Multimedia Tools and Applications (2024) 83:82069–82087
https://doi.org/10.1007/s11042-024-18778-9

Towards smarter hiring: resume parsing and ranking with YOLOv5 and DistilBERT

Shakti Kinger[1] · Divija Kinger[2] · Shivam Thakkar[2] · Devashish Bhake[2]

Received: 17 August 2023 / Revised: 9 February 2024 / Accepted: 25 February 2024 /
Published online: 12 March 2024

Abstract

In the contemporary landscape of recruitment, the Applicant Tracking System (ATS) plays a pivotal role in automating the screening and shortlisting of candidates. However, the prevailing ATS encounters challenges such as imprecise data extraction, erroneous keyword selection, and a lack of standardized criteria for comparison. As a result, many deserving applicants are turned away, highlighting the necessity for a more complex and human-centered strategy. In response to these limitations, our research introduces an innovative Resume Parsing and Ranking solution. Leveraging advanced natural language processing techniques and machine learning algorithms, our system provides a customized experience for the automated screening process. The naive methods underscore the distinct advantages of our innovative approach, emphasizing the need to build a robust and accurate model for Resume Parsing and Ranking. Notably, it addresses discrepancies arising from diverse resume structures, ensuring a standardized and equitable evaluation of all applicants. The main contribution of our work lies in the development of a state-of-the-art Resume Parser that enhances efficiency, reduces bias, and optimizes candidate selection outcomes. Our proposed method integrates cutting-edge technologies to refine the existing ATS process, offering a tailored and precise approach to resume evaluation. The primary problem addressed is the lack of precision and standardization in thecurrent ATS, leading to sub-optimal candidate shortlisting. Our solution tackles this by introducing a comprehensive framework that mitigates the impact of varied resume structures, thereby promoting fair and consistent candidate assessment. Through empirical validation, our obtained results showcase an accuracy of 96.2% in resume parsing, thereby significantly improving the efficiency of the candidate selection process.

Keywords Natural Language Processing (NLP) · DistilBERT · Resume Parsing · Resume Ranking

✉ Shakti Kinger
shakti.kinger@mitwpu.edu.in

[1] School of Engineering and Technology, Dr. Vishwanath Karad MIT-WPU, S.No.124, Paud Road, Kothrud, Pune 411038, Maharashtra, India

[2] K. J. Somaiya Institute of Technology, Somaiya Ayurvihar Complex Eastern Express Highway Near Everard Nagar, Sion East, Mumbai 400022, Maharashtra, India

1 Introduction

Automating the extraction of critical data from resumes or CVs and converting unstructured data into a structured format for effective processing and analysis is the main goal of a resume parser. During this process, important information like the candidate's name, contact details, employment history, education, and skills must be identified and extracted. By using machine learning and natural language processing (NLP) techniques, resumes can be quickly screened and compared, expediting the hiring process and saving time and money.

A resume parser can be used to quickly and accurately filter large volumes of resumes; it can also automate the labor-intensive manual review process and improve hiring and recruitment decision-making in general. Many different types of organizations, such as recruiting managers, HR departments, and job boards, have embraced this technology. In order to extract information, modern resume parsers use techniques like named entity recognition, keyword extraction, and syntactic analysis. Some even use machine learning techniques to gradually increase the accuracy of information extraction and classification.

Our proposed solution aims to elevate the resume parsing process by harnessing the capabilities of modern natural language models and computer vision techniques. Extensive research led us to NLP models like ALBERT [21], MobileBERT [35], ELECTRA [8], DeBERTa [14], RoBERTa [25], and DistilBERT [32], which have exhibited superior performance compared to the widely used BERT [10] model. DistilBERT, in particular, emerged as the optimal choice for classification and named entity recognition, offering efficiency with reduced computational demands.

Recognizing the potential complexity of implementing an advanced system, especially for organizations with limited technical resources, we emphasize a user-friendly and adaptable design. In our paper, we recognize the challenges posed by big data storage and cloud computing in information management, echoing the concerns addressed in the paper [26]. Just as task allocation management is crucial for enterprise sustainability, our research acknowledges the significance of efficiently parsing and ranking resumes for effective talent acquisition processes. Leveraging advanced techniques such as YOLOv5 and DistilBERT, we propose an optimization approach to address the challenges of processing large volumes of resume data and leveraging cloud computing resources.

Keeping in mind the wide range of resume styles and job specifications, our solution is fundamentally flexible. The system is always being tested and improved upon to make sure it works with different resume formats and can adapt to the ever-changing world of different job requirements. With careful validation and iteration, we hope to provide a solution that successfully addresses the complex requirements of various sectors and occupations.

2 Literature review

This section provides a comprehensive exploration of diverse methodologies and advancements in the field of resume parsing and ranking. In the contemporary realm of recruitment, the critical process of resume parsing, aimed at extracting essential information from resumes, has witnessed significant enhancements through the integration of cutting-edge technologies such as Natural Language Processing (NLP), Named Entity Recognition (NER), Neural Networks, BERT-based methods, and innovative Hybrid approaches. This section is structured into sub-sections, each dedicated to a specific technological facet, delving into seminal studies and their unique contributions within the respective domains.

2.1 Resume parsing using NLP and named entity recognition

In the realm of modern recruitment, the pivotal process of resume parsing, extracting essential information from resumes, is significantly enhanced through the incorporation of Natural Language Processing (NLP) and Named Entity Recognition (NER) technologies. Several studies have delved into this intersection, each providing unique insights and methodologies. The paper [3] underscores the utilization of NLP techniques to efficiently extract crucial information from resumes, contributing to the streamlined shortlisting of candidates and facilitating companies in selecting high quality employees. Another study [33] focuses on keyword identification, clustering, and the presentation of the most relevant resumes to employers based on sophisticated keyword matching, incorporating user interaction for verification. Meanwhile, [9] implements NLP-driven resume parsing, leveraging an entity linking paradigm to enhance the matching process based on job criteria, including a unique pie chart display for candidates. In a comprehensive approach involving text segmentation, NER, named entity clustering, and text normalization is employed, providing a detailed understanding of resume content through the identification and grouping of entities. Lastly, [20] proposes a system automating candidate recruitment through the extraction of specific data points from resumes, with job matching relying heavily on the accuracy of the text extraction process. These studies highlight how NLP (Natural Language Processing), NER (Named Entity Recognition), and resume parsing work together to improve candidate shortlisting, extract information efficiently, and enhance job matching processes.

