially, the resume was divided into sections containing related data and then named entities were categorized using NER. Additionally, text normalization was carried out to ensure that the extracted entities were consistent and trustworthy. Finally, o-reference resolution was employed to clarify the abbreviations in the resume and examine their linguistic expressions. The resume parser's results were utilized to rank resumes based on a specific job description.

C. Skill Extraction Methods:

Various researchers have focused on skill extraction in resume parsing, particularly aiming to identify implicit skills. Kameni and NoHugging Face [14] devised a method utilizing a CNN-based multi-label classification model to predict high-level skills from unprocessed CV content. They transformed the content into matrices and employed binary CNN classifiers to categorize resumes into occupation classes. Similarly, Chifu et al. [12] proposed an ontology-driven system to extract professional skills from natural language resumes. Their method incorporated pattern induction and skill detection modules to identify unfamiliar skills mentioned in resumes. Additionally, they addressed the ambiguity of skill phrases using Wikipedia..

D. Deep Learning Based Methods:

Zu and Wang [6] proposed a text block classification method for precise resume information extraction. Utilizing neural network-based text classifiers, their approach focused on accurately categorizing text blocks and lines within them. They introduced a novel block segmentation method leveraging line details and word representations. This paired with line type and label classification facilitated precise text block categorization. Meanwhile, Nguyen et al. [14] introduced a four-phase method for resume data extraction. It involved text segmentation, rule-based NER entity extraction, and DNN-based entity sequencing. Despite effectiveness, limited training data hindered DNN performance. Additionally, the study targeted IT resumes, potentially impacting applicability across different fields.

III. PROPOSED FRAMEWORK

This part discusses the input, output, and thorough explanation of the suggested resume parsing framework. Fig.1 displays the process of parsing resumes.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303223|

A .Input and Output:

This pipeline aims to extract information from resumes without taking into account the document's font styles, font sizes, or standard structure.

1. Input : The resume parser takes in unstructured documents of various file types such as pdf, docx, and doc files.
2. Output : Output is presented in a structured format detailing all relevant information and details about a candidate obtained from their resume.

B . Text Extraction:

Resumes exhibit diverse formats, including document type, layout, writing style, and structure, attributed to the lack of standardized formats. To enable text mining techniques, raw text extraction from resumes is necessary. The proposed method utilizes PDFbox to retrieve unprocessed text from PDF resumes, as depicted in Fig. 2. However, this approach is limited by PDFbox's inability to extract text from non-PDF document formats, necessitating conversion to PDF before extraction.

Fig. 2: Conversion of PDF file to Txt file

1) Data Preprocessing: The gathered information undergoes preprocessing to eliminate noise introduced by PDFbox, such as non-ascii characters, unnecessary punctuation, missing and extra spaces, and newlines.

i) Missing or Extra Spaces: Because of the varied font styles and sizes found in resumes, the extracted words may have additional spaces or lack spaces altogether. For example, if spaces are missing, 'I am a competent person' will become 'Iamacompetentperson' or 'I a m a c o m p e t e n t p e r'. In the event of additional spaces. This creates disturbances in the text and is removed prior to classifying text blocks.

ii)*Punctuation Marks:* The extracted text was full of punctuation marks, particularly at the start and in the middle. They create disturbance in the data by altering the meaning of a word which hinders the process of extracting information. For example, Adobe Illustrator and Adobe Illustrator are considered the same skill, but are recognized as two distinct words by the model. It is eliminated during data-preprocessing to prevent bias in similar types of text.

iii) *Non-Ascii characters:* PDFbox produces certain non-ascii characters in the resulting text document when extracting text in regards to bullet points, emoticons, or icons. Before the text block classification training, any non-ascii characters in the extracted text are eliminated. An instance of non-ascii characters being produced during text extraction can be seen in Figure 3.

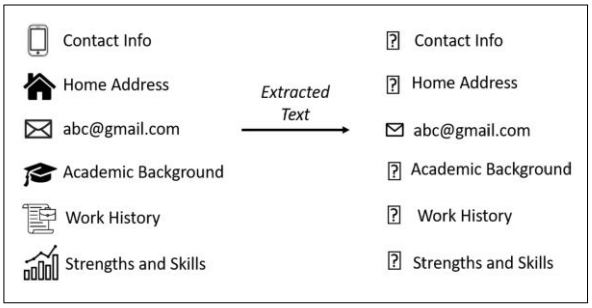

Fig. 3: Ascii characters generation

C . *Text Block Classification:*

Resume parsing relies heavily on categorizing text blocks, crucial for subsequent modules. Resumes typically follow a hierarchical structure, with blocks encompassing diverse topics like personal details, employment history, skills, and

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303223|

interests. Each block contains specific information related to its category, such as location in personal details or educational background in academic history. This study focuses on five key categories, including academic background and skills, as indicators of an individual's qualifications. While both machine learning and deep learning methods are employed for block categorization, neural network models excel in automatic feature identification and pattern recognition. However, substantial data is essential for effective neural network training.

