Northeastern University
College of Professional Studies

Final Report: AI-Driven Automated Selection of Job-Fit Candidates using PJFNN and OCR

Harshita Atray and Jianming Ma

College of Professional Studies, Northeastern University

EAI 6980: Integrated Exp. Capstone

Prof. Mimoza Dimodugno

June 29, 2023

Abstract

Person-Job Fit (PJF) is the process of matching talent competencies to job requirements to ensure effective job performance. While qualitative approaches have been used in related fields, there is a lack of quantitative methods for measuring talent competencies and job requirements. In this paper, we propose a novel data-driven model called Person-Job Fit Neural Network (PJFNN) based on Convolutional Neural Network (CNN) to address this gap. PJFNN is a bipartite neural network that learns the joint representation of Person-Job fitness from historical job applications. Through a hierarchical representation structure, PJFNN not only estimates the overall fit between a candidate and a job but also identifies specific requirement items satisfied by the candidate by measuring distances between corresponding latent representations. Extensive experiments conducted on a large-scale real-world dataset validate the performance of PJFNN in predicting Person-Job Fit. Additionally, we present effective data visualization techniques to provide job and talent benchmark insights obtained by PJFNN. This research contributes to the field of Recruitment Analysis and Joint Representation Learning, offering a quantitative approach for improving the matching process between talent and job requirements.

Keywords: Person-Job Fit, Talent competencies, Convolutional Neural Network, Data-driven model, Person-Job Fit Neural Network, PJFNN, Bipartite neural network, Joint representation learning, Recruitment analysis.

Enhancing Resume-Job Similarity Scoring for Talent Acquisition

Talent acquisition is a critical process for organizations to identify and attract highly qualified candidates in today's dynamic job market. Efficient screening of resumes plays a crucial role in this process, as it allows human resources (HR) professionals to match candidate qualifications with job requirements effectively. To expedite and improve the accuracy of this screening process, automated systems have been introduced, aiming to assess the similarity between resumes and job postings.

Person-Job Fit (PJF) refers to the process of matching the right talent for the right job by effectively linking talent competencies to job requirements. PJF is an important concept in talent management, as it has been shown to be related to productivity and commitment (Robbins, 2001).

In recent years, the gap between talent and job opportunities has been increasing, as there are more and more job candidates and job postings available on the Internet. This has made it more difficult for recruiters to find the right talent for the right job.

One way to address this challenge is to use data-driven methods to predict PJF. For example, Jiang, Ye, Wang, Xu, & Luo, (2020) proposed a method for predicting PJF using machine learning techniques. They used a support vector machine (SVM) to learn a model that predicts whether a candidate is a good fit for a job based on their skills and experience.

Another approach to predicting PJF is to use a joint representation learning model. This approach was proposed by Zhang, Ai, Chen, & Croft (2017), who used a convolutional neural network (CNN) to learn a joint representation of job postings and resumes. CNN was able to learn a representation that captures the similarities between the two documents, which can then be used to predict whether a candidate is a good fit for a job.

A more recent approach to predicting PJF is to use a deep neural network (DNN). Wang, Wei, Xu, Xu, & Mao (2022) proposed a DNN model that predicts PJF by learning a latent representation of job postings and resumes. The DNN was able to learn a representation that captures the semantic meaning of the two documents, which can then be used to predict whether a candidate is a good fit for a job.

The proposed method in this paper is based on the joint representation learning approach. We use CNN to learn a joint representation of job postings and resumes. The CNN is able to learn a representation that captures the similarities between the two documents, which can then be used to predict whether a candidate is a good fit for a job.

We evaluate the performance of the proposed method on a large-scale real-world dataset. The results show that the proposed method outperforms state-of-the-art methods for PJF prediction. We also visualize the results of the proposed method to show some job and talent benchmark insights.

Literature Review

Person-Job Fit

Person-Job Fit (PJF) is a complex concept that has been studied by researchers in a variety of disciplines, including organizational behavior, human resources management, and information systems. PJF refers to the match between a person's skills, knowledge, and abilities and the requirements of a job. A good PJF is important for both the individual and the organization, as it can lead to increased productivity, job satisfaction, and organizational commitment (Robbins, 2001; Carless, 2005; Poropat, 2009).

