Skill Scanner: Connecting and Supporting Employers, Job Seekers and Educational Institutions with an AI-based Recommendation System

Koen Bothmer and Tim Schlippe

IU International University of Applied Sciences
tim.schlippe@iu.org

Abstract. Usually employers, job seekers and educational institutions use AI in isolation from one another. However, skills are the common ground between these three parties which can be analyzed with the help of AI: (1) Employers want to automatically check which of their required skills are covered by applicants' CVs and know which courses their employees can take to acquire missing skills. (2) Job seekers want to know which skills from job postings are missing in their CV, and which study programs they can take to acquire missing skills. (3) In addition, educational institutions want to make sure that skills required in job postings are covered in their curricula and they want to recommend study programs. Consequently, we investigated several natural language processing techniques to extract, vectorize, cluster and compare skills, thereby connecting and supporting employers, job seekers and educational institutions. Our application *Skill Scanner* uses our best algorithms and outputs statistics and recommendations for all groups. The results of our survey demonstrate that the majority finds that with the help of *Skill Scanner*, processes related to skills are carried out more effectively, faster, fairer, more explainably, and in a more supported manner. 89% of all participants are not averse to apply our recommendation system for their tasks. 67% of job seekers would certainly use it.

Keywords: artificial intelligence in education, upskilling, recommender systems, clustering, natural language processing.

1 Introduction

Access to education is one of people's most important assets and ensuring inclusive and equitable quality education is goal 4 of United Nations' Sustainable Development Goals. This goal should not only refer to general education, but also to specific education in the professional environment. If people have the right education for the professional environment, they have a better chance to get jobs that allow them to have a good life. Unfortunately, there are often still gaps between the skills that are needed in the job market, the skills that job seekers have and the skills that are taught in educational institutions like schools, universities, online platforms, massive open online courses (MOOCs), etc. [1].

To solve this problem, all three players—employers, job seekers[1], and educational institutions—need to be aligned. There are already natural language processing (NLP) approaches to extract text data from job seekers' CVs (curriculum vitae), employers' job postings or educational institutions' learning curricula and give recommendations to one of these players. However, this way all three parties use AI in isolation from one another. For example, [2] present a Word2Vec-based [3] system which informs employers how well job seekers' CVs fit job postings. LinkedIn has a system that recommends employers' jobs to job seekers based on their personal profile [4]. [5] investigate how AI-based recommendations help job seekers find study programs based on their profile. [6] use a combination of knowledge graph and BERT [7] for helping employers find suitable candidates in a corpus of CVs.

Our approach leverages similar NLP methods [8], but it benefits not only one, but all three players involved. Connecting and supporting them all allows the greatest possible exchange of information and satisfies their needs as illustrated in Figure 1:

(1) Employers want to automatically check which of their required skills are covered by applicants' CVs (*Find and Select*) and know which courses their employees can take to acquire missing skills (*Upskill Workforce*).
(2) Job seekers want to know which skills from job postings are missing in their CVs (*Fit to Demand*), and which study programs they can take to acquire missing skills (*Find Program*).
(3) Educational institutions want to ensure that the skills required in job postings are covered in their curricula (*Fit to Demand*), recommend study programs and advise students (*Advise*).

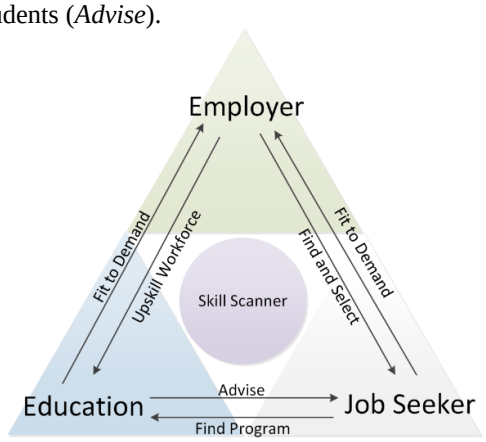


Fig. 1. Connecting and supporting employers, job seekers, and educational institutions.

Since skills are the common ground between these three players, we developed the application *Skill Scanner*[2] which combines NLP techniques to extract, vectorize, clus-

[1] The term "job seeker" refers to current applicants and individuals who wish to advance towards a position.
[2] https://github.com/KoenBothmer/SkillScanner

ter and compare skills in a pipeline and outputs statistics and recommendations for all three players in form of reports. Our goal was to help employers, job seekers and educational institutions adapt to the job market's needs. Consequently, we used job postings, which represent the job market's needs as reference. Our recommendation system determines which skills in the job market's job postings are covered and which skills are missing. These representative skills, which we draw from a large set of job postings, are referred to as "*market skills*" in this paper.