The pipeline we suggest uses the Boolean Naive Bayes (BNB) algorithm, a probabilistic model, to categorize sections of resumes. The text data that has been pre-processed is utilized to create the vocabulary for the five specified information fields, which are then provided as input to the text block classification module.

1)Boolean Naive Bayes Alogorithm: BNB is a probabilistic model that is supervised and makes classifications based on the highest probability from the possible classes. Moreover, it eliminates duplications in the text prior to determining probabilities. It relies on Bayes' theorem assuming that features are independent, meaning each feature is not related to any other feature when considering the class label. The calculation of the probability for each class in Naive Bayes is shown in equation (1) .

$$\hat{C} = argmax_{c \in C} \frac{P(x|c) \times P(c)}{P(x)} \quad (1)$$

P(x|c) is the likelihood probability of variable z given class c and P(c) is the prior probability of a certain class c among the set of classes C where C = c_1 ,c_2,....,c_n.

2) BNB in resume parsing: The method involves training 5 language models for 5 classes (education, skill, experience, interest, and personal information) using the multinomial boolean naive bayes algorithm. Each model categorizes particular data sections found in the resumes. Table I provides the specifics of these information blocks.

TABLE I: **Details of Information Blocks**

Models	**Information Block**
Education Model	Academic Qualification, Certifications, Publica- tions
Experience Model	Work History, Professional Experience, Organi- zation, Responsibilities
Skills Model	Linguistic Skills, Technical Skills, Soft Skills, Computer Skills
Interests Model	Hobbies, Interests, Extra Curricular Activities
Personal Info Model	Personal Data, References

3) *Training:* The goal of BNB training is to create five language models using the five different vocabularies - education, skills, personal information, experience, and interests - produced during data-preprocessing, as illustrated in Fig. 4.

All of these words are combined together to create one comprehensive vocabulary V. Each word w in the vocabulary has a conditional probability P(w/c) computed for a specific class c, with this process being repeated for all classes C. The probability of a word given a certain class P(w/c) is determined using Laplace Smoothing (LS) formula instead of a simple count ratio as demonstrated in Eq (2). LS has the benefit of ensuring that no word will be assigned a zero probability for any particular class. If a word is not found in the vocabulary of a specific class c out of the five classes

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303223|

C, a small probability is still assigned to avoid having a total zero probability during testing. The probability of each word in the vocabulary is calculated conditionally as follows:

$$P(w_i|c) = \frac{count(w_i, c) + 1}{count(w, c) + |V|} \quad (2)$$

Count of a specific word in a class's vocabulary is represented as count(w_i , C), while the total number of words in the vocabulary of that class is represented as count(w, c). |V| signifies the total number of words in the overall vocabulary.

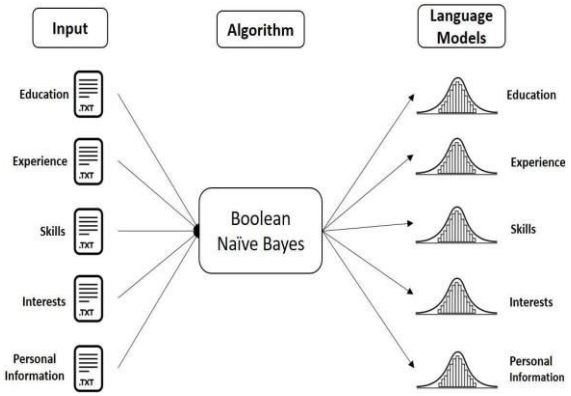

Fig. 4: Training of Boolean NB Algorithm

A probability is computed for the given word in the vocabulary for each class and saved in its corresponding model. This leads to the creation of 5 models of the same size as the shared vocabulary, each with unique probabilities.

4) Testing: Initially, the unprocessed text from a potential candidate's resume is extracted and prepared in the same manner as described in the previous section. Later, each section in the resume is categorized into one of the five classes. At first, the BNB algorithm eliminates repeated words in each block as required. Five language models are used to determine the conditional probability P(w|c) of each unique word w in a given block b to be classified.