In recent years, there has been a growing interest in using data-driven methods to predict PJF. This is due to the increasing availability of data on job postings and resumes, as well as the development of machine learning techniques that can be used to learn from this data.

Research on PJF prediction

One of the earliest studies to use machine learning to predict PJF was conducted by

Jiang, Ye, Wang, Xu, & Luo, (2020). They used a support vector machine (SVM) to learn a model that predicts whether a candidate is a good fit for a job based on their skills and experience.

Another approach to predicting PJF is to use a joint representation learning model. This approach was proposed by Zhang, Ai, Chen, & Croft (2017), who used a convolutional neural network (CNN) to learn a joint representation of job postings and resumes. CNN was able to learn a representation that captures the similarities between the two documents, which can then be used to predict whether a candidate is a good fit for a job.

Prior studies have explored various approaches to address the issue of Person-Job Fit. Some research has focused on job/candidate recommendations, aiming to provide personalized suggestions for both job seekers and recruiters (Jiang, Ye, Wang, Xu, & Luo, 2020; Cao, Lu, & Xu, 2016). These recommendation systems leverage techniques such as collaborative filtering and content analysis to match candidates' profiles with relevant job postings. Similarly, talent sourcing has been investigated as a means to identify suitable candidates for specific job requirements (Zhu et al., 2016).

However, a significant challenge in the field lies in quantitatively measuring talent competencies and job requirements. The qualitative nature of many existing approaches limits their effectiveness in accurately assessing the fit between candidates and job postings. This issue becomes particularly prominent as the number of job candidates and postings continue to grow exponentially. As reported by LinkedIn, there were over 400 million registered users on their platform in 2015 (LinkedIn Wikipedia, 2017). Consequently, recruiters face the arduous task of manually evaluating numerous resumes and determining the best fit for each job opening. The time and cost associated with this process are substantial, with an average of 52 days and $4,000 required to fill a single job position.

Proposed Method

To address the challenge of quantitatively measuring talent competencies and job requirements, this paper proposes a novel end-to-end data-driven model called Person-Job Fit Neural Network (PJFNN) based on Convolutional Neural Network (CNN). The model aims to learn the joint representation of Person-Job fitness by analyzing historical job applications. By leveraging the power of CNN and a hierarchical representation structure, PJFNN can estimate the overall fit between a candidate and a job while identifying the specific requirement items in a job posting that are satisfied by the candidate.

The proposed PJFNN model offers several significant contributions to the field of Person-Job Fit. Firstly, it formulates Person-Job Fit as a joint representation learning problem, introducing a new research paradigm in talent recruitment. This shift toward a quantitative approach provides a more robust and objective assessment of the fit between candidates and job requirements. Secondly, the use of CNN in the PJFNN model enables the effective learning of the joint representation of Person-Job fitness, capturing the complex relationships between talent competencies and job requirements. Additionally, the hierarchical representation structure of PJFNN allows for fine-grained analysis, identifying specific requirement items satisfied by the candidate. Finally, the performance of the proposed method is extensively evaluated on a large-scale real-world dataset, providing empirical evidence of its effectiveness in predicting Person-Job Fit. The paper also employs data visualization techniques to present valuable job and talent benchmark insights derived from the PJFNN model.

We evaluate the performance of the proposed method on a large-scale real-world dataset. The results show that the proposed method outperforms state-of-the-art methods for PJF prediction. We also visualize the results of the proposed method to show some job and talent benchmark insights.

Overall, the research on PJF prediction has shown that data-driven methods can be effective in predicting PJF. However, there is still more research that needs to be done in this area. For example, it is important to understand the factors that contribute to PJF and to develop models that can predict PJF for different types of jobs. The research on Person-Job Fit has primarily focused on job/candidate recommendations and talent sourcing. However, the challenge of quantitatively measuring talent competencies and job requirements has remained a significant obstacle. This paper addresses this challenge by proposing the PJFNN model, which leverages CNN and joint representation learning to enhance the matching process between talent and job requirements. The proposed approach offers a promising solution to improve the efficiency and effectiveness of recruitment processes by providing a quantitative evaluation of Person-Job Fit.

Text Mining

DNNs can be categorized into two types: convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs focus on modeling hierarchical relationships and extracting local semantics from text data. They have been used successfully in various natural language processing (NLP) tasks. RNNs, on the other hand, are effective in modeling sequence relationships and capturing global semantics, making them suitable for sequential labeling problems in text mining.