As companies hiring data scientists state that it is increasingly difficult to find a so-called "unicorn data scientist" [9], we conducted our analyses using companies' job postings for a data scientist position, job seekers' CVs for that position, and a curriculum from a master's program in data science. However, our investigated methods and our final recommendation system can be applied to other job positions as well.

Finally, we present our detailed analysis of the feedback from employers, job seekers, and educational institutions on the reports generated with *Skill Scanner*, demonstrating the potential benefits of finding covered and missing skills with the help of our cluster-based algorithms to all three parties.

2 Related Work

In this section we will present the latest approaches of recommendation systems for employers, job seekers, and educational institutions.

2.1 Recommendation Systems for Employers

Automatically ranking CVs is a valuable tool for employers. For example, [10] rank candidates for a job based on semantic matching of skills from LinkedIn profiles and skills from their job description, relying on a taxonomy of skills. They determine the semantic similarity of the skills reported in an applicant's LinkedIn profile to the skills required for a job using the node distance, i.e., the distance to the lowest common ancestor in the taxonomy of skills. Recent NLP techniques offer opportunities to improve these methods: [2] use word embeddings from Word2Vec [3] to match CVs to jobs. [6] combine a knowledge graph and BERT to find suitable candidates in a corpus of CVs.

Our system also works with embeddings—however with sentence embeddings—to vectorize the skills. In addition, we use a cluster approach to find synonymous skills.

2.2 Recommendation Systems for Job Seekers

Recommendation systems for job seekers have been investigated by [11,12,13]. As in the systems for employers, text data from social media profiles such as LinkedIn or Facebook is usually processed [5,14]. Researchers at LinkedIn [15] have built a taxonomy of 35k standardized skills and use semantic matching to measure the similarity in job descriptions and job seekers' profiles.

The benefit of our clustering approach compared to a taxonomy is that our model can pick up new skills without the need to update a taxonomy.

2.3 Recommendation Systems for Educational Institutions

[16] give a systematic review of recent publications on course recommendation. Most related work focuses on recommending courses to potential students. They report a growing popularity of data mining techniques in those systems. To cope with the challenges of different levels of abstraction and synonyms in the course materials and students' documents, some researchers first cluster the content, which they can then compare. K-means is usually used for clustering. To help employers recommend appropriate courses for their employees, [17] suggest a framework called "Demand-aware Collaborative Bayesian Variational Network (DCBVN)".

Compared to the related work, we propose courses for students and employees based on K-means clustering extended with additional steps to detect outliers in the clusters. While the job market is not considered in the recommendation process of other approaches, we use information from employers' job postings—denoted as *market skills* in this paper—as valuable information to enhance our recommendations.

3 Extracting, Vectorizing, Clustering and Comparing Skills

Our goal was to help employers, job seekers, and educational institutions adapt to the *market skills*. Our recommendation system *Skill Scanner* determines which skills in the job postings are covered and which skills are missing. In this section we will describe our pipeline to extract, vectorize, cluster, and compare skills.

5.1 Our Pipeline to Extract, Vectorize, Cluster and Compare Skills

For a certain job position, (1) *Skill Scanner* takes a CV, a job posting or a learning curriculum as input, (2) extracts the skills of the provided document, (3) compares the document's extracted skills to a skill set which represents the market's needs (*market skills*) and (4) returns information of which *market skills* are covered or missing in the provided document compared to the market's needs [8].

To be able to compare the skills in the provided document to the *market skills*, we need to cope with the challenges of different levels of abstraction and synonyms among the skills in the uploaded document and the *market skills*. Consequently, we apply the following 4 steps when we gather the *market skills* and when we upload a document to be analyzed which are visualized in Figure 2:

1. *Skill Extraction*: Extract skill requirements.
2. *Vectorization*: Map skill requirements to a semantic vector space, where skills with similar meanings are closer together and skills with different meanings are farther apart.
3. *Clustering*: Cluster skill requirements to cope with the challenges of different levels of abstraction and synonyms.

4. *Comparison and Analysis*: Compute intersections among the skill sets of the provided documents and the *market skills* and visualize recommendations based on covered and missing *market skills*.