Algorithm 1 Test BooleanNB(C, t, $prior$, $condprob$, B)

Input:
C: Set of classes
t: All words in a single block
B: Set of blocks in a resume
$prior$: Prior probability of a certain class
$condprob$: Conditional probability of a word given a certain class

procedure BNBTest(C,t,$prior$,$condprob$,B)
for each $b \in B$ **do**
  $W \leftarrow ExtractUniqueWordsfromBlock$(t,b)
  for each c $\in C$ **do**
    $score[c] \leftarrow log(prior[c])$
    for each $w_i \in W$ **do**
      $score[c]$ += $log(condprob[w_i][c])$
  Ĉ ←argmax$_{c \in C}$ $score[c]$
  Assign label Ĉ to b

Output:
Set of class labels assigned to B

Moreover, the likelihood probability for a given block is calculated by multiplying the conditional probabilities of each unique word in the block. It is then combined with the previous probability P(c) of a specific class e in order to

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303223|

calculate the ultimate probability score prevent a specific group from taking action. Nevertheless, the issue of underflow arises due to the multiplication of probabilities. The log adjusted formula is employed to determine the ultimate probability score of a specific block for each model in order to address this limitation. Furthermore, the block is assigned the class with the highest probability.

D. *Resume Facts Identification:*

Following the classification of the text blocks, distinct blocks are generated for personal details, educational history, skills, job experience, and hobbies. NER utilizes them to identify the named entities in the text passages. NER is a crucial component of the process of extracting information, aiming to identify, extract, and categorize named entities in text into specific groups . The chosen entities for this research project are:

- Education - title of degree and area of study for the applicant on their resume
- Experience - duration a candidate was employed at a particular job and name of the organization
- Skills - the technical abilities listed on the CV
- Job Title - candidate's most recent job title
- Location - the job seeker's place of location

Initially, NER methods heavily depended on rule and dictionary-based strategies, which required rule sets created by experts in specific fields. Nevertheless, the regulations do not encompass every instance of language use, the creation procedure is excessively lengthy, and is less probable to be transferable . Lately, the progress of deep learning in numerous Natural Language Processing (NLP) activities has sparked a transformation in the information extraction field, obtaining state-of-the-art outcomes without requiring manually created features. Additionally, transfer learning has proven to be successful in various NLP tasks by leveraging pre-trained models to apply previously acquired knowledge to tasks specific to a particular domain. HUGGING FACE is a language model that utilizes transformer architecture and focuses on context. Positional embedding is utilized in lieu of positional encoding to produce the contextualized embedding of a full sentence [11]. In this document, a HUGGING FACE model that has been fine-tuned is employed to identify and categorize the named entities described earlier. The pre-trained HUGGING FACE language model parameters are extensively trained with annotated data to encode information about our problem domain. Different types of data sets are utilized for training various sections of a resume to prevent class imbalance issues.

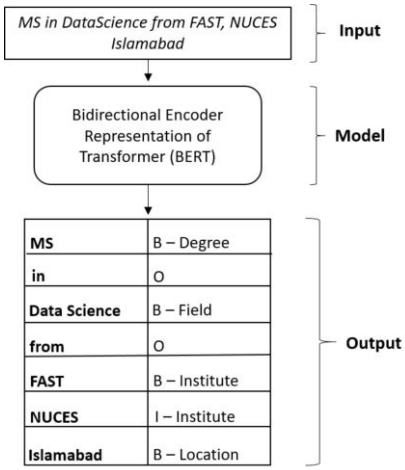

Fig. 5: Process of Named Entity Recognition

Following the adjustment of the HUGGING FACE model, the unstructured input goes through processing and is labeled with IOB tags as well as the tags employed during the training of the HUGGING FACE model. The outcome was a structured output where each word was labeled with the named entity, as depicted in Figure 5. IOB formatting, which stands for Inside, Outside, Beginning, is a format for tagging tokens that is commonly utilized in applications like NER. They are akin to part-of-speech tags, but they also provide details on where an entity is situated in the text.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303223|

IV. RESULT

The suggested resume parsing framework showed efficient performance in retrieving organized data from various resume styles and designs. Using the Boolean Naive Bayes algorithm for text block classification accurately sorted resume sections into categories such as education, skills, work experience, personal details, and interests. The categorization models, which were trained using specialized vocabularies for each category, successfully achieved a high level of accuracy in dividing the unorganized resume text into separate labeled sections that correspond to various types of information.

Using a fine-tuned HUGGING FACE model, the named entity recognition module successfully detected and extracted important entities such as degree names, organizations, skills, locations, and job titles from the categorized text blocks. Including a personalized skill ontology improved the identified skill entities by creating semantic connections and inferring additional skills not directly mentioned in the resume. This additional data enrichment step boosted the depth of the candidate skill profiles produced by the framework. In general, the outcomes confirmed the effectiveness of the resume parsing process in managing different resume formats and creating organized, enhanced candidate profiles that are appropriate for ranking and matching in electronic recruitment systems.