2.2 Resume parsing using neural networks

In the rapidly evolving landscape of resume parsing, the infusion of neural networks has emerged as a formidable strategy, offering sophisticated algorithms for extracting valuable information. This literature review delves into pivotal findings and methodologies from a series of studies that focus on resume parsing through the lens of neural networks. The paper titled [2] introduces a comprehensive system utilizing Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi- LSTM), and Conditional Random Field (CRF) models. The CNN classifies segments, while the CRF and Bi-LSTM-CNN models handle sequence labeling for tagging entities, achieving segmentation and extraction of 23 fields. Another study [24] emphasizes the development of efficient parsing algorithms and highlights the superior performance of the Convolutional Recurrent Neural Network (CRNN), showcasing a remarkable 96% accuracy. Additionally, the study focuses on parsing original resumes by splitting them into words and utilizing a word base for job suitability prediction. In [18], the proposed model not only extracts skill-related terms but also identifies high level skills, providing a holistic understanding of skills expressed in resumes. Despite the model's evaluation on a set of IT resumes, the paper lacks insights into scalability and generalizability. The study [11] specifically addresses the education section, employing a semi-supervised deep learning model with initial training on annotated sections, resulting in improved accuracy. Lastly, [28] proposes a hierarchical approach using CNN with Glove-Word Embedding for resume classification, creating a structured hierarchy for enhanced categorization. Collectively, these studies underscore the potency of neural networks in refining the efficiency and accuracy of resume parsing, shedding light on the intricate task of extracting meaningful insights from resumes.

2.3 Resume parsing using BERT-Based methods

In the dynamic landscape of resume parsing, the integration of BERT-based methods has emerged as a prominent and innovative approach, leveraging Bidirectional Encoder Representations from Transformers to enhance contextual understanding and refine the extraction of information from resumes. This comprehensive literature review delves into key findings and methodologies from several noteworthy studies that harness the power of BERT in the realm of resume parsing. The paper titled [5] introduces a two-step hybrid Information Retrieval methodology, utilizing Boolean Naive Bayes with Laplacian smoothing and a tri-gram approach for text block classification and incorporating BERT-cased for entity recognition. Despite achieving an average F1 score of 0.52 in BERT-based entity recognition, the paper acknowledges certain limitations in accurately recognizing named entities. In [36] the study proposes the adoption of BERT vectorization to identify text contextually and enhance resume parsing through contextual meaning extraction. This approach aims to capture nuanced contextual information from resumes, contributing to improved parsing accuracy. Another significant contribution comes from the paper [4] which presents an end-to-end solution for ranking candidates based on their suitability for a job description. The authors develop a robust resume parser that extracts comprehensive information from candidate resumes, utilizing BERT sentence pair classification for ranking. BERT's incorporation enhances contextual understanding, leveraging past job experiences mentioned in resumes for more accurate approximation. Lastly, the study [22] introduces a fine-tuned Siamese Sentence-BERT model for matching jobs and job seekers. The model incorporates textual information from vacancies and resumes, outperforming unsupervised and supervised baselines. However, the paper lacks a comprehensive discussion on potential biases in the labeled dataset used for model training and the interpretability of the model's embeddings. Collectively, these studies underscore the efficacy of BERT-based methods in advancing the field of resume parsing, providing a nuanced understanding of contextual information, and contributing to the improvement of candidate ranking and job matching processes.

2.4 Resume parsing using hybrid methods

The evolution of resume parsing techniques has witnessed a surge in the exploration of hybrid methods, combining diverse approaches to enhance the accuracy and efficiency of information extraction from resumes. The study titled [31] investigates the integration of rule-based, supervised, and semantics-based methods for accurate fact extraction from resumes. The approach involves initial raw text extraction, text block classification for separating blocks, and named entity recognition enriched with ontology for entity extraction. While promising, the study acknowledges challenges such as varied resume styles impacting data mining operations and the heavy dependence on large annotated datasets, leading to knowledge incompleteness. In the paper [1], the authors propose an automated Resume Classification System (RCS) leveraging Natural Language Processing (NLP) techniques and Machine Learning (ML) algorithms. The study focuses on preprocessing resumes into a corpus and employing vectorized representation through NLP techniques for effective classification. The contribution lies in the integration of NLP and ML for accurate resume categorization, addressing the challenges of diverse resume formats. The study [37] introduces an end-to-end pipeline for resume parsing utilizing neural network-based classifiers and distributed embeddings. The novel text block segmentation algorithm efficiently segments resumes into predefined text blocks. The comparative evaluation highlights the superiority of BLSTM CNNs-CRF in named entity recognition, showcasing the effectiveness of hybrid methods. [23] combines a parametric form of linear chain Conditional Random Field (CRF) with BERT-based models for accurate resume information extraction. The incorporation of BERT pre-training, bidirectional long and short-term memory networks (BiLSTM), and CRF aims to address the challenges of accurately extracting information, especially in the presence of polysemism. The paper acknowledges the limitations of the model in handling words with multiple meanings.

2.5 Resume ranking and matching

The literature on resume ranking and matching encompasses diverse approaches to enhance the recruitment process. In [27], a Semantic Web-based recommendation system called Smart Applicant Ranker is introduced, utilizing ontology to match candidate resumes with job requirements efficiently. However, the paper lacks detailed information on profile modeling and a comprehensive evaluation of the ranking algorithm. The paper [17] proposes the use of Sentence-BERT for resume shortlisting, demonstrating its superiority over BERT. Yet, concerns about dataset diversity and addressing language variations or cultural differences are not thoroughly explored. Another paper [12] leverages a domain-specific ontology and machine learning for a fine-grained relationship between skills. However, the paper does not discuss challenges in extracting job-related content [34]. The methodology in the paper [30] uses dynamic parameter optimization strategies to perform task scheduling. Utilizing the task execution time distribution methodology helps handle uncertainties in resume ranking, enhancing accuracy in processing varying work experiences and educational backgrounds. The paper [7] employs Chinese text segmentation and stop-word filtering, potentially limiting its ability to accurately assess candidate suitability. Lastly, [13] proposes a supervised learning approach without benchmark comparisons. Each approach contributes to the evolving landscape of resume ranking and matching, but further exploration and evaluation are warranted.

Table 1 Best average performance metrics achieved by each approach

Section	Best Avg. Accuracy	Best Avg. F-Measure	Best Avg. Precision	Best Avg. Recall
Natural Language Processing and NER [9]	93.98%	93.99%	95.44%	92.52%
Neural Network based Approach [18]	-	-	-	96.79%
BERT based Approach [4]	97%	-	-	-
Hybrid Approaches [23]	-	89.69%	91.41%	88.17

In conclusion, this literature review has provided a comprehensive overview of diverse approaches employed in the field of resume parsing, highlighting the significance of Natural Language Processing (NLP), Named Entity Recognition (NER), Neural Networks, BERT-based methods, and Hybrid approaches. Table 1 shows the best average performance metrics achieved by each approach.

While the presented studies have significantly advanced our understanding of resume parsing, the below identified research gaps underscore the need for further exploration.