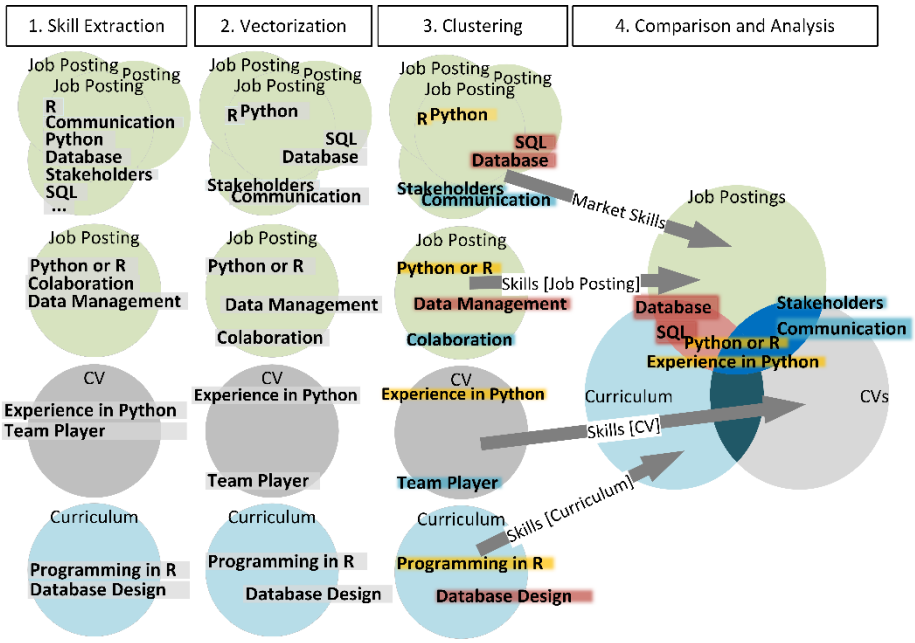


Fig. 2. Our Pipeline to Extract, Vectorize, Cluster and Compare Skills.

5.2 Implementation

In job postings, CVs and learning curricula, skills are usually expressed in bullet points. Therefore, in step 1 *Skill Extraction*, we developed keyword- and rule-based techniques to extract bullet points from these sources. For job postings, we used the BeautifulSoup package [18] to gather and extract 21.5k bullet points from 2,633 job postings for data scientists in English from Indeed.com and Kaggle.com. In this work, we refer to this representative set of skills as the *market skills*. Since some bullet points in a job posting are not skill requirements, we analyzed methods to deal with outliers that are not skill requirements in step 3 *Clustering*.

Like [2,6], we experimented with word embeddings to vectorize the skills in step 2 *Vectorization*. To represent the skills which usually consist of several words, we investigated stacking and averaging the word embeddings in a skill after they were produced with Word2Vec [3] and GloVe [19]. In addition, we explored sentence embeddings. As Bidirectional Encoder Representations from Transformers (BERT) [7] models are successful in NLP tasks, we also experimented with Sentence-BERT [20], a modification of the pre-trained BERT transformers. Sentence-BERT (44.2%) outperformed word embedding like GloVe (39.5%) by 12% in Silhouette score [21] at the end of our pipeline.

The benefit of our clustering approach compared to a taxonomy like in [17] is that our model can pick up new skills without the need to update a taxonomy. While hierarchical clustering approaches have not proven to be robust against outliers [22], K-means clustering has been successfully used in clustering word embeddings [23] and is adaptable and scalable [24]. Consequently, we used K-means to cluster our 768-dimensional vectors with the cosine distance as the distance metric. K was chosen as 31 with the highest Silhouette score of 44%. To remove outliers in the vectorized skills and allow our clustering techniques to perform better, we experimented with combinations of PCA [25], UMAP [26], and DBSCAN [27]. Using UMAP to reduce the vectorized skills to two dimensions and DBSCAN to remove outliers in the 2-dimensional space performed best according to our manual checks and reduced the 21.5k potential skills retrieved with our web scraper to 18.8k skills.

After retrieving clusters and vectors representing the skill of each cluster, we perform mathematical operations to find covered and missing *market skills*. For example, Figure 3 shows a section of a report in Skill Scanner where overlaps between skills in job postings and learning curricula were calculated. There, using the visualizations, educational institutions are shown the importance of certain skills in data scientist profiles along with the lack of coverage of those skills in their curricula.