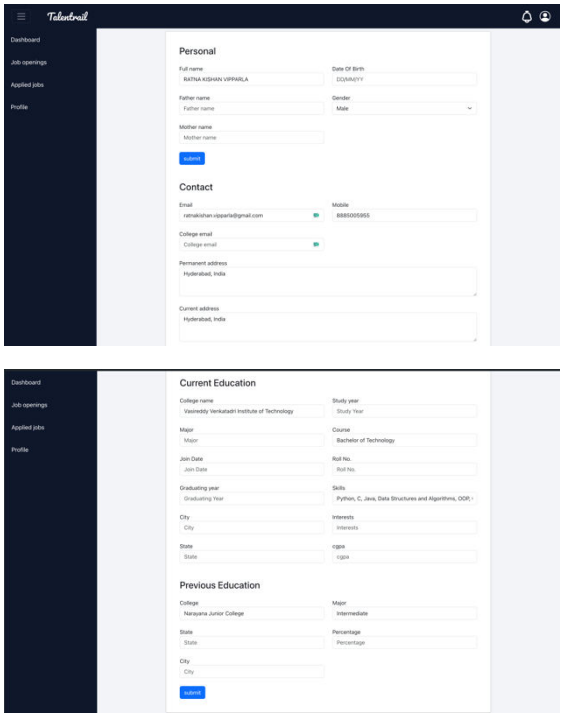

Fig. 6: Parsed Data

V. CONCLUSION

This paper introduces a framework for parsing resumes that can extract information from resumes with a variety of text formats and layouts. The objective of this study is to improve the precision of extracting information from individual resumes to help in selecting the most suitable candidate. In the suggested framework, the key processing steps include extracting text, classifying text blocks, and identifying resume facts using NER. Moreover, the skills that are taken out are enhanced with a specifically designed ontology. This framework streamlines the laborious task of manually creating custom features for text block classification and NER by saving time. The second addition is the method utilized for categorizing text blocks, involving the employment of Boolean Naive Bayes to categorize blocks into various classes. Additionally, individual NER training is conducted for each type of block to prevent class imbalance. Finally, the ultimate achievement is the creation of a personalized skill ontology for enhancing data. The focus is on various areas to encompass diversity and enhance reusability in the realm of e-recruitment. Plans for future work include

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303223|

incorporating an experiments and evaluation section to showcase results of the proposed framework using actual e-recruitment data.

REFERENCES

The template will number citations consecutively within the bracket

[l] W. Abdessalem and S. Amdouni, "E-recruiting supp011 system based on text mining methods," *International Journal of Knowledge and Learning,* vol. 7, pp. 220-232, 2011.

[2] E. K. Ryland and B. Rosen, "Personnel professionals' reactions to chronological and functional resume formats," *Career Development Quarterly,* vol. 35, no. 3, p. 228-238, 1987.

[3] S. Alotaibi and Y. Mourad, "Job recommendation systems for enhancing e-recruitment process," in *Proceedings of the International Conference on Ii1formatio11 and Knowledge Engineering,* 2012.

[4] "Analysis and shortcomings of e-recruitment systems: Towards a semantics-based approach addressing knowledge incompleteness and limited domain coverage," *Journal of lr!formation Science,* vol. 45, no. 12, p. 016555151881144, 2018.

[5] S. K. Kopparapu, "Automatic extraction of usable information from unstructured resumes to aid search," in *2010 IEEE International Con- ference on Progress in b(fonnatics and Computing,* vol. 1, pp. 99-103, 2010.

[6] S. Zu and X. Wang, "Resume information extraction with a novel text block segmentation algorithm," *Linguistics,* vol. 8, no. 5, pp. 29-48, 2019.

[7] J. Chen, C. Zhang, and Z. Niu, "A two-step resume information extraction algorithm," *Mathematical Problems in Engineering,* 2018.

[8] V. Bhatia, P. Rawat, A. Kumar, and R. R. Shah, "End-to-end resume parsing and finding candidates for a job description using bert," 2019.

[9] K. Yu, G. Guan, and M. Zhou, "Resume information extraction with cascaded hybrid model," in *Association for Computational Linguistics,*
p. 499-506, 2005.

[10] A. Singh, C. Rose, K. Yisweswariah, V. Chenthamarakshan, and
N. Kambhatla, "Prospect: a system for screening candidates for recruit- ment," in *Proceedings of the 19th ACM international conference on ltiformation and knowledge management,* p. 659-668, 2010.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL- HLT,* 2019.

[12] J. Chen, L. Gao, and Z. Tang, "Information extraction from resume documents in pdf format," in *Document Recognition and Retrieval,* 2016.

[13] 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 Cryptogra- phy,* vol. 23, no. 1, pp. 157-165, 2020.

[14] K. F. F. Jiechieu and N. Tsopze, "Skills prediction based on multi-label resume classification using cnn with model predictions explanation," *Neural Computing Applications,* vol. 33, pp. 5069-5087, 2021.

[15] E. S. Chifu, V. R. Chifu, I. Popa, and I. Salomie, "A system for detecting professional skills from resumes written in natural language," in *2017 13th IEEE International Co1iference on Intelligent Computer Communication and Processing (ICCP),* pp. 189-196, 2017.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details