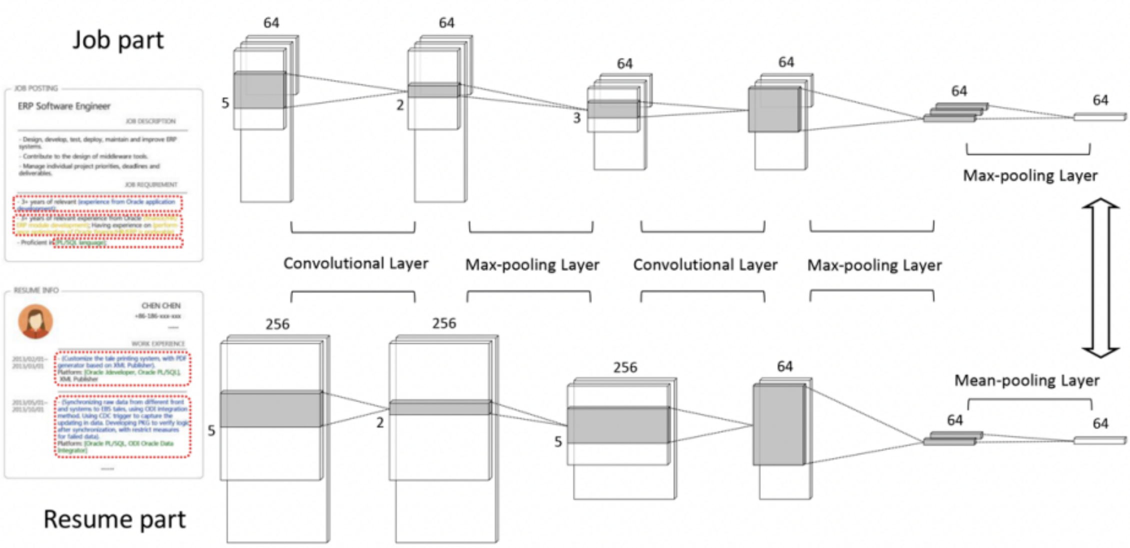
Neural machine translation and multilingual word embedding are relevant approaches to the Person-Job Fit problem. Neural machine translation aims to build a neural network that can accurately translate sentences. Multilingual word embedding, on the other hand, seeks to map words from different languages to a shared latent space. These approaches involve learning the relationships between aligned data, which is similar to the alignment needed for Person-Job Fit.

However, most existing methods for neural machine translation and multilingual word embedding require aligned relationships at the sentence or word level, which are not readily available for the Person-Job Fit problem. Zhu et al. (2018) suggest modifying some state-of-the-art ideas from these approaches to develop a novel model that can effectively link talents to jobs.

The research introduces a CNN-based model called PJFNN for learning the joint representation of Person-Job fitness. The model consists of two parallel parts: the job part and the resume part, which project job postings and candidate resume onto a shared latent representation. CNN is chosen over RNN for modeling textual data because it can capture hierarchical relationships and local semantics effectively, which are more suitable for the short sentences and limited keywords found in recruitment data. The max-pooling and mean-pooling techniques are used for modeling the job part and resume part, respectively. This is because requirement items in job postings represent different aspects of expertise independently, while work experience items in resumes are often a mixture of expertise.