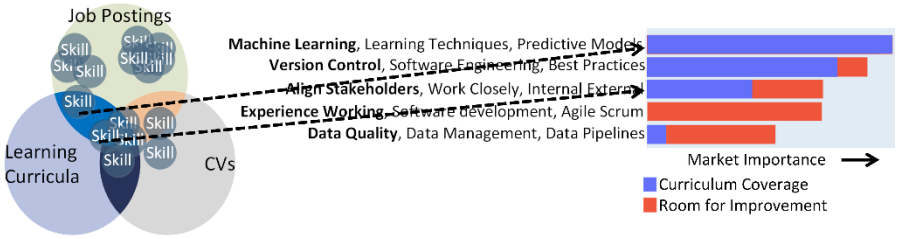


Fig. 3. Skill Comparison and Analysis between Job Postings and Learning Curricula.

4 Reports for Employers, Job Seekers and Educational Institutions

In this section we will present reports for employers, job seekers, and educational institutions which *Skill Scanner* outputs after analyzing the skills of the provided document and the *market skills*.

4.1 CV-Market Report

Figure 4 demonstrates an excerpt of the *CV-Market Report* which is generated if *Skill Scanner* receives a CV as input given the *market skills*. Alternatively, the report can be generated when a CV and a single job posting are provided and then compared. The *CV-Market Report* visualizes the coverage and importance of the skills in the CV which supports (1) job seekers to find out which skills are still missing, and which are already covered in their CV to be used for the application for a desired position, and

(2) employers to find out which skills the applicant is still missing, and which are already covered when applying for an advertised position.

In case of a single provided job posting, the bars show exclusively the skills specified in this provided job posting. Otherwise, the bars show all *market skills*. In both cases the skills are sorted by importance in the *market skills* which is represented by the bar lengths. Each skill is described by the 3 most frequent bigrams used in the *market skills*. The blue part in each bar demonstrates how well the skills in the provided CV match, whereas the red part shows how much is missing indicating the room for improvement. This representation of the skills with the bigrams, the coverage and the room for improvement is used consistently in all reports.

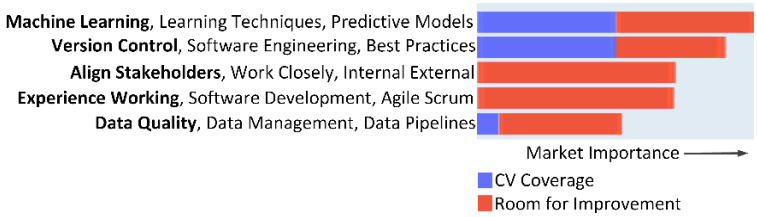


Fig. 4. CV-Market Report.

Technically, the bigrams at each bar on the y-axis are the most common bigrams that are located in a *market skill* cluster gained in step 3 *Clustering* of *Skill Scanner*'s pipeline (see Figure 2). How well a skill in the provided CV matches a skill in the job posting or in the *market skills* is determined by the distance of the skill's vector specified in the CV to the centroid of the cluster.

4.2 CV-Curriculum Report

Figure 5 demonstrates an excerpt of the *CV-Curriculum Report* which displays learning modules that best cover the skill gaps of an input CV given the *market skills* and a set of learning modules. The *CV-Curriculum Report* supports (1) job seekers to find a study module for the targeted expansion of skills with regard to a desired job position, (2) employers to find study modules for the targeted upskilling of employees, and (3) educational institutions to attract and advise students.

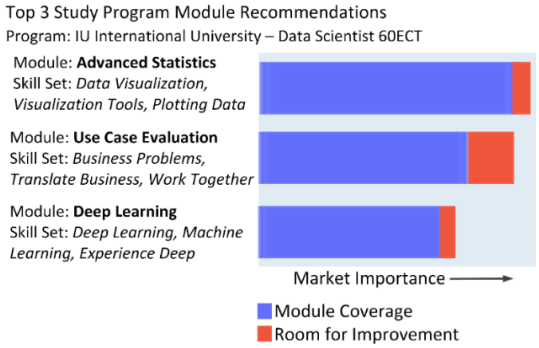


Fig. 5. CV-Curriculum Report.

4.3 Curriculum-Market Reports

Figure 6 shows an excerpt of the *Curriculum-Market Report* which is generated if *Skill Scanner* receives a learning curriculum as input given the *market skills*. The *Curriculum-Market Report* displays the coverage and importance of the skills in the curriculum and the *market skills* which helps educational institutions adapt the taught content with regard to the skills required in the job market.