3 Research gaps

While the reviewed literature provides valuable insights into various aspects of resume parsing, several research gaps and areas for further exploration emerge:

1. Lack of Generalizability: Across Different Industries with Distinct Resume Styles: Most studies narrowly focus on the application of NLP and NER in the context of IT resumes, neglecting to explore how these technologies perform across different industries with distinct resume styles and terminology. This limitation hinders the generalizability of the findings and restricts the applicability of NLP and NER to a broader range of professional sectors.
2. Limited Exploration of User Interaction and Verification in NLP-Driven Parsing: The studies emphasizing user interaction for verification in NLP-Driven parsing are scarce, resulting in a lack of insights into the effectiveness and user acceptance of such interactive approaches.
3. Resource-Intensive Nature of CNN and BERT Models: The literature review often overlooks the practical implications of implementing CNN and BERT-based models, which are known for their resource-intensive nature. There is a discernible lack of discussion on the challenges users may face when attempting to run these computationally demanding models on standard personal computers. This gap raises concerns about the accessibility and feasibility of deploying such models in real-world scenarios, especially for users with limited computational resources.
4. Inadequate Capture of Contextual Nuances in Neural Network-based Parsing Models: While neural network-based models, especially CNN, Bi-LSTM, and CRNN, showcase high accuracy, there is a significant gap in understanding how well these models capture contextual nuances in resumes. The failure to address this limitation raises questions about the adaptability of these models in real-world scenarios with diverse contextual variations.
5. Errors Introduced due to Reliance on Simple ML Algorithms: As discussed in [29], we understand that ML models can be prone to errors, face difficulties in acquiring high-quality training data, and require substantial computational resources. Our aim is to bridge these gaps by introducing an innovative approach that leverages DistilBERT, a more computationally efficient variant of BERT, addressing the resource-intensive nature of traditional BERT models. Moreover, we propose a novel hybrid model that integrates DistilBERT with carefully designed neural network components, aiming to enhance both efficiency and accuracy in resume parsing. Through this approach, we aspire to contribute to the field by offering a practical solution that not only narrows existing research gaps but also provides a more accessible and efficient framework for users with limited computational resources.

4 Dataset and processing

4.1 Data collection

To extract sectional data from resumes, we sought a dataset featuring marked bounding boxes around different resume sections. Utilizing an open-source dataset with resume images and corresponding bounding boxes, we ensured diversity in resume for- mats while mitigating potential biases associated with proprietary data sources. This widely recognized dataset is prevalent among researchers and provides a foundation for our study.

For text classification into distinct sections, we curated a dataset comprising sections from resumes categorized into ten classes: Personal Details, Education, Skills, Interests, Experience, Projects, Languages, Soft Skills, Certifications, and Achievements. This dataset was constructed using 5000 resumes segmented by our Computer Vision model and text extraction system. Classification was performed by a Zero Shot Classifier [34] (DeBERTa v3 [16]), an enhanced version of DeBERTa [15]. Our team performed thorough manual cross-verification to guarantee the quality and reliability of the dataset. We carefully validated each classification produced by the Zero Shot Classifier against the ground truth labels to eliminate any possible misclassifications. This extensive verification process enhances the credibility and suitability of the acquired data for training and testing our models.

For Named Entity Recognition, we developed a dataset with manual tagging of resumes, encompassing classes such as PERSON, LOCATION, SKILLS, CONTACT, EMAIL, ROLE, ORGANISATION, DURATION, CERTIFICATE, PROJECTS, EDUCATION, INSTITUTION, and LINKS (Fig. 1).

4.2 Data preprocessing

To assure its quality and compatibility, the acquired resume data underwent a variety of preprocessing stages before being used for training and analysis. The collected resume text underwent a cleaning process to improve the precision of subsequent natural language processing operations. Any special characters, punctuation, and superfluous spaces had to be eliminated. Data balancing strategies were used in cases where the acquired dataset showed class imbalance, with particular resume sections having

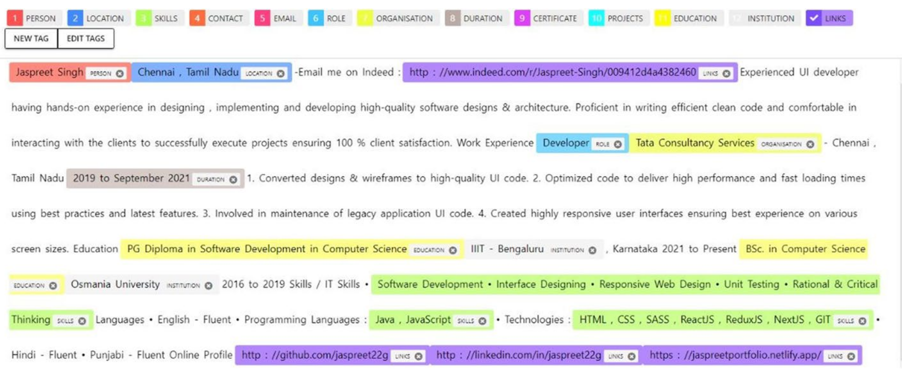


Fig. 1 Annotation process of a sample resume for the Named Entity Recognition using DistilBERT Model

noticeably fewer instances than others. To make sure that each area of the resume was sufficiently represented during training, this included either oversampling the minority courses or under-sampling the majority classes. Some of the resumes were empty or incomplete; some had incorrect information, and in many such cases, these resumes had to be dropped to maintain the consistency of the dataset.

4.3 Addressing input data quality concerns

The development of our Resume Parser and Ranker is confronted with a critical challenge revolving around the quality of input data. While computer vision techniques, exemplified by the YOLOv5 model, showcase impressive accuracy in identifying and classifying resume sections, the system's efficacy is heavily contingent upon the precision and comprehensiveness of the submitted resume data. Concerns arise, particularly from potential errors or deviations in input data, which could significantly impact the system's overall performance. Of notable concern is the impact of incorrect formatting on data precision and effectiveness, as unique bounding box assumptions and section classifications may falter with poorly formatted resumes. To address these concerns, robust pre-processing procedures are imperative to accommodate diverse resume formats and mitigate extraction inaccuracies. Additionally, ongoing initiatives aim to counter potential biases, incorporate diverse resume formats, and regularly update training datasets to enhance the system's adaptability and resilience to varying input qualities and formats, ensuring its applicability across a wide spectrum of real-world scenarios.

5 Methodology

This section provides a brief description of the proposed methodology for the Resume Parser and Ranker based on computer vision and DistilBERT [32] architecture (Fig. 2).

5.1 Text extraction

Our approach consists of multiple steps that leverage computer vision and Natural Language text extraction techniques to accurately extract text from different sections of the resume. The main processes in the extraction process are briefly described below:

1. Bounding Box Estimation and Tag Prediction: In this process, we use a computer vision object detection model to identify and enclose different sections of the text in bounding boxes. This allows us to isolate the text extraction and avoid interference with the other data in the Named Entity Recognition process. We also predict the probable tag or category for each section of the text using the contextual and visual features of the text.
2. Text Extraction from Bounding Boxes: In this process, we extract all the text from the bounding boxes and store them separately for further processing.
3. Section Classification: In this process, we classify the extracted text according to the section of the resume that it belongs to.