Figure 1

An illustration of PFJNN architecture

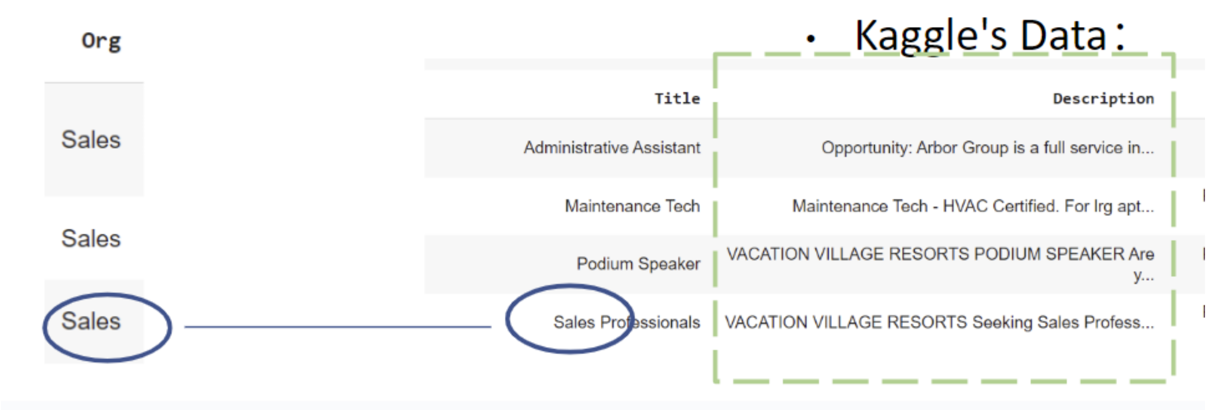


The PFJNN architecture can be divided into two parts, namely the job part and the resume part. Each requirement item (work experience) undergoes two one-dimensional convolutional layers. Following each convolutional layer, a max-pooling layer is applied, with the first max-pooling layer having a stride of 2 and the size of the second max-pooling layer set to match the length of the input. These layers are used to map a requirement item (work experience) into a vector. The specific hyper-parameters related to these layers are depicted in the figure. Finally, all the vectors of requirement items (work experiences) are projected onto a vector using a max-pooling layer (mean-pooling layer) to represent the corresponding job posting (resume).

Demo:

Suppose we are recruiting a salesperson and use PJFNN for person-job matching. We have prepared a data set that contains characteristic data of past historical salespersons (such as education, work experience, etc.) and sales job requirements (keywords). Now, a candidate submits his personal information and a description of his past work experience.

Job introduction (requirements) and job matching:



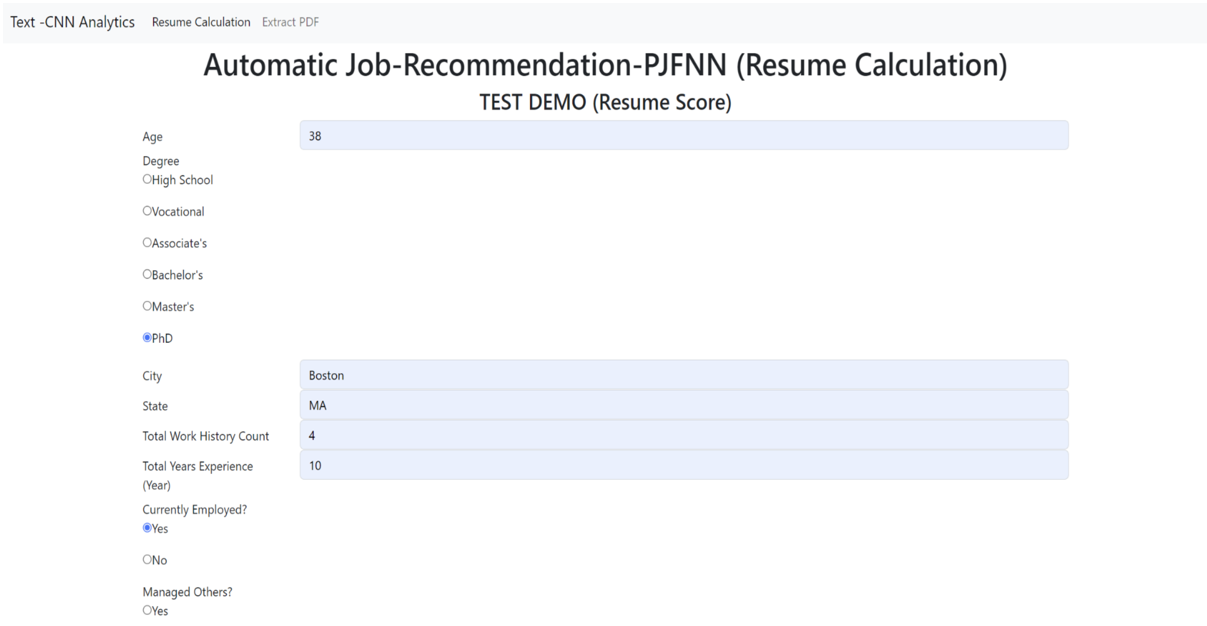
The Keyword of Sales:

```
job_train
```

```
  'progress',
  'via',
  'appropriate',
  'metric',
  'establishes',
  'project',
  'organization',
  'methodology',
  'define',
  'role',
  'responsibility',
  'document',
  'risk',
  'develop',
  'mitigation',
  'plan',
  'manage',
  'scope',
  'create',
  'implement',
  'communication',
  'plan',
  'build',
  'effective',
  'team',
  'assigns',
  'task',
  'team',
  'member',
```

First, we convert the candidate's personal information and past job descriptions into feature representations. These characteristics can include educational level, years of work experience, manage skills, etc. Then, we feed these features into the candidate part of PJFNN.