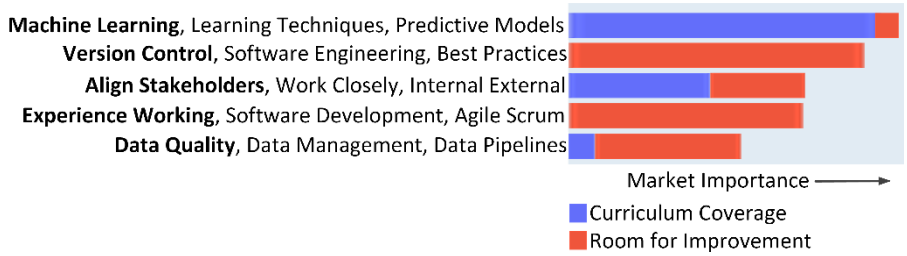


Fig. 6. Curriculum-Market Report.

4.4 CV-CVs Report

Figure 7 demonstrates an excerpt of the *CV-CVs Report* which is generated if *Skill Scanner* receives a CV given other CVs, a job posting, and the *market skills*. The *CV-CVs Report* visualizes a comparison of skill coverage in one CV to skills in other CVs which supports employers to select the best candidate from a group of applicants by processing each CV and comparing the scores.

How well a skill in the provided CV matches a skill in the job posting is determined by the distance of the skill's vector specified in the CV to the centroid of the skill cluster gained in step 3 *Clustering* of *Skill Scanner*'s pipeline (see Figure 2). The average of these distances determines the applicant's score. The score is computed for each CV, then the CVs are ranked by their scores.

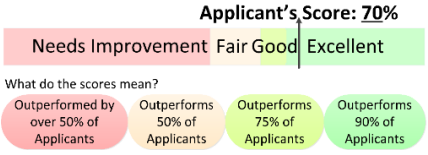


Fig. 7. CV-CVs Report.

5 Feedback from Employers, Job Seekers, and Educational Institutions

In this section we describe the design and results of our survey, in which we asked for feedback on our reports.

Table 1. Overview of which reports were shown in which questionnaire**.**

Job seekers	*CV-Curriculum Report*	*CV-Market Report*
Employers	*CV-Curriculum Report*	*CV-CVs Report*
Educational institutions	*CV-Curriculum Report*	*Curriculum-Market Report*

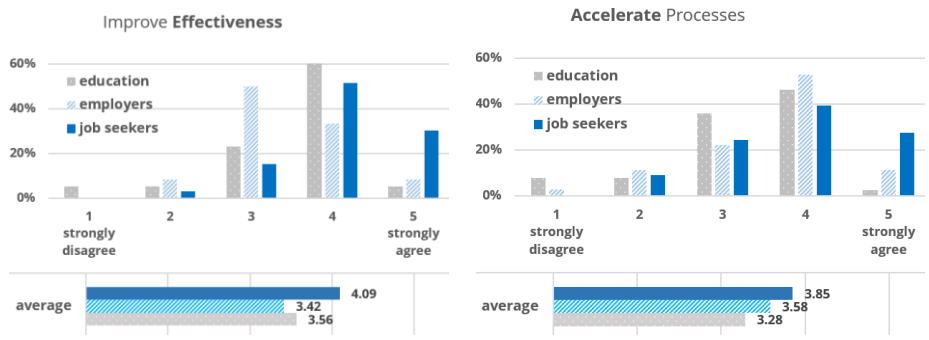
5.1 Experimental Setup

As described *Skill Scanner* receives a set of skills from a document, compares it to the job market's demands and returns reports based on the input document. To figure out if with the help of these reports processes related to skills are carried out more effectively, faster, fairer, more explainably, and in a more supported manner, we analyzed the feedback on the reports with 3 questionnaires—1 questionnaire for representatives of job seekers, 1 questionnaire for representatives of employers, and 1 questionnaire for representatives of educational institutions. Table 1 gives an overview of which reports were shown in which questionnaire.