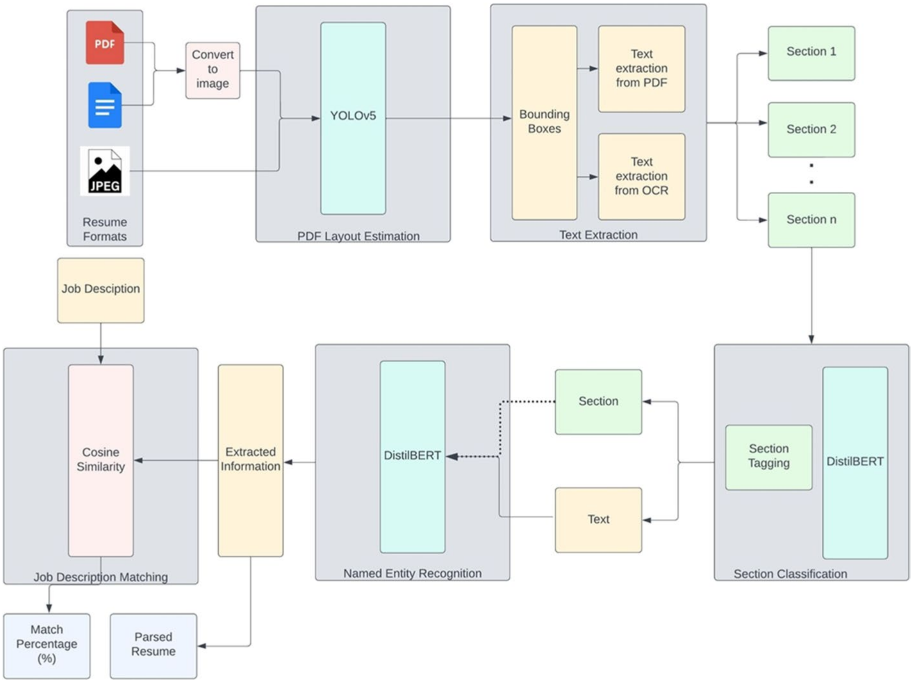


Fig. 2 General Architecture depicting the overall pipeline that showcases the steps involved in the resume parser

4. Named Entity Recognition: In this process, we extract all the relevant information from the text segments using a Named Entity Recognition model. This ensures that the most accurate information is displayed.
5. Matching: In this process, we match the extracted information of the candidate with the job description of the company using a similarity measure. This helps us find the best match.

5.1.1 Bounding box generation & tag estimation

The extraction process begins by enclosing different sections of the resume in bounding boxes. We trained the YOLOv5 model on an opensource dataset that consisted of resumes with annotated sections and tags. The model was able to accurately identify the various relevant sections in the resume, such as personal information, education, skills, work experience, projects, and certificates. These sections helped us to divide the resumes into separate segments, each of which was processed individually.

The YOLOv5 [19] model also estimated the probable tag or category for each piece of the resume along with the bounding boxes. The model gave an estimated tag to each segment by considering the contextual data within the bounding boxes, providing a preliminary classification of that section. The YOLOv5 framework is based on deep learning and builds on the YOLO (You Only Look Once) family of models. It uses a single neural network to detect objects in an image, providing bounding box coordinates and confidence scores. YOLOv5 is based on CSP-Darknet53 [6], which is a convolutional neural network (CNN) backbone that uses cross-stage partial (CSP) connections to reduce the number of parameters and increase the speed of the model. YOLOv5 also uses a new neck structure called PANet (Path Aggregation Network), which enhances the feature fusion between different layers of the network. YOLOv5 adopts a new head structure called SPP (Spatial Pyramid Pooling), which enlarges the receptive field and improves the performance of the model. It uses a very specific type of annotation technique that involves providing not only the class to which a specific entity belongs but also the exact coordinates of the mask that bound the object in question. An annotation looks like:

(c, x1, y1, x2, y2, . . . , xn, yn), where c is the class to which it belongs and the numbers following it will be the coordinates of the bounding box.

We used the trained YOLOv5 model to estimate the bounding boxes around the key sections of the resumes. For this, the model required images of the resumes as input. Convolutional neural networks (CNNs) were then used to extract the visual features and predict the bounding box coordinates. The YOLOv5 model also classified the estimated sections into specific categories besides bounding box estimation. The model assigned an estimated tag or category to each section by integrating the textual data within the bounding boxes with the visual features obtained by the CNNs. The model achieved an accuracy of 96.3% in correctly drawing the bounding boxes for the sections of the resume (Fig. 3).

5.1.2 Text extraction

We superimposed the bounding boxes that we created in the previous step on the input resume pdf pages. These bounding boxes helped us split the pdf into smaller pieces for

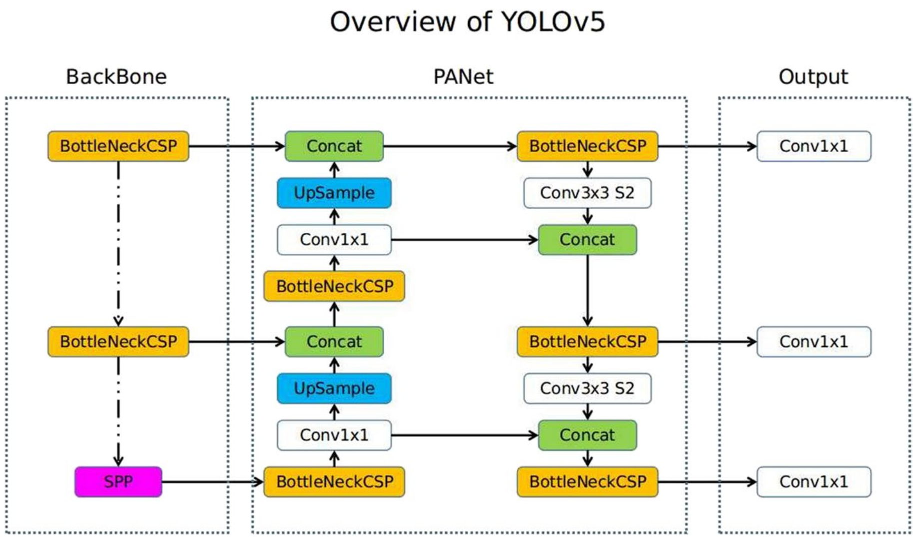


Fig. 3 Yolov5's network architecture. There are three components to it: CSPDarknet for the backbone, PANet for the neck, and Yolo Layer for the head. Before being sent to PANet for feature fusion, the data is first fed to CSPDarknet for feature extraction. Yolo Layer then outputs the results of the detection (class, score, position, and size)

text extraction. This method was applied to isolate the information that was inputted into the Named Entity recognizer. This isolation helps the model to concentrate on one piece of information that is semantically coherent and distinct from the other parts. We used the PyMuPDF Python module to extract the text. PyMuPDF enables us to extract text from text boxes by providing the coordinates, height, and width of the box. We obtained this information from the bounding boxes that we generated earlier.

5.2 Section classification

In this phase, we create a model for performing Section Classification. For this, we fine-tune the DistilBERT model on the dataset that we created for section categorization (Fig. 4). The dataset consists of resume excerpts from various categories, such as personal information, education, skills, experience, projects, certifications, and more. We also include the tags that correspond to each category, such as PERSON, LOCATION, SKILLS, CONTACT, EMAIL, ROLE, ORGANIZATION, DURATION, CERTIFICATE, PROJECTS, EDUCATION, INSTITUTION, and LINKS. These tags are used to label the resume excerpts and train the section tag classifier.