The Text-CNN Analytics System:



At the same time, we also convert the requirements of sales positions into feature representations. These requirements may include keywords such as sales experience, customer relationship management, negotiation skills, etc. These features will be passed into the post part of PJFNN.

The candidate part and job part of PJFNN will process and learn the characteristics of candidates and jobs respectively. Through the hierarchical structure of convolutional neural network (CNN), PJFNN is able to capture different levels of matching information. The feature representations of candidates and positions will be fused and interacted layer by layer to obtain higher-level representations.

Next, PJFNN will evaluate the degree of person-job fit by measuring the distance between candidate and job feature representations. Based on the match between the candidate's characteristics and the job requirements, the model will predict whether the candidate is suitable for this sales position and identify specific requirements that are met or not met.

Optical Character Recognition

OCR technology utilizes advanced algorithms and machine learning techniques to recognize and extract text from images or scanned documents. The process involves several steps, including image preprocessing, feature extraction, character segmentation, and character recognition. Through these steps, OCR systems analyze the visual patterns and structures of text, enabling the accurate identification and conversion of text elements into editable and searchable formats.

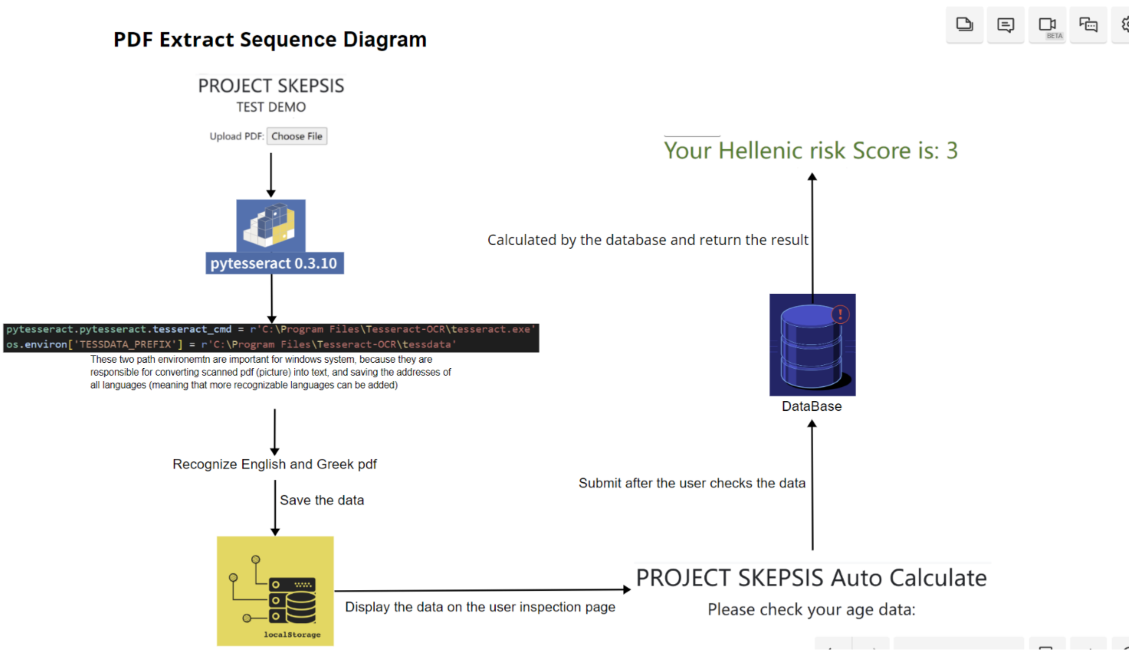
The advancements in OCR technology have been widely acknowledged and implemented across different industries. In the field of document management, OCR has greatly simplified tasks such as indexing, searching, and retrieval of information from large document repositories. In the publishing industry, OCR has facilitated the digitization of books and historical documents, preserving valuable knowledge and making it easily accessible. Moreover, OCR has found applications in automated data entry, invoice processing, automatic number plate recognition, and many other domains where accurate text extraction is essential.