In each questionnaire, we asked questions about the reports presented. The participants evaluated most questions with a score. The score range follows the rules of a forced choice Likert scale, which ranges from (1) *strongly disagree* to (5) *strongly agree*. Each questionnaire was designed in English and translated to Dutch and German. In total, 108 participants (54 female, 54 male) filled out our questionnaires. Most of them live either in the Netherlands (68.52%) or in Germany (28.71%). Participants were evenly distributed among the three user groups: 33 stated that they were job seekers (30.56%), 36 reported working in the human resources department of employers (33.33%) and 39 reported working for an educational institution (36.11%).

5.2 Effectiveness

First, we asked if the participants agree that *Skill Scanner* would help performing their skill-related tasks more effectively. Figure 8 (a) illustrates the feedback of our 36 representatives of *employers*, 33 representatives of *job seekers*, and 39 representatives of educational institutions (*education*). While *effectiveness* was rated best with 4.09 (*agree*) on average by *job seekers*, it was rated with 3.56 (between *neutral* and *agree*) by *education* and 3.42 (between *neutral* and *agree*) by *employers*. The feedback from the *job seekers* is 20% better than from *employers* and 15% better than from *education*.

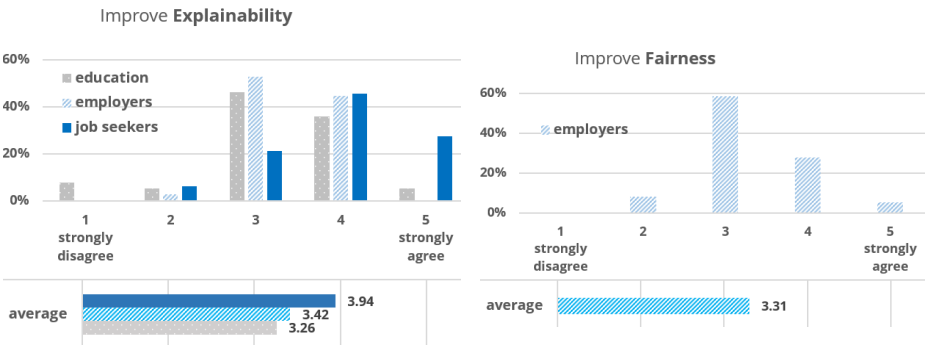


(a) Effectiveness (b) Speedup

Fig. 8. Feedback on *effectiveness* and *speedup*.

5.3 Speedup

Figure 8 (b) shows our results in relation to the potential speedup of skill-related tasks with *Skill Scanner*. The feedback from *job seekers* is again the best with 3.85 on average (*agree*). This time the feedback from *employers*, with an average of 3.58 (between *neural* and *agree*), is better than that from *education* with 3.28 (better than *neutral*). The feedback from *job seekers* is 8% better than from *employers* and 17% better than from *education.*



(a) Explainability (b) Fairness

Fig. 9. Feedback on *explainability* and *fairness*.

5.4 Explainability

Then we asked the participants if they agree that *Skill Scanner* helps explain strengths and weaknesses in CVs and curricula. The results are demonstrated in Figure 9 (a). While the feedback from *job seekers* is again rated highest with 3.94 (*agree*) on average, *employers* rated with 4.42 (between *neutral* and *agree*) and *education* with 3.26 (*neutral*) on average. The feedback from *job seekers* is 15% better than from *employers* and 21% better than from *education*.

5.5 Fairness

To find out if *Skill Scanner* can contribute to a fairer selection of applicants based on the analyzed CVs, we asked only the employers if they agree, since their HR departments look through the CVs. The feedback is visualized in Figure 9 (b) and is a little better than *neutral* with an average of 3.31. *Skill Scanner*'s impact on fairness is not as great as for the other aspects we asked about, since sympathy and soft skills also play a role in hiring an applicant fairly.

5.6 Usage

89% of all participants are not averse to apply our recommendation system. As with all other questions, job seekers were the most agreeable respondents when asked about usage. 67% of job seekers would certainly use *Skill Scanner*. This might be explained by the ease with which job seekers could adopt *Skill Scanner* in an application process. For employers and educational institutions, the introduction of our recommendation system would mean that they would have to change many processes, which is why they are likely to be more critical.

6 Conclusion and Future Work

The labor market dictates what job seekers should learn, and educational institutions should teach. Therefore, *Skill Scanner* processes skills in job postings, CVs, and curricula and outputs recommendations for employers, job seekers, and educational institutions based on present and missing skills and their importance to employers. *Skill Scanner*'s reports were shown to 108 representatives of our 3 parties in a survey. The majority finds that with our system, skill-related processes can be more effective, faster, fairer, more explainable, more autonomous and in a more supported manner. After these initial estimates of Skill Scanner's potential, further analysis could include measuring time and cost savings. Future work may also be to apply our pipeline to other job positions and expand it to other domains. In addition, we would like to extend *Skill Scanner* with further reports, e.g., based on a comparison of job posting and *market skills*, which is easily possible with our clustering pipeline.