The section tag classifier is a DistilBERT model that has an encoder-decoder architecture. The encoder captures the contextual representations of the input text using a transformer-based model that employs self-attention mechanisms. The self-attention mechanisms allow the model to weigh the relevance of each token or word in relation to other tokens in the text. This enables the model to extract rich contextual information from the text. The decoder generates the output tags for each input text using a linear layer and a softmax function.

We use the section tag classifier to assign tags to the text that we extracted from the bounding boxes in the previous step. The section tag classifier categorizes the resume

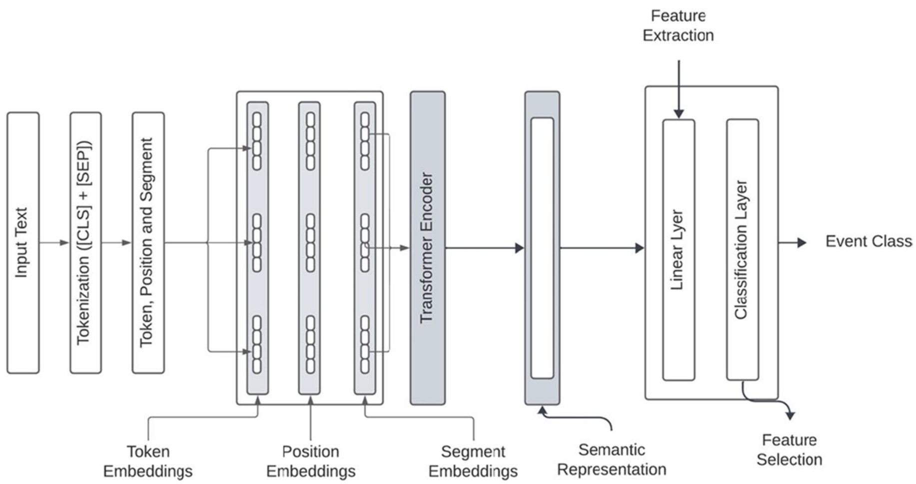


Fig. 4 Overview of the DistilBERT architecture that has facilitated the Named Entity Recognition on Resumes. The first section is the lexicon encoder and the input encoder that will encode the text for use as input in the transformer encoder attention head. The next is the semantic encoder, which will encode and preserve the semantics when generating new text using the next section. The final section is the feature extraction layer, which selects the best features after fine-tuning Table 2

Table 2 Training Parameters for DistilBERT

Learning Rate	2e-05
Training Batch Size	8
Evaluation Batch Size	8
Seed	42
Optimizer and epsilon	Adam with betas = (0.9, 0.999) = 1e-8
Number of Epochs	20

excerpts into different sections and assigns them the appropriate tags. This step is added to the standard resume parsing procedure to reduce the semantic ambiguity that would arise from using the whole document's text alone. The section tag classifier achieves an accuracy of 96.21%, which shows how well it can tag the resume excerpts correctly (Fig. 5).

5.3 Named entity recognition

The next step was entity extraction in the sectional data. We applied the DistilBERT model for Named Entity Recognition in the sectional data. We trained a different model for this task, using a dataset that we created on our own. The dataset contained labeled text segments with thirteen entities that corresponded to the tags that we used in the section classification step, such as PERSON, LOCATION, SKILLS, CONTACT, EMAIL, ROLE, ORGANIZATION, DURATION, CERTIFICATE, PROJECTS, EDUCATION, INSTITUTION, and LINKS.

All this data was then converted to a form that is accepted for training the DistilBERT model using the SpaCy library. Using the custom transformer parameters from the library, we were able to train the model using a GPU. The DistilBERT model

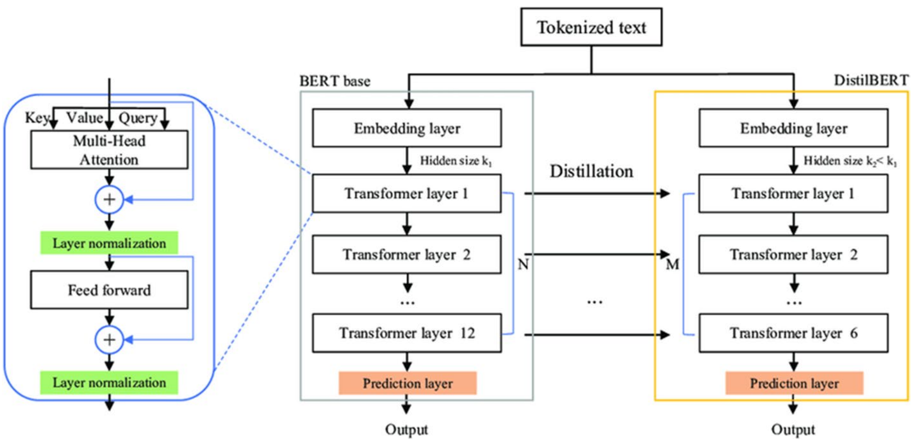


Fig. 5 The DistilBERT model architecture and components. The model uses the concept of distillation, wherein the data from the training is used to train the model, and only the updated weights are taken and the knowledge is transferred to a smaller and efficient model. Which means that a large BERT model is taken and distilled into a smaller and efficient DistilBERT model

learned to identify and extract these specific entities or categories from the input text. The DistilBERT model followed a different process for Named Entity Recognition. First, the input text was tokenized into discrete tokens or sub-word units using WordPiece tokenization, which handled out-of-vocabulary terms and improved generalization. Then, the tokenized text was embedded into high-dimensional vector representations using an embedding layer. Next, the DistilBERT models processed the tokens using transformer layers, which grasped the contextual relationships between the tokens in the input sequence through self-attention mechanisms. The DistilBERT models also used a distillation technique to transfer knowledge from a larger pre-trained model, such as BERT, during training. This reduced the size and complexity of the DistilBERT models. Finally, the DistilBERT models used a classification head that consisted of additional layers that were specifically designed for the Named Entity Recognition task.

5.4 Matching with job description

By evaluating the degree of alignment between the resumes and the given Job Description, the Resume Parser is essential in expediting the resume screening procedure. This matching percentage serves as a valuable metric for ranking and sorting resumes, enabling the identification of the most qualified candidates for the position. We use cosine similarity to achieve this. The semantic similarity between the vectors representing the job description and the vectors representing the resume is calculated using this method. The approach assesses the degree of similarity by measuring the cosine of the angle between these vectors. When the similarity is higher, the vectors are more aligned, and when it is lower, there is a greater angular divergence between them. A greater similarity score on a resume indicates a closer match with the essential education, training, and experience listed in the job description.

6 Results

When considering the individual accuracies of each class, the Precision-Confidence curve for the YOLOv5 computer vision model reveals that the model has the best overall accuracy attainable. The model has been trained for 50 epochs using the stochastic gradient descent (SGD) optimizer before it reached the best accuracy it currently has (96.3%).