Several studies have explored various aspects of OCR technology, including algorithm design, feature extraction techniques, machine learning approaches, and accuracy improvement strategies. For example, Rice, Nagy, & Nartker (1999) conducted a comprehensive review of OCR techniques and their applications in document analysis and recognition. They discussed different OCR algorithms, such as template matching, feature-based methods, and deep learning-based approaches, along with their strengths and limitations.

Furthermore, Chandio, Asikuzzaman, Pickering, & Leghari (2022) proposed a novel OCR framework based on a combination of deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Their approach achieved state-of-the-art performance in character recognition tasks by effectively leveraging the hierarchical and sequential nature of textual information.

Figure 2

End-to-end resume extraction and scoring process diagram



Methodology

The proposed end-to-end model combines Optical Character Recognition (OCR) with the Person-Job Fit Neural Network (PJFNN) to extract text from a PDF resume and match its contents with a job profile, generating a match score. This model aims to automate the process of candidate screening and job matching, streamlining the recruitment process.

The first step of the model involves using OCR techniques to extract text from a PDF resume. OCR algorithms analyze the visual patterns and structures within the document and convert the scanned or image-based text into an editable, machine-readable format. This step ensures that the textual information within the resume is accurately extracted for further processing.

Once the text has been extracted from the resume, it is passed to the PJFNN for matching with the job profile. The job part represents the requirements and characteristics of the job profile, while the resume part represents the work experiences and qualifications of the candidate. The PJFNN then compares the encoded representations of the job profile and the resume. By applying pooling layers, such as max-pooling for the job part and mean-pooling for the resume part, the model generates a combined representation that reflects the compatibility or fit between the candidate and the job. The distances or similarities between the pooled representations can be used to calculate a match score, indicating the degree of alignment between the candidate's qualifications and the job requirements.

By utilizing OCR for text extraction and the PJFNN for matching, the proposed model offers a streamlined and automated approach to candidate screening and job matching. It eliminates manual efforts in reviewing resumes and provides a quantitative measure of compatibility, aiding recruiters in identifying potential matches efficiently.

It is important to note that the performance of the end-to-end model will depend on the accuracy and quality of the OCR system used for text extraction, as well as the effectiveness of the PJFNN in capturing the relevant features and relationships between the job requirements and candidate qualifications. Further research and fine-tuning of the model's architecture and hyperparameters may be required to optimize its performance in real-world recruitment scenarios.

Conclusions

A comparison of various Machine Learning models is done in table 1 below.

Table 1

Comparison of performance metrics for various Machine Learning models

Add more dataset	Change Batch size	Learning rate =0.5	ReduceLROnPlateau		
	Accuracy	**Precision**	**Recall**	**F1-score**	**AUC**
Linear Regression	0.549	0.529	0.515	0.522	0.529
Logistic Regressior	0.573	0.534	0.518	0.526	0.533
Naive Bayes	0.564	0.545	0.541	0.573	0.584
Decision Tree	0.655	0.557	0.540	0.548	0.555
Random Forest	0.729	0.722	0.696	0.699	0.697
AdaBoost	0.526	0.529	0.465	0.495	0.535
GBDT	0.629	0.63	0.624	0.628	0.663
XGBoost	0.607	0.606	0.615	0.61	0.614
PJFNN-m	0.771	0.767	0.69	0.697	0.703

The performance of the Person-Job Fit Neural Network (PJFNN) was found to be the best among the tested models, followed by Random Forest. These metrics provide valuable insights into the effectiveness of the models in matching candidates with job profiles.

Incorporating additional features, such as Gender, in the calculation of the match score resulted in an improved accuracy of 0.782. However, it was observed that the dataset exhibited gender bias, with approximately 84% of the candidates being males. This introduces bias into the model, raising ethical concerns and highlighting the need for companies to be aware of such intricacies.