References

1. Palmer, R.: Jobs and Skills Mismatch in the Informal Economy. 978-92-2-131613-8 (2017)
2. Fernández-Reyes, F.C., Shinde, S.: CV Retrieval System Based on Job Description Matching Using Hybrid Word Embeddings, Computer Speech & Language, vol 56 (2019)
3. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. ICLR (Workshop Poster) (2013)
4. Geyik, S.C., Guo, Q., Hu, B., Ozcaglar, C., Thakkar, K., Wu, X., Kenthapadi, K.: Talent Search and Recommendation Systems at LinkedIn: Practical Challenges and Lessons Learned. SIGIR (2018)

5. Guruge, D.B., Kadel, R., Halder, S.J.: The State of the Art in Methodologies of Course Recommender Systems—A Review of Recent Research Data, 6(2), 18 (2021)
6. Wang, Y., Allouache, Y., Joubert, C.: Analysing CV Corpus for Finding Suitable Candidates using Knowledge Graph and BERT. DBKDA (2021)
7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019)
8. Bothmer, K., Schlippe, T.: Investigating Natural Language Processing Techniques for a Recommendation System to Support Employers, Job Seekers and Educational Institutions. The 23rd International Conference on Artificial Intelligence in Education (AIED) (2022).
9. Baškarada, S., Koronios, A.: Unicorn Data Scientist: The Rarest of Breeds, Program: Electronic Library and Information Systems, Vol. 51 No. 1, pp. 65–74. (2017)
10. Faliagka, E., Iliadis, L. Karydis, I., Rigou, M., Sioutas, S., Tsakalidis, A. Tzimas, G.: Online Consistent Ranking on E-Recruitment: Seeking the Truth Behind a Well-Formed CV. Artif Intell Rev 42, pp. 515–528 (2014)
11. Si-ting, Z., Wenxing, H., Ning, Z., Fan, Yang: Job Recommender Systems: A Survey. ICCSE (2012)
12. Hong, W., Zheng, S., Wang, H., & Shi, J.: A Job Recommender System Based on User Clustering. J. Comput., 8, 1960-1967 (2013)
13. Alotaibi, S: A Survey of Job Recommender Systems. Int. J. Phys. Sci. (2012)
14. Diaby, M., Viennet, E., Launay, T.: Toward the Next Generation of Recruitment Tools: An Online Social Network-Based Job Recommender System. ASONAM (2013)
15. Li, J., Arya, D., Ha-Thuc, V., Sinha, S.: How to Get Them a Dream Job? Entity-Aware Features for Personalized Job Search Ranking. SIGKDD (2016)
16. Deepani B. Guruge, Rajan Kadel, and Sharly J. Halder: The State of the Art in Methodologies of Course Recommender Systems—A Review of Recent Research. Data 6, no. 2: 18. (2021)
17. Wang, C., Zhu, H., Wang, P., Zhu, C., Zhang, X., Chen, E., Xiong, H.: Personalized and Explainable Employee Training Course Recommendations: A Bayesian Variational Approach. ACM Trans. Inf. Syst. (2021)
18. Hajba, G.L.: Using Beautiful Soup. In: Website Scraping with Python. Apress (2018)
19. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for Word Representation. EMNLP (2014)
20. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, EMNLP-IJCNLP (2019)
21. Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics. 20: 53–65. (1987)
22. Rani, Y., Rohil, H.: A Study of Hierarchical Clustering Algorithm. International Journal of Information and Computation Technology (Vol. 3, Issue 10) (2013)
23. Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G.; Does Deep Learning Help Topic Extraction? A Kernel K-Means Clustering Method with Word Embedding. Journal of Informetrics, 12 (4), 1099–1117 (2018)
24. Lloyd, S.P.: Least Squares Quantization in PCM. Techn. Report RR-5497, Bell Lab (1957)
25. Pearson, K.: On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine. 2 (11): 559–572 (1901)
26. McInnes, L., Healy J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv, abs/1802.03426 (2018)
27. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD. AAAI Press, 226–231 (1996)