The accuracy of the Yolov5 model as stated above is 96.3%. Fig. 6 shows the confusion matrix for the 14 classes that were used in the Bounding Box creation using the YOLO V4 model.

The confusion matrix above demonstrates that the model can accurately classify and extract tags such as CONTACT, EXPERIENCE, NAME, PROFILE, IMAGE and SKILLS from the resume. These tags represent the information that the model can recognize and categorize from the resume, such as the contact information, the work experience, the name of the candidate, the profile summary, the image of the candidate, and the skills of the candidate. The model has a high number of true positives and a low number of false positives and false negatives for these tags, which means that the model can correctly identify and label these tags most of the time and rarely miss or mislabel them. The reason why the model can learn these tags well is that these tags appear frequently in the PDFs of the resumes, which provide the model with enough data and examples to train and test on. The

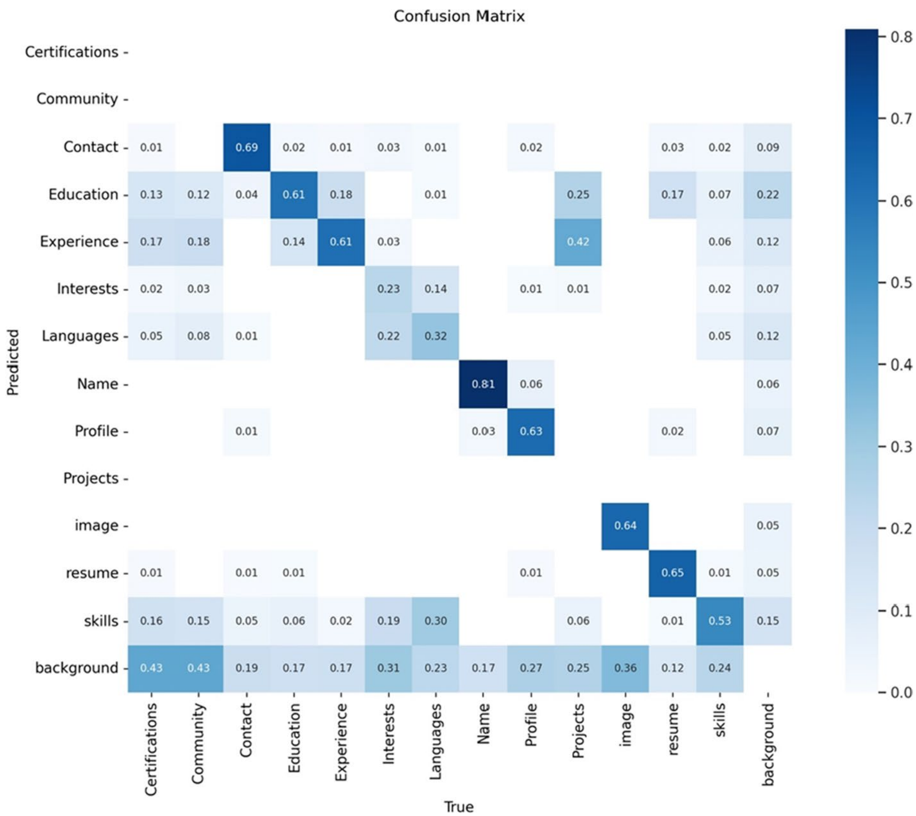


Fig. 6 Confusion matrix for the 14 classes used in the Bounding Box creation for the YOLOv5 Model

model can capture the patterns and features of these tags from the PDFs of the resumes and use them to classify and extract the information from the resumes.

Based on the provided graphs, the model exhibits remarkable accuracy in nearly all of the classes within the resumes (Fig. 7). The open-source dataset does, however, have comparatively less examples for other variables, including image, interest, projects, community, and certifications. Expanding the dataset size will help to effectively address the model's potential for greater specialization in these particular tags.

Enabling the extraction of pertinent text that can then be further categorized using the section classifier model, the model successfully identifies the important regions of interest in resumes. This strategy improves overall accuracy and makes it easier to categorize resume sections precisely Table 3.

In order to reliably identify the sections of resumes, the section classifier model was put into use, enhancing the capabilities of the computer vision model. The results shown above clearly demonstrate how effective the section classifier model is in identifying the proper section for a given input text. This effective performance fills in the holes the computer vision model left by offering a complete answer for section identification in resumes.

The loss graph (Fig. 8) illustrates the learning curve of the DistilBERT Named Entity Recognition Model, which is a natural language processing model that can recognize and categorize named entities in a text, such as persons, locations, organizations, dates, etc.

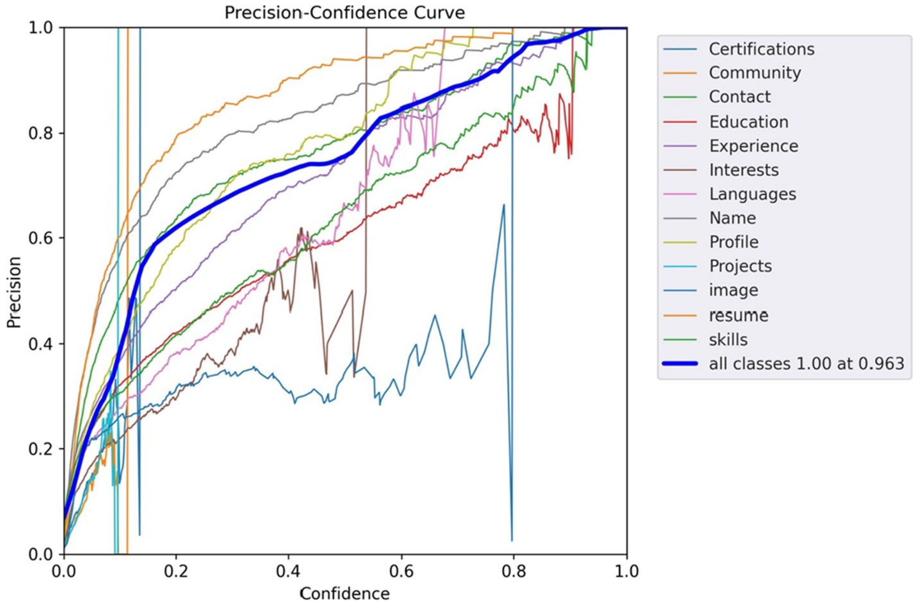
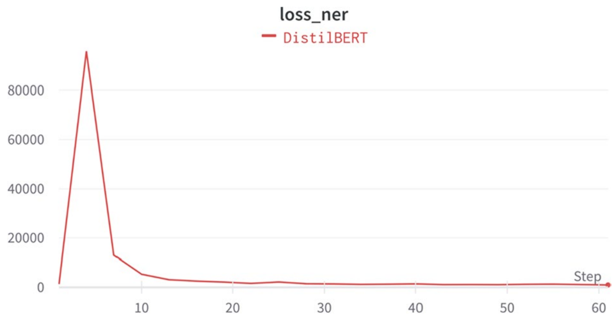