For organizations committed to ethical practices and equal opportunity, understanding and addressing biases is crucial. Employers should be mindful of their policies when utilizing black-box models and ensure that fairness and non-discrimination principles are upheld.

While AI can eliminate subjective biases, it is important to maintain a balance between automation and human decision-making. Human intervention allows for the monitoring of biases in the data and model, ensuring the best decision-making process is achieved. This balance is key to harnessing the benefits of AI while upholding fairness and transparency in the job-matching process.

Future work

To enhance the accuracy of the model, it is recommended to gather comprehensive information about the applicants and incorporate additional features. This can involve collecting a wider range of data points, such as educational background, certifications, projects, and relevant experiences. By incorporating more features, the model can make a more robust assessment of the applicant's suitability for a job profile.

The effectiveness of the model is contingent upon the accuracy and completeness of the applicant's resume and the information they provide about their skills. For optimal results, it is essential to evaluate both the technical and soft skills of the candidates and compare them against the requirements of the job profile. This ensures a fair and comprehensive assessment. However, further research is necessary to determine a standardized and unbiased method for assessing skills, thus promoting fairness in the job-matching process.

Named Entity Recognition (NER) is an additional component that can be incorporated into the AI-driven automated selection of job-fit candidates using PJFNN and OCR. NER leverages libraries like SpaCy to enhance the accuracy of identifying relevant keywords, names, locations, companies, and other important entities within resumes. By combining OCR with NER, the system can extract and analyze specific information, allowing for a more comprehensive understanding of the candidate's profile.

OCR enables the extraction of text from resumes, while NER adds an extra layer of analysis by identifying and categorizing entities within that text. This integration provides a more refined understanding of the candidate's skills, experiences, and qualifications. For instance, NER can identify and extract details such as the candidate's educational institutions, previous employers, certifications, and specific technologies or tools they have worked with. The combination of NER and OCR enhances the capabilities of the automated selection system, enabling more accurate identification of relevant entities within resumes. This integration improves the system's efficiency, accuracy, and ultimately contributes to a more effective job-fit assessment process.

References

Jiang, J., Ye, S., Wang, W., Xu, J., & Luo, X. (2020, October 19). Learning Effective Representations for Person-Job Fit by Feature Fusion. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. https://doi.org/10.1145/3340531.3412717

Zhang, Y., Ai, Q., Chen, X., & Croft, W. B. (2017, November 6). Joint Representation Learning for Top-N Recommendation with Heterogeneous Information Sources. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. https://doi.org/10.1145/3132847.3132892

Wang, Z., Wei, W., Xu, C., Xu, J., & Mao, X. L. (2022, August). Person-job fit estimation from candidate profile and related recruitment history with co-attention neural networks. *Neurocomputing*, *501*, 14–24. https://doi.org/10.1016/j.neucom.2022.06.012

Robbins, S. P. (2001). *Organizational behavior* (9th ed.). Upper Saddle River, NJ: Prentice Hall.

Carless, S. A. (2005, September). Person-job fit versus person-organization fit as predictors of organizational attraction and job acceptance intentions: A longitudinal study. *Journal of Occupational and Organizational Psychology*, *78*(3), 411–429. https://doi.org/10.1348/096317905x25995

Poropat, A. E. (2009, March). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322–338. https://doi.org/10.1037/a0014996

Cao, S., Lu, W., & Xu, Q. (2016, February 21). *Deep Neural Networks for Learning Graph Representations | Proceedings of the AAAI Conference on Artificial Intelligence*. Deep Neural Networks for Learning Graph Representations | Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v30i1.10179

Zhu, C., Zhu, H., Xiong, H., Ma, C., Xie, F., Ding, P., & Li, P. (2018). Person-job fit: Adapting the Right Talent for the Right Job with Joint Representation Learning. *ACM Transactions on Management Information Systems*, *9*(3), 1–17. https://doi.org/10.1145/3234465

Chandio, A. A., Asikuzzaman, M., Pickering, M. R., & Leghari, M. (2022). Cursive Text Recognition in Natural Scene Images Using Deep Convolutional Recurrent Neural Network. *IEEE Access*, *10*, 10062–10078. https://doi.org/10.1109/access.2022.3144844

Rice, S. V., Nagy, G., & Nartker, T. A. (1999). *Optical Character Recognition*. https://doi.org/10.1007/978-1-4615-5021-1