Fig. 7 Precision Confidence Curves for the YoloV5 model that was used for creating the bounding boxes and identifying the various sections

Table 3 Results of DistilBERT for section classification

Loss	F1	ROC	Accuracy
0.0369	0.9652	0.9808	0.9621

Fig. 8 Loss graphs for the DistilBERT model for Named Entity Recognition



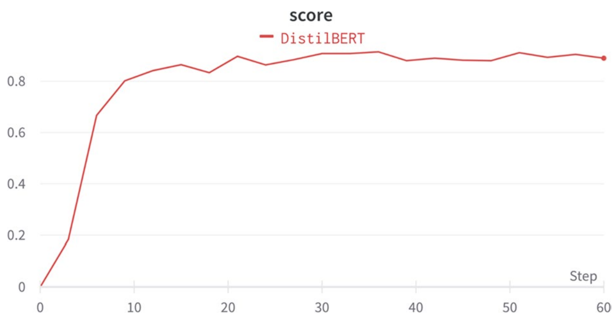
The loss graph shows how the loss function of the model changed over the training and validation steps, which are the cycles of the learning process. The loss function of the model indicates the error or discrepancy between the predicted output and the actual output of the model. The loss graph has two curves, one for the training loss and one for the validation loss. The training loss is the loss function computed on the training data, which is the data that the model learns from. The validation loss is the loss function computed on the validation data, which is the data that the model is tested on. The x-axis of the loss graph represents the number of steps, and the y-axis represents the value of the loss function. The loss graph shows that the model had a very high loss value at the beginning of the training, which means that the model had a poor performance and made many errors in its predictions. However, the loss value decreased rapidly after approximately 10 steps, which means that the model improved its performance and made fewer errors in its predictions. The loss value continued to decrease until the end of 20 steps, which means that the model reached a good performance and made accurate predictions. The loss graph suggests that the model learned and converged well over the training and validation steps, and achieved a high accuracy in the Named Entity Recognition task.

The accuracy graph (Fig. 9) demonstrates the performance of the DistilBERT Named Entity Recognition Model, which is a natural language processing model that can identify and classify named entities in a text, such as persons, locations, organizations, dates, etc. The accuracy graph shows how the accuracy score of the model changed over the training and validation steps, which are the iterations of the learning process. The accuracy score of the model measures how well the model can correctly predict or classify the output. The accuracy graph has two curves, one for the training accuracy and one for the validation accuracy. The training accuracy is the accuracy score computed on the training data, which is the data that the model learns from. The validation accuracy is the accuracy score computed on the validation data, which is the data that the model is tested on. The x-axis of the accuracy graph represents the number of steps, and the y-axis represents the value of the accuracy score. The accuracy graph shows that the model had a very low accuracy score at the beginning of the training, which means that the model had a poor performance and made many errors in its predictions. However, the accuracy score increased rapidly after approximately 10 steps, which means that the model improved its performance and made fewer errors in its predictions. The accuracy score continued to increase until the end of 20 steps, which means that the model reached a good performance and made accurate predictions. The accuracy graph indicates that the model learned and converged well over the training and validation steps, and achieved a high accuracy in the Named Entity Recognition task (Fig. 10).

7 Conclusion

In this paper, we presented a novel architecture to improve all the aspects of a traditional resume parser right from the extraction of text which was based on a variety of different mechanisms such as using YoloV5 for creating bounding boxes to demarcate the sections properly and later extracting the text from the same along with tag

Fig. 9 Average Accuracy Score For the DistilBERT model for the Named Entity Recognition Task



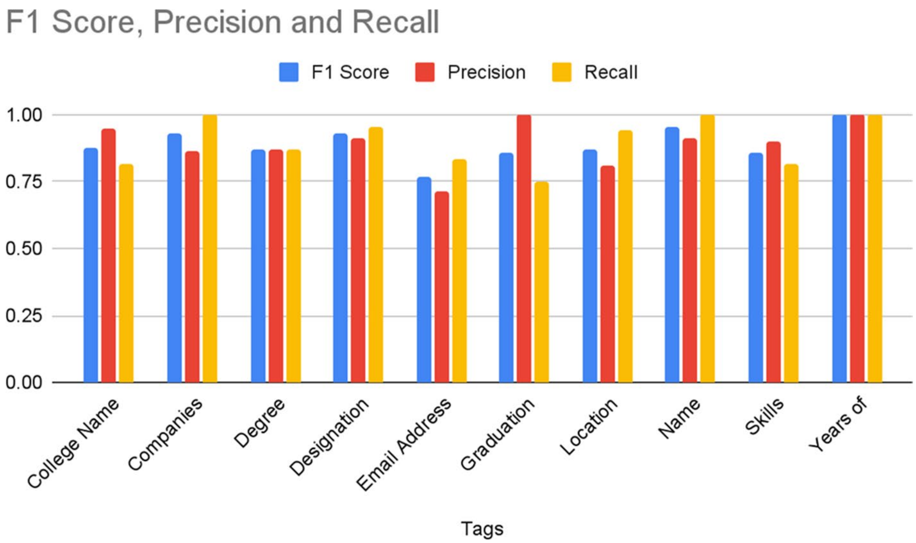


Fig. 10 F1 score, Precision and Accuracy for individual tags used to train the DistilBERT Named Entity Recognition model

estimation to understand this section belongs to which tag. This ensures that all the text in one section is meaningful and has the same context. Later on extracting all the important tags (Named Entity Recognition) from a section with the help of DistilBERT. Every part of the process is improved and trained in such a way to provide the maximum accuracy possible. This would significantly aid recruiters by streamlining the hiring process, enabling them to efficiently identify relevant candidate information and make informed decisions with enhanced accuracy Table 4.

8 Future work

A strategy that involves creating transcripts from the video footage and training a deep neural network to extract pertinent information from those transcripts can be developed to improve the process of parsing video resumes. Incorporating a single deep neural

Table 4 Tabular Representation for the F1 score, Precision and Recall for all the individual tags used to train the DistilBERT model.

Tags	F1 Score	Precision	Recall
College Name	0.88	0.95	0.82
Companies worked	0.93	0.86	1.00
Degree	0.87	0.87	0.87
Designation	0.93	0.91	0.95
Email Address	0.77	0.71	0.83
Graduation Year	0.86	1.00	0.75
Location	0.87	0.81	0.94
Name	0.95	0.91	1.00
Skills	0.86	0.90	0.82
Years of Experience	1.00	1.00	1.00

network to handle numerous tasks, such as making bounding boxes and extracting pertinent tags, can also streamline the current approach. This unified strategy has a number of benefits, including increased productivity and smooth job integration. The model is capable of conducting an efficient analysis of video information and deriving insightful knowledge from it by utilizing the power of deep learning.

Declarations

Competing interests The authors declare that they have no confict of interest.

References

1. Ali I, Mughal N, Khan ZH, et al (2022) Resume Classification System using Natural Language Processing and Machine Learning Techniques. Mehran University Research Journal of Engineering and Technology 41(1):65–79. https://doi.org/10.22581/muet1982.2201.07
2. Ayishathahira CH, 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). IEEE. https://doi.org/10.1109/cetic4.2018.8530883
3. Babu MBVS, Bharath R, Parvez S, et al (2022) Resume Parser using Natural Language Processing. International Journal of Advanced Research in Science, Communication and Technology pp 161–167. https://doi.org/10.48175/ijarsct-7616
4. Bhatia V, Rawat P, Kumar A, et al (2019) End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT. 10.48550/ ARXIV.1910.03089
5. Bhatti SUR, Shaiq MD, Sajid H, et al (2022) Resumizer: Hybrid Resume Information Retrieval System. In: 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM), pp 1–7. https://doi.org/10.1109/IMCOM53663.2022.9721722
6. Bochkovskiy A, Wang CY, Liao HYM (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. https://doi.org/10.48550/ARXIV.2004.10934
7. Chou YC, Chao CY, Yu HY (2019) A R´esum´e Evaluation System Based on Text Mining. In: 2019 International Conference on Artificial Intelligence in Informa- tion and Communication (ICAIIC). IEEE. https://doi.org/10.1109/icaiic.2019.8669066
8. Clark K, Luong MT, Le QV, et al (2020) Pre-training text encoders as discriminators rather than generators. arXiv. https://doi.org/10.48550/ARXIV.2003.10555
9. Deepak G, Teja V, Santhanavijayan A (2020) A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm. J Dis- crete Mathematical Sci Cryptography 23(1):157–165. https://doi.org/10.1080/09720529.2020.1721879
10. Devlin J, Chang MW, Lee K, et al (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1810.04805
11. Gaur B, Saluja GS, Sivakumar HB et al (2020) Semi-supervised deep learning based named entity recognition model to parse education section of resumes. Neural Comput Appli 33(11):5705–5718. https://doi.org/10.1007/s00521-020-05351-2
12. Guo S, Alamudun F, Hammond T (2016) R´esuMatcher: A personalized r´esum´e- job matching system. Expert Syst Appli 60:169–182. https://doi.org/10.1016/j.eswa.2016.04.013
13. Habous A, Nfaoui EH (2021) Combining Word Embeddings and Deep Neural Networks for Job Offers and Resumes Classification in IT Recruitment Domain. Int J Adv Comput Sci Appli 12(7). https://doi.org/10.14569/IJACSA.2021.0120774
14. He P, Liu X, Gao J, et al (2020) DeBERTa: Decoding-enhanced BERT with Disentangled Attention (Version 6). arXiv. https://doi.org/10.48550/ARXIV.2006.03654
15. He P, Liu X, Gao J, et al (2020) "DeBERTa: Decoding-enhanced BERT with Disentangled Attention." *ArXiv* abs/2006.03654 (2020): n. pag.
16. He P, Gao J, Chen W (2021) DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. ArXiv
17. James V, Kulkarni A, Agarwal R (2023) Resume Shortlisting and Ranking with Transformers. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35081-88

18. Jiechieu KFF, Tsopze N (2020) Skills prediction based on multi-label resume classification using CNN with model predictions explanation. Neural Comput Appli 33(10):5069–5087. https://doi.org/10.1007/s00521-020-05302-x
19. Jocher G, Chaurasia A, Stoken A, et al (2022) ultralytics/yolov5: v7.0- YOLOv5 SOTA Realtime Instance Segmentation. https://doi.org/10.5281/ZENODO.7347926
20. Kopparapu SK, Vandana Jayaraj, Automated Filtering of Relevant Resumes. Int J Comput Appli. 154, 6 (Nov 2016) 34-36. https://doi.org/10.5120/ijca2016912162
21. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. ArXiv. 2020. abs/1909.11942
22. Lavi D, Medentsiy V, Graus D (2021) conSultantBERT: Fine-tuned Siamese Sentence-BERT for Matching Jobs and Job Seekers. https://doi.org/10.48550/ARXIV.2109.06501
23. Li X, Shu H, Zhai Y, et al (2021) A Method for Resume Information Extraction Using BERT-BiLSTM-CRF. In: 2021 IEEE 21st International Conference on Communication Technology (ICCT). IEEE. https://doi.org/10.1109/icct52962.2021.9657937
24. Liu J, Shen Y, Zhang Y, et al (2021) Resume Parsing based on Multi-label Classification using Neural Network models. In: 2021 6th International Conference on Big Data and Computing. ACM. https://doi.org/10.1145/3469968.3469998
25. Liu Y, Ott M, Goyal N, Du J, Joshi, M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1907.11692
26. Ma L, Gupta RK, Onyema EM (2022) Optimization of Intelligent Network Infor- mation Management System under Big Data and Cloud Computing. Scalable Computing: Practice and Experience 23(3):91–101. https://doi.org/10.12694/scpe.v23i3.2001
27. Mohamed A, Bagawathinathan W, Iqbal U, et al (2018) Smart Talents Recruiter -Resume Ranking and Recommendation System. In: 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS). IEEE. https://doi.org/10.1109/iciafs.2018.8913392
28. Nasser S, Sreejith C, Irshad M (2018) Convolutional Neural Network with Word Embedding Based Approach for Resume Classification. In: 2018 International Conference on Emerging Trends and Innovations In Engineering And Tech- nological Research (ICETIETR). IEEE. https://doi.org/10.1109/icetietr.2018.8529097
29. Onyema EM, Almuzaini KK, Onu FU et al (2022) Prospects and Challenges of Using Machine Learning for Academic Forecasting. Comput Intel Neurosci 2022:1–7. https://doi.org/10.1155/2022/5624475
30. Onyema EM, Gude V, Bhatt A, et al (2023) Smart Job Scheduling Model for Cloud Computing Network Application. SN Computer Science 5(1). https://doi.org/10.1007/s42979-023-02441-5
31. Sajid H, Kanwal J, Bhatti SUR, et al (2022) Resume Parsing Framework for E- recruitment. In: 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM). IEEE. https://doi.org/10.1109/imcom53663.2022.9721762
32. Sanh V, Debut L, Chaumond J and Wolf T (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (Version 4). arXiv. https://doi.org/10.48550/ARXIV.1910.01108
33. Sanyal, Satyaki & Hazra, Souvik & Ghosh, Neelanjan & Adhikary, Soumyashree. (2017). Resume Parser with Natural Language Processing. https://doi.org/10.13140/RG.2.2.11709.05607
34. Schopf T, Braun D, Matthes F (2022) Evaluating Unsupervised Text Classification: Zero-shot and Similarity-based Approaches. Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval
35. Sun Z, Yu H, Song X, Liu R, Yang, Y, Zhou D. (2020). MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2004.02984
36. Tallapragada VVS, Raj VS, Deepak U, et al (2023) Improved Resume Parsing based on Contextual Meaning Extraction using BERT. In: 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE. https://doi.org/10.1109/iciccs56967.2023.10142800
37. Zu S, Wang X (2019) Resume Information Extraction with A Novel Text Block Segmentation Algorithm. Int J Natural Lang Comput 8(5):29–48. https://doi.org/10.5121/ijnlc.2019